Full title: Development and Quality Appraisal of a new English Breast Screening linked dataset as part of the Age, test Threshold and frequency of Mammography screening (ATHENA-M) study

Short title: Development and Quality Appraisal of the ATHENA-M dataset

Type of Manuscript: Original Research

Journal: *British Journal of Radiology*

Author names:

Julia Brettschneider(*)[1]PhD, Breanna Morrison(*)[2] PhD, David Jenkinson[3] PhD, Karoline Freeman[3] PhD, Jackie Walton[4] PhD, Alice Sitch[2] PhD, Sue Hudson[5] BSc MSc, Olive Kearins[4] MSc, Alice Mansbridge[3] BSc, Sarah E Pinder[6,7] FRCPath, Rosalind Given-Wilson [8] MBBS, FRCR, Louise Wilkinson[9] BMBCh, FRCR, Matthew G Wallis[10] MBChB, FRCR, Shan Cheung[4], MPhil Statistics, Sian Taylor-Phillips[3] PhD

(*) These authors contributed equally

Correspondence to s.taylor-phillips@warwick.ac.uk

Author affiliations:

1. Department of Statistics, University of Warwick, Coventry, UK
2. University of Birmingham, Edgbaston, Birmingham, UK
3. Warwick Medical School, University of Warwick, Coventry, UK
4. Screening Quality Assurance Service, NHS England, UK
5. Peel & Schriek Consulting Ltd, London, UK
6. School of Cancer & Pharmaceutical Sciences, King's College London, London, UK
7. Comprehensive Cancer Centre at Guy's Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK
8. St George's University Hospitals NHS Foundation Trust, London, UK
9. Oxford Breast Imaging Centre, Churchill Hospital, Oxford, UK
10. Cambridge Breast Unit and NIHR Cambridge Biomedical Research Centre, Cambridge University Hospitals NHS Trust, Cambridge, UK

Acknowledgments: We gratefully acknowledge the significant contribution of Jackie Charman and the team at the National Disease Registration Service (NDRS) within NHS England for providing expert guidance on appropriate data to include, developing and delivering systems for data linkage, and expertise and advice in understanding data variables and quality.

Data source and access:

This project involves data derived from patient-level information collected by the NHS, as part of the care and support of cancer patients. The Cancer Registration data are collated, maintained and quality assured by the National Disease Registration Service, which is part of NHS England. The Screening data are collated, maintained and quality assured by the Screening Quality Assurance Service at NHS England. All prospective and retrospective English studies are evaluated by the Breast Screening Research Advisory Committee (RIDAC). Researchers wishing to access data used in this study should contact the corresponding author who will guide them to the contemporary processes.

Ethics approval: The Observational study of Age, test THreshold and frequency on English NAtional Mammography screening outcomes (ATHENA-M) study has NHS ethical approval (ATHENA-M: 21/LO/0120) and Public Health England Office for Data Release Approval (ATHENA-M:

ODR1920_283). This work was undertaken on the instruction of the NHS Breast Screening Programme to provide evaluation of efficacy. Public Health England processed personal data of eligible women to assure the quality and delivery of the breast screening programme under Section 251 of the NHS Act 2006, with BSP Research Advisory Committee giving their approval (ATHENA-M: BSP RAC 089).

Funding: This report is independent research arising from an NIHR Career Development Fellowship for STP (the POSTBOx study, CDF-2016-09-018) and the ATHENA-M project (NIHR130107), funded from the NIHR's Health and Social Care Delivery (HS&DR) Research programme. The study sponsor is the University of Warwick. In addition, STP is funded by an NIHR Research Professorship (NIHR302434), KF was funded by the NIHR through a Development and Skills Enhancement Fellowship and this research was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health and Research or the Department of Health. The funder has no role in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication.

Competing interests: STP was funded by the NIHR through a previous career development fellowship (CDF-2016-09-018) and by a current NIHR Research Professorship (NIHR302434). KF was funded by the NIHR through a Development and Skills Enhancement Fellowship.

Abstract

Objectives: To build a dataset capturing the whole breast cancer screening journey from individual breast cancer screening records to outcomes and assess data quality.

Methods: Routine screening records (invitation, attendance, test results) from all 79 English NHS breast screening centres between 1st January 1988 and 31st March 2018 were linked to cancer registry (cancer characteristics and treatment) and national mortality data. Data quality was assessed using comparability, validity, timeliness, and completeness.

Results: Screening records were extracted from 76/79 English breast screening centres, 3/79 were not possible due to software issues. Data linkage was successful from 1997 after introduction of a universal identifier for women (NHS number). Prior to 1997 outcome data are incomplete due to linkage issues, reducing validity. Between 1st January 1997 and 31st March 2018, a total of 11,262,730 women were offered screening of whom 9,516,953 attended at least one appointment, with 139 million person-years of follow-up (a median of 12.4 person years for each woman included) to 86,009 breast cancer diagnoses and 995,657 any-cause deaths. Comparability to reference datasets and internal validity were demonstrated. Data completeness was high for core screening variables (>99%) and main cancer outcomes (>95%).

Conclusions: The ATHENA-M project has created a large high-quality and representative dataset of individual women's screening trajectories and outcomes in England from 1997 to 2018, data before 1997 are lower quality.

Advances in knowledge: This is the most complete dataset of English breast screening records and outcomes constructed to date, which can be used to evaluate and optimise screening.

Introduction

Data collected through screening programmes can support studies on the epidemiology of breast cancer[1,2], the effectiveness of screening programmes,[3,4] the variation in cancer prevention practice due to technology or process,[5-7] cost-effectiveness, [8-10] potential biases,[11] the suitability for application of AI in screening image analysis[12], and the potential and implementation of risk-stratification[13,14].

Descriptions of individual breast screening observational databases in several countries have been published, including in the USA[15,16,17], Denmark[18], and Korea[19,20]. However, to the best of our knowledge, to date, there is no publication reporting the data quality of routine breast screening data. Available studies focus on the quality and audit of the breast screening programme rather than the screening data itself [21,22]. This is also true for other cancer screening[23,24].

Three features that make English screening datasets particularly attractive are the volume of data (up to 30 years follow-up for 13 million women), inclusion of large parts of the eligible population and the relatively homogeneous organisation under the umbrella of a national health system. Less systematic approaches bear risks of bias such as distortion linked to accessibility heterogeneity[25,26], which applies naturally in countries where health care provision is associated with higher socio-economic status. Scandinavian countries have relatively homogenous access to health care and have a tradition of maintaining excellent records, but these datasets are smaller and the populations are less ethnically diverse, limiting generalisability and transferability. A key to delivering on the promises implied by the characteristics of the English dataset is their quality and successful linkage of the 83 separate parts of the database including 79 screening centre datasets and four datasets about cancer outcomes, invitation records, socio-economic background, and mortality.

In 2009, Bray and Parkin published guidance on the practical aspects and techniques for addressing data quality at the cancer registry, considering comparability, validity, timeliness and completeness[27,28]. This framework has been used for the evaluation of the Swedish breast cancer registry[29] and cancer registries more generally in the UK[30], Iceland[31], Finland[32], Norway[13,33], Bulgaria[34], Ukraine[35], and Singapore[36]. Other studies examining the quality of cancer registry data focused on completeness only[37-40], or on completeness and timing[41]. The UK government has recently laid out a data quality framework based on the Bray and Parkin framework (Gov guidelines, 12/2020), but this is the first time such a framework has been applied to breast cancer screening data. Data quality assessment for observational health studies has come under the spotlight due to the risk of misclassification, bias, and hence potential irreproducibility observed; for example, with the use of electronic health records, real-world evidence in pragmatic clinical trials, and repositories such as UK Biobank[42-45].

This first aim of this paper is to describe the construction of the ATHENA-M dataset by combining 83 existing datasets from different sources. Through comprehensive linkage of individual women's trajectories, we provide a rich resource for future studies improving the quality and effectiveness of screening programmes. The second aim of this paper is to assess the ATHENA-M from a data quality perspective to ensure reproducibility of findings. We set up a framework based on four common pillars for data quality, along with concrete quality checks tailored to a composite longitudinal data repository for cancer screening.

Methods

ATHENA-M is a unique composite dataset created from repositories of the National Breast Screening Service (NBSS) at 79 separate breast screening centres, the national invitation system Breast Screening Select, Office of National Statistics mortality data via the Public Health England Mortality and Birth Information System (PHE-MBIS), the National Cancer Registration and Analysis Service (NCRAS), and the Index of Multiple Deprivation derived from postcode (IMD) – see Table 1 below for more information.

Population: Inclusion criteria

In ATHENA-M we included all women invited to at least one breast screening appointment in England between ages 47 and 73, between 1st January 1988 and 31st March 2018, based on date of

first offered appointment. We excluded women without a screening invitation accompanied by date information, and women whose appointment was not part of the standard breast screening programme (for example women who self-referred with symptoms). Women who opted out of having their data being held on the National Disease Registration Service (NDRS) registers had already been applied to the cancer registry data. In line with the National Data Opt-Out policy, this opt out was not applied to the screening data as no confidential patient information was shared with any organisation external to Public Health England (PHE) (details in Appendix C2; the rate of national data opt out at that time was 5.3%).

Population: Screening protocol in England from its initial roll-out until today

The roll-out of the national breast screening programme for women aged 50–64 began in 1988 in selected areas and was extended to cover the whole of England from 1990. Each woman is invited once every three years. Changes in the programme's operation include extensions of ages eligible for screening, the increased involvement of a second reader to search for signs of cancer on the mammograms, harmonisation of the administrative systems, technological changes, and some modifications to breast cancer classification[46]. Extensions of the invited population may affect the prevalence of cancer, and improvements in medical diagnostics may affect detection and observed characteristics of cancers. For example, technological developments such as the rollout of digital mammography has increased the rate of DCIS diagnosis[47] which may also have played a role in the change in reported DCIS grade classification (less low, more high grade) that has been observed in parallel[48-50], and more accurate node staging may lead to the detection of more metastases of smaller size. Similarly, the evolution of audit and quality assurance processes, and key performance indicators have driven changes in practice[51,52]. The modifications most notable for this study are visualised in Figure 1 and described in detail in Appendix A.

Data source/s and pre-processing

The ATHENA-M dataset draws on several *a priori* independent data repositories with their own history that need to be characterised and pre-processed. Table 1 summarises these data sources with more details in Appendix B1.

Table 1: Data sources and preprocessing for ATHENA-M

National Breast Screening Service (NBSS)	In the first decade of the programme IT support was decided regionally leading to the use of NBSS and 4 other administrative systems operating locally until NBSS became the nationwide standard in 2004–2005 and has remained largely unchanged since. Data were collected between November 2018 and May 2019 from each of the 79 breast screening services using a standalone set of extract programmes written using SAP® Crystal Reports® and an Open Database Connectivity (ODBC) interface (standard with the NBSS system implementation), saved in text format and sent to Public Health England for collation and cleaning. Three extracts were taken from each centre: details of eligible women invited to screening (NBSS-women), details of eligible routine screening episodes (invitation for screening and all of the associated actions that happen as a result) (NBSS-episode) and clinical details of screening episodes where the woman was recalled for further tests (NBSS-feature).
National Cancer Registration and Analysis Service (NCRAS)	The systematic collection of cancer and tumour disease data in England is managed by NCRAS with over 300,000 cases of all cancers collected annually, including patient details, cancer type, and information on severity and received treatment. Data from health care providers, histopathology and haematology services, radiotherapy departments, screening services, general practitioners and other services are matched and merged to build a complete picture of the cancer incidence in England and to understand how cancer patients are diagnosed, treated and what their outcomes are. Once all expected records for any one

	incidence year have been received, validated, and quality assured, NCRAS takes a snapshot of the dataset providing a single, consistent source of cancer registrations. We used the August 2021 snapshot for linking screening data to registered patients.
Breast Screening Select (BS Select)	The dataset of the national invitation system for the NHS breast screening programme in England (BS Select) dates back to the beginning of the programme in 1988 and contains women registered with a general practitioner in England. It is used to automatically send a list of women who are due an invitation, based on the parameters set when creating a batch, to the responsible screening office who imports this list into NBSS. It receives in turn a screen outcome for each of these episodes. NBSS includes all routine screening call and recall appointments, as per study inclusion criteria, but initial data cleaning suggested missing data at a subset of centres. BS Select was used to check whether women eligible for routine call and recall were recorded as other appointment types: self-referrals; general practitioner referrals; higher risk referrals; and non-routine early recall appointments.
Index of Multiple Deprivation (IMD)	As a frequently used compositive measure for relative deprivation in small areas, the IMD captures components such as income, employment, education, health, crime, housing and services, and living environment. In England, it is revised every few years by the UK Ministry of Housing, Communities and Local Government (MHCLG)[53]. ATHENA-M includes the quintiles of the income domain using the women's postcode at the time of her last screening appointment. To reflect revisions, both a score based on the IMD current (at that time) and on IMD 2015 are included.
Office for National Statistics (ONS) Death Records	The Public Health England Mortality and Birth Information System (PHE-MBIS) was created to streamline the sharing, storage and dissemination of ONS birth and death registration by PHE. The data was released under the control of the PHE Office for Data Release. Recording of death data on PHE-MBIS started in 1997 (month unknown).

Data linkage

Figure 2 shows the variables used for linkage between datasets. Linkage between datasets was primarily based on (pseudonymised) NHS number, a unique identifier used across all NHS services for each woman which became universal in 1997 (Figure 5).

Records belonging to the same woman using different centres could be matched, but records of the same film reader operating across centres could not be linked. Details about scoring systems used where necessary and record matching counts are shown in Appendix B (Tables B3.2-5). Whilst linkage between the cancer registry and NBSS for each woman could utilise NHS number, there was no identifier linking cancer records for a woman to screening records, and at the point of data extraction the cancer registry did not contain reliable data about whether a cancer was screen detected. NBSS episodes with screen detected cancer will have the data from the Cancer Registry linked to it if diagnosis date in the Cancer Registry was between 7 days before and 100 days after the 'screening date' or 'date taken'. The remaining 559,443 cancers in the Cancer Registry are classified as non-screen detected cancer.

Quality assessment of ATHENA-M dataset

Central pillars for data quality in cancer registries are comparability, validity, timeliness, and completeness[27,28]. A rich collection of generic data quality indicators related to these concepts has been suggested to cover a wide range of observational health studies[54] expanding existing frameworks developed for electronic heath records. We adopt these frameworks to assess the quality of composite data about cancer outcomes and the screening journey preceding it. Table 2 shows data quality pillars tailored to the architecture of the ATHENA-M dataset and lists the main criteria we used to assess

them. We give particular attention to the technological and procedural changes to the screening programme, modifications of administrative processes, changes in cancer classification, and heterogeneity of the invited population. The distinction into crude and qualified missingness addressing nonresponse, drop-out, and other specific reasons[54] is particularly suited to addressing the missingness occurring through attrition and failed linkage.

Table 2: Pillars and criteria to assess data quality in the context of ATHENA-M

Completeness	Refers to the extent to which screening records, cancer outcomes and sociodemographic information are included in the database. Missingness can relate to the lack of inclusion of potentially relevant variables or to the lack of values in included variables. In the longitudinal context of screening journeys, missingness needs to be considered across the whole time period.	• Missingness in cancer registry variables (Appendix Table C1) • Excluded centres (Appendix Table C2.1-3) • Missingness of age over the whole study period (percentage) • Missingness in variables from NBSS pre 1997 and later (Table 3, Figure 4) • Failed linkage to cancer registry pre 1997 and later (Table 3) • Missing NHS number and failed linkage by year (Figure 5, Appendix Table B3.1) • Missingness in NCRAS variables pre 1997 and later (Table 3) • Non-attended invitations (Table Appendix Table C4)
Comparability	Assesses the calibration of the generation of statistics from different population groups associated with different centres, regions, socio-economic status and demographic characteristics. In our longitudinal setting comparability needs to be addressed over time as well. A basic requirement is standardisation of definitions and practices concerning classification and coding of screening and cancer outcomes.	• Benchmarking of NBSS data against KC62 data for numbers of screens, recalls, and cancers over the course of the study period (Figure 6) • Benchmarking of NBSS data against ONS data for cohort age over the course of the study period (median and IQR)
Validity (accuracy, plausibility, correctness)	Refers to the proportion of cases in the screening results and potential subsequent cancer related outcomes for which given characteristics truly have that attribute. It depends on the precision of the diagnostic process and the level of expertise in abstracting, coding, and recording.	• Discordance between cancer indicator and recorded action following screen (percentage) • Consistency between screening and mammography date percentage) • Recalls for screen detected cancers (percentage)
Timeliness (currency)	Reflects the degree of updating speed in the screening records and, if applicable, cancer-related outcomes.	• Release timelines after censoring (numerical) • Time course of screening patterns on the cohort level (Figure 7) • Attendance at second screening appointment over time (Appendix Table C5, Figure C6)

Data quality assessment is the basis for reproducible results and has a long tradition in cancer registration originally based on the first three pillars[55]. Timeliness was added because in its absence no accurate trends can be estimated. Completeness is the most straight forward to assess superficially, but its potential implications for statistical inference depend on whether the missing values followed any systematic patterns or not. In the latter case, the occurrence of missing values can be tolerated at relatively high levels, but if the occurrence of missing values in one variable is linked to other variables this seriously impact conclusions. There is a well-developed body of literature defining different notions of random versus systematic missingness, reasons for this, detection strategies, and remedies ranging from imputation techniques in the case of missingness at random to model-based approaches involving knowledge about the missingness patterns[56]. In practice, optimising data quality can involve trade-offs such as between timely data and the extent to which they are complete and accurate. Table 2 summarises how the four pillars of data quality were assessed giving a conceptual explanation as well as technical criteria used to derive the findings listed in the result section.

Completeness was assessed at four levels. Firstly, the number of centres that contributed data, secondly, missing data in the NBSS dataset, thirdly, missing linkage and related to that, fourthly, missing data on cancer and mortality information. In a wider understanding of the assessment of completeness we also include the issue of uptake of screening appointments. While driven by women's choices rather than by technical or administrative causes, high levels of appointments where the woman chose not to attend could potentially create similar limitations on the usability of the dataset as other types of missing data.

Comparability was assessed by benchmarking the NBSS data against two other data sources. Firstly, we compared numbers of screens, recalls and cancers in the NBSS data with the NHS Breast Screening Programme Central Return Data Sets (or KC62 data). As mandatory requirement, screening centres in England annually submit NHS Breast Screening Programme Central Return datasets (KC62)[57] with information about processes and outcomes. These are used to monitor management, progression towards achieving targets about cancer diagnosis, and numbers of women screened per centre[58]. Centres check data completeness in NBSS before running standard extractions for KC62, these extracts are in turn checked by regional quality assurance teams and finally by the national data analytical team. To investigate the amount of missingness, we compared the number of screened women between NBSS and KC62 records by centre annually, between April 2004 and March 2018, inclusive. Specifically, to make records comparable the number of screened episodes from the NBSS dataset were grouped by financial year (April to March) and screening service at the time when the screening appointment was sent out. Between April 2004 and March 2018, the age range 50–64 was used. The KC62 includes additional appointment types such as self-referral which were excluded from our analysis, so whilst we may expect consistent systematic differences this comparison enables identification of any major issues in data extraction or transfer. Secondly, we compared age of women in the NBSS data to ONS data for the relevant age range (50–70 years) for the years 2001–2018 for which ONS data were available. As a population-wide screening programme, this should be similar to national statistics in terms of age distribution.

Validity of the dataset was assessed in terms of concordance between different measurement methods for whether a cancer was detected and date of detection, and through logical consistency that every screen detected cancer should be preceded by a decision to recall.

Timeliness was assessed by quantifying the time required for ethical and other approvals, and for data extraction and linkage. Time course profiles are used to study patterns in the number of screening episodes on the cohort level. As in the case of completeness we adopt a slightly wider understanding of data quality by including aspects of timeliness driven by women's choices. Specifically, we assess overall screening uptake and attendance to second screening appointments.

Results

Exclusions

The initial dataset contained records for 13,260,132 women with a total of 53,471,265 screening episodes. As part of the data preparation, all or parts of the records of 276,353 of these women were converted from an invalid or old-style NHS number to a valid 10-digit NHS number using the tracing service, but this service did not work for all[59]. The screening dataset (NBSS-episode) was subject to exclusion steps (Figure 3) some of which also affected the NBSS-women dataset. The first four exclusions related to duplicated entries (N=8,451, 0.02%), technically inadequate mammograms that were subsequently repeated (N=411,979, 0.77%), and other multiple entries per screening appointment (N=60,999, 0.11%). The following four exclusions related to appointment dates classified as uninvited (e.g. due to cancellation by the centre), missing date information, dates out of range of the study period (N=258,023, 0.48%), and appointments for women outside the standard age range at screening (younger than 47 years or older than 73 years, N=375,892, 0.70%). Three specific centres had data collection issues, described in Appendix C. Women who only had data from these centres had all their screening appointments removed (N=596,379, 1.12%), but 345,578 screening appointments at these centres were kept in the dataset for women who also used other centres, to facilitate more complete screening records for these women. Table C2 shows that the three excluded centres have very similar characteristics to the other centres which ensures that their removal has a very limited effect on conclusions drawn from this dataset. The final dataset contained records for 13,094,122 women and 51,759,542 invitations to screening appointments, of which 38,319,093 (74.0%) were attended, resulting in 38,185,530 screens (73.8% of invitations). 2,271,367 (17%) women did not attend any episodes. The initial dataset did not contain non-routine appointments as they were not recorded as part of NBSS. However, taking into account also the BS Select records showed that the vast majority (95.1%) of all screening appointments were indeed routine appointment (NBSS records after exclusions). This is followed by self-referrals (3.8%) and GP-referrals (0.7%) as detailed with yearly breakdowns in Appendix Table C3. Overall, we had 139 million person-years of follow-up in this dataset (a median of 12.4 person years for each woman included).

Completeness (pillar 1)

Completeness in terms of screening centres was affected by the three centres that were excluded from the analysis, as discussed in the exclusions section. However, there were no extreme differences between the excluded centres and the included centres regarding screening outcomes (see Appendix Table C2).

Age information is nearly complete except for the first few years. While IMD was nearly complete (missingness rates of at most 1.9% in all time periods according to Appendix Table C2), ethnicity data is very sparse (only collected by a small number of centres in later phases of the programme).

Table 3: Missingness in screening process, linkage, and cancer characteristics

	Overall	1997 and later	Pre 1997
Successful screening episodes	**38,185,530**	**31,963,548**	**6,221,982**
Reader 1 recall decision			
Other *	58,934 (0.15%)	33,356 (0.10%)	25,578 (0.41%)
Missing	7,218 (0.02%)	4,655 (0.02%)	2,563 (0.04%)
Reader 2 recall decision			
Other*	34,146 (0.09%)	22,825 (0.07%)	11,321 (0.18%)
Missing**	7,336,330 (19.2%)	3,584,826 (11.2%)	3,751,504 (60.3%)
Final recall decision			
Invalid code/Missing	129,843 (0.34%)	79,175 (0.25%)	50,668 (0.81%)
Needle biopsy follow up tests			
Missing	6,941 (0.02%)	4,059 (0.01%)	2,882 (0.05%)
Cancer detected at screening	271,380 (0.71%)	238,922 (0.75%)	32,458 (0.52%)
Not linked to registry	14,614 (5.4%)	9,000 (3.8%)	5,614 (17.3%)
Linked to registry	**256,766 (94.6%)**	**229,922 (96.2%)**	**26,844 (82.7%)**
DCIS	**49,208**	**45,266**	**3,942**
Grade missing/invalid	19,112 (38.8%)	15,461 (34.2%)	3,651 (92.6%)
Size missing	32,986 (67.0%)	30,105 (66.5%)	2,881 (73.1%)
Invasive	**207,558**	**184,656)**	**22,902**
Grade other/missing	17,401 (8.4%)	8,144 (4.4%)	9,257 (40.4%)
Size Missing	37,026 (17.8%)	26,612 (14.4%)	10,414 (45.5%)
Node info missing	74,039 (35.7%)	55,791 (30.2%)	18,248 (79.7%)
ER Status missing	119,867 (57.8%)	97,215 (52.6%)	22,652 (98.9%)
PR Status missing	165,008 (79.5%)	142,183 (77.0%)	22,825 (99.7%)
HER2 Status missing	122,828 (59.2%)	100,026 (54.2%)	22,802 (99.6%)
Numerical Stage missing	69,309 (33.4%)	51,781 (28.0%)	17,528 (76.5%)
T Stage missing	75,218 (36.2%)	58,354 (31.6%)	16,864 (73.6%)
N Stage missing	74,222 (35.8%)	56,852 (30.8%)	17,370 (75.8%)
M Stage missing	143,229 (69.0%)	122,906 (66.6%)	20,323 (88.7%)

*Other refers to reader decisions that cannot be classified as recall or no recall, including technically inadequate mammograms, and recall with a shorter screening interval, ** A missing decision for reader 2 is often not missing data but represents a screening pathway where there is only one reader examining each womans mammograms. ER refers to estrogen receptor, PR refers to progesterone receptor, and HER2 refers to Human Epidermal Growth Factor Receptor 2.

Data such as biopsy information, film reader recall decision, and final recall decision should be present for every screening appointment. The percentage of records with those variables missing is listed in the upper part of Table 3. Of the 38,185,530 screening appointments, the decision about recall for further tests by the first reader was missing for only 7,218 (0.02%) and there was not a valid reader identifier for 676,785 (1.77%). The second reader's decision is missing for 7,336,330 (19.21%) appointments, but this reflects the gradual introduction of second readers. The final recall decision is only missing for 129,843 appointments (0.34%). Figure 5 shows missingness of these variables post 1997 at centre level for all sufficiently large centres. Missingness of reader and final recall decisions and biopsy information in these centres is generally below 1% and apart from a few outliers even under 0.05%.

A major driver of data completeness was linkage accuracy. In 1988, an NHS number used for record linkage, was only available for 7,438/13,019 (57.1%) of women, even after using the tracing service, however, by 1997, missingness was low with 208,240/209,319 (99.5%) of women having an NHS number available (Figure 5 top). The screenings in which a cancer was detected which could not be linked to cancer registry records and for whom therefore data items on the characteristics of the cancer were missing was 59/153 (38.6%) in 1988. After 1996 with mandated use of NHS number this was 9,000/238,922 (3.8%) (Figure 5 bottom). As a direct consequence of missing linkage, cancer type (DCIS or invasive) was missing for 14,613 (5.4%) records overall (Table 3).

For further cancer characteristics, missingness was a product of both invalid data linkage and missing data in the cancer registry itself which was substantial before 1997 (Figure 5). For instance, before 1997 information on grade was missing for 3,651/3,942 (92.6%) DCIS cases and 9,257/22,902 (40.4%) invasive cancers, while in the time period from 1997 onwards this reduced to 15,461/45,266 (34.2%) for DCIS and 8,144/184,656 (4.4%) for invasive cancer. Missing information on lesion size reduced from 2,881/3,942 (73.1%) to 30,105/45,266 (66.5%) for DCIS and from 9,257/22,902 (40.4%) to 8,144/ 184,656 (4.4%) for invasive cancers. No data are available to explain the greater missingness for DCIS than invasive cancer data, but we do know invasive cancer characteristics were used for quality assurance and clinical management decisions which may have increased completeness of reporting. Information on node involvement, receptor status, and numerical stage of invasive tumours was rarely reported at all before 1997. Further details about missingness in cancer registry variables are listed in Appendix Table C1.

Table C4 shows the distribution of women in the screening dataset by the number of non-attended invitations. The majority of women (53.7%) attended all the appointments they were invited to, while 32.5% did not attend one or two, and only 13.8% did not attend more than two.

Comparability (pillar 2)

The use of NBSS records created under the umbrella of the national health system has led to a high degree of consistency in procedures, variables names, and their meaning. It is matched by a similar level of standardisation of codes used at the cancer registry. Details can be found in the corresponding sections in Appendix A. Successful linkage of these repositories led to an unparalleled level of comparability of the data across the whole geographical area covered by the screening programme.

Benchmarking of the extracted NBSS data against KC62 data showed that the latter had overall slightly higher numbers of screens, recalls, and cancers (Figure 6). The difference can mostly be explained by women who self-refer or are referred by their GP for screening, rather than as part of the standard call-recall system, as these women are included in the KC62 data but not in the screening data. The KC62 data recorded an average of 90,663 more screens per year than the NBSS data, 5,857 more recalls, and 773 more women with cancer. KC62 numbers of screens, recalls, and cancers exceeded the NBSS count by at most 7.8%, 11.9%, and 9.6%. The systematic difference was consistent over time and aligns with the expected difference as the ATHENA-M dataset excludes self-referral appointment types which are included in KC62.

Comparison of women in the NBSS data with those in the ONS data for the years 2001–2018 shows that from 2006 onwards, the median woman's age for both the ONS data and the screening cohort

from NBSS was 59 years (IQR 54–64). Prior to 2006, women in the NBSS were slightly younger by a maximum of 2 years (2001 median 56 years (IQR 53–60) vs 58 years (IQR 54–64)).

Validity (pillar 3)

The cancer indicator from NBSS (whether cancer was detected following a screen) was discordant to the recorded action taken following screening from NBSS in 641/38,185,530 (0.002%) of cases. The date of screening and the date of mammography were identical in 38,164,605/38,169,905 (99.986%) of cases. Of the 271,380 screening appointments where cancer was detected, 1,602 (0.59%) did not appeared to have decided to recall the woman for the tests required to detect cancer. This is logically inconsistent, but in practice may occur rarely when a woman attends screening and symptomatic service in the same time period, or when unusual pathways are followed after a technical recall.

Generally, validity concerns were low and improved over time for all indicators assessed.

Timeliness (pillar 4)

The ATHENA-M dataset was censored in 2018 and released to researchers in 2022. The process of receiving NHS ethical approval took 4 months, the process of data extraction from all 79 breast screening centres took 9 months, the process of data linkage to other datasets took 11 months and approvals for data release took a further 2 years, partly impacted by the COVID-19 pandemic and reorganisation of healthcare structures. These delays provide some of the limitations to timeliness. Further timeliness is challenging to achieve if long term follow-up to outcomes is required for cohorts receiving 20 years of screening, in the context of changing tests and treatments.

Patterns of screening episodes for cohorts of women by year of first invitation are shown in Figure 7. The two first cohorts show no visible patterns due to their small size. A complete screening history of all 7 screening invitations is only available in those initially screened in 1998 and earlier as it covers a timespan of more than 20 years. There have been significant changes in screening technology, cancer prevalence and treatment effectiveness since then limiting generalisability of results to modern screening. Patterns show triannual cycles with some delays and decline in participation over the years. The fraction of women who attended second screening appointments within the expected time frame was initially less than 60%, but quickly rose in the 1990s to plateau around 78-80% in the decade following 1997 after which is slightly increased and stayed at levels 80-82% (Appendix Table C5 and Figure C6).

Discussion and conclusions

The creation of the ATHENA-M dataset involved three phases: acquisition and pre-processing of five raw data sources; linkage based on pseudonymised NHS numbers and scoring systems; and exclusions of a very small number of centres and of redundant or erratic individual episode records in other centres. The overall data quality of ATHENA-M is very good. Completeness in the screening journey core variables such attendance and reader decisions is excellent. Age and IMD score are also nearly complete. Cancer type (DCIS or invasive) is missing in about 5% of cases, but further cancer details have very high missingness before 1997 with moderate improvements afterwards. Ethnicity has only been collected sporadically in a small number of centres. The relatively high level of technical completeness of the data is to a large extent a reflection of the high level of women's participation in the screening programme (as evidenced by more than 86.2% of women attending all except up to two of the screening appointments offered to them). It is worth noting that there will also be unknown missing data, for example if a woman emigrated from England to another country we would have no records; however we expect these numbers to be very small.

Comparing ATHENA-M to mandatory KC62 records from 2004 onwards shows consistently smaller numbers in screens, recalls, and cancers, but the difference is consistent over time and explainable as missing non-routine cases. Using ONS data as a benchmark, there are small differences in age in the early 2000s, but no noticeable differences from 2006 onwards. Several rounds of exclusions and data cleaning have ensured that records are valid and unique. Validity for screening process variables was confirmed by very low discordance between findings at the screening appointment, resulting actions, and relevant dates. Reader information has a small percentage of invalid values until 2005 but is nearly perfect afterwards which coincides with introduction of automated data entry. Timeliness has been limited by delays in data release processes, and by the nature of the dataset where the

intervention lasts for up to 20 years. Timeliness in the wider sense as measured by attendance at the second screening appointment within the expected time interval was low initially but quickly rose and plateaued around 1997 (between 78% and 82% in all years since 1997).

ATHENA-M is a large composite dataset involving women's records drawn from two levels, centres and screening episodes representative of the English population eligible for breast cancer screening. The data is longitudinal with long follow-up time, especially for the older records, and benefits from using the same NBSS system across the same centres with standardised categorical data collection, large amounts of which are automated. A weakness is the only sporadic inclusion of ethnicity information rendering it unsuitable to address study questions around the role of ethnicity. Another limitation is the high missingness in details about the cancers (grade, size etc), especially in the early phases of the study period. This could lead to biased conclusions and confounding.

A US dataset with similar aims is presented by Lehman *et al*[17]. It used the powerful SEER platform and repositories[60], but it only covers 7 years of data in specific geographic areas, which may not be representative of the population of the whole country. They do not have a whole population call-recall system of systematic invitation for all eligible women, so it also may not be generalisable to all women within the geographic area. A Swedish dataset of women eligible for screening linked to breast cancers (from the Swedish Cancer Registry) and breast cancer deaths (from the National Cause of Death Register) was established for the evaluation of breast cancer mortality in Swedish breast cancer screening programmes[61]. One of the strengths of this Swedish screening dataset is the high attendance, which is rarely matched, but there are limitations arising from the relative homogeneity of that population.

ATHENA-M is suitable to take on the role of a reference dataset for cancer screening evaluation and research. For most objectives, we advise excluding the pre-1997 period when there was no universal unique identifier to ensure complete linkage to outcomes including cancer detection and mortality. Conclusions related to IMD also need to be drawn with care. While the data on IMD is fairly complete, it is based on the woman's most recent postcode which may not always best reflect the woman's socioeconomic status (e.g., not be up to date, not reflecting where she lived most of her life).

From a data maturity perspective, we identified a set of recommendations for future data collection in this and other population-wide cancer screening contexts:

- Development of a standardised customised data entry format with a user-friendly interface allowing frequent monitoring to ensure and improve data quality;
- Systems of instant data entry by clinical staff in predefined categories presented as user-friendly drop-down lists, without processes requiring clerical staff, and fully automated data collection for those fields where it is possible (such as image metadata, breast density, exposure factors, equipment, compression, reader identifier, reading time);
- Harmonisation of definitions related to the screening journey and outcomes to be used across centres, screening records, cancer registry, and in electronic health records in primary and secondary care;
- Introduction of unique reader identifiers to allow linkage of all screens looked at by the same reader within and across centres;
- Use of unique identifiers as suitable surrogates to ensure complete linkage between screening records and cancer registry for both women, and cancer episodes within women
- Improvement of cancer registry data completeness;
- Data linkage to or collection of accurate ethnicity information;
- Improvement of data sharing and accessibility to researchers and health care providers, whilst maintaining ethical and data governance standards and ensuring sufficient contextualisation to avoid concerns voiced in the context of AI[62];
- Building in mechanisms to integrate technical innovations, modifications of protocols, or inclusion of additional variables (e.g. individual risk factors) in a timely manner.

There are huge potential benefits in data linkage between screening programmes and outcome data, which can be used for research, quality assessment and service improvements, and to underpin data

collection for prospective research. In this retrospective study of population-wide English data we have demonstrated such linkage is possible on a large scale. Using a data quality assessment framework customised to screening journey and outcome data we found that ATHENA-M has an overall high level of quality. This work can also serve as a guide on how to construct similar datasets for other longitudinal screening programmes.

Declarations

The RECORD checklist (and extension of the STROBE statement for observational studies using routinely collected health data) has been completed.

References

1. Travis RC, Balkwill A, Fensom GK, et al. Night shift work and breast cancer incidence: three prospective studies and meta-analysis of published studies. *JNCI: Journal of the National Cancer Institute* 2016;108(12):djw169.
2. Evans DG, Brentnall AR, Harvie M, et al. Breast cancer risk in a screening cohort of Asian and white British/Irish women from Manchester UK. *BMC Public Health* 2018;18(1):1-7.
3. Massat NJ, Dibden A, Parmar D, et al. Impact of screening on breast cancer mortality: the UK program 20 years on. *Cancer Epidemiology and Prevention Biomarkers* 2016;25(3):455-62.
4. Blanks R, Moss S, McGahan C, et al. Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990-8: comparison of observed with predicted mortality. *Bmj* 2000;321(7262):665-69.
5. Taylor-Phillips S, Wallis MG, Jenkinson D, et al. Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection: a randomized clinical trial. *Jama* 2016;315(18):1956-65.
6. Does time of day influence cancer detection and recall rates in mammography? Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment; 2017. SPIE.
7. Blanks R, Given-Wilson R, Cohen S, et al. An analysis of 11.3 million screening tests examining the association between recall and cancer detection rates in the English NHS breast cancer screening programme. *European Radiology* 2019;29(7):3812-19.
8. Khan SA, Hernandez-Villafuerte KV, Muchadeyi MT, et al. Cost-effectiveness of risk-based breast cancer screening: A systematic review. *International journal of cancer* 2021;149(4):790-810.
9. Pharoah PD, Sewell B, Fitzsimmons D, et al. Cost effectiveness of the NHS breast screening programme: life table model. *Bmj* 2013;346
10. Morton R, Sayma M, Sura MS. Economic analysis of the breast cancer screening program used by the UK NHS: should the program be maintained? *Breast Cancer: Targets and Therapy* 2017;9:217.
11. Lawrence G, Wallis M, Allgood P, et al. Population estimates of survival in women with screen-detected and symptomatic breast cancer taking account of lead time and length bias. *Breast cancer research and treatment* 2009;116(1):179-85.
12. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *bmj* 2021;374
13. Clift AK, Dodwell D, Lord S, et al. The current status of risk-stratified breast screening. *British Journal of Cancer* 2021:1-18.
14. Dent T, Jbilou J, Rafi I, et al. Stratified cancer screening: the practicalities of implementation. *Public Health Genomics* 2013;16(3):94-99.
15. Lee CS, Bhargavan-Chatfield M, Burnside ES, et al. The National Mammography Database: Preliminary Data. *American Journal of Roentgenology* 2016;206(4):883-90. doi: 10.2214/AJR.15.14312
16. Lee CS, Parise C, Burleson J, et al. Assessing the Recall Rate for Screening Mammography: Comparing the Medicare Hospital Compare Dataset With the National Mammography Database. *American Journal of Roentgenology* 2018;211(1):127-32. doi: 10.2214/AJR.17.19229
17. Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* 2017;283(1):49-58. doi: 10.1148/radiol.2016161174

18. Mikkelsen EM, Njor SH, Vejborg I. Danish Quality Database for Mammography Screening. *Clin Epidemiol* 2016;8:661-66. doi: 10.2147/clep.S99467 [published Online First: 20161025]
19. Kang SY, Kim YS, Kim Z, et al. Basic Findings Regarding Breast Cancer in Korea in 2015: Data from a Breast Cancer Registry. *J Breast Cancer* 2018;21(1):1-10. doi: 10.4048/jbc.2018.21.1.1 [published Online First: 20180323]
20. Hong S, Song SY, Park B, et al. Effect of Digital Mammography for Breast Cancer Screening: A Comparative Study of More than 8 Million Korean Women. *Radiology* 2020;294(2):247-55. doi: 10.1148/radiol.2019190951 [published Online First: 20191203]
21. Results from the UK NHS breast screening programme 2000–05. *Journal of Medical Screening* 2007;14(4):200-04. doi: 10.1258/096914107782912068
22. Gathani T, Bull D, Green J, et al. Breast cancer histological classification: agreement between the Office for National Statistics and the National Health Service Breast Screening Programme. *Breast Cancer Res* 2005;7(6):R1090-R96. doi: 10.1186/bcr1352 [published Online First: 2005/11/09]
23. Elfström KM, Arnheim-Dahlström L, von Karsa L, et al. Cervical cancer screening in Europe: Quality assurance and organisation of programmes. *European Journal of Cancer* 2015;51(8):950-68. doi: https://doi.org/10.1016/j.ejca.2015.03.008
24. Lee TJW, Rutter MD, Blanks RG, et al. Colonoscopy quality measures: experience from the NHS Bowel Cancer Screening Programme. *Gut* 2012;61(7):1050-57. doi: 10.1136/gutjnl-2011-300651
25. Conti B, Bochaton A, Charreire H, et al. Influence of geographic access and socioeconomic characteristics on breast cancer outcomes: A systematic review. *PLoS One* 2022;17(7):e0271319. doi: 10.1371/journal.pone.0271319 [published Online First: 20220719]
26. Lundqvist A, Andersson E, Ahlberg I, et al. Socioeconomic inequalities in breast cancer incidence and mortality in Europe-a systematic review and meta-analysis. *Eur J Public Health* 2016;26(5):804-13. doi: 10.1093/eurpub/ckw070 [published Online First: 20160523]
27. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *European Journal of Cancer* 2009;45(5):747-55. doi: https://doi.org/10.1016/j.ejca.2008.11.032
28. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *European Journal of Cancer* 2009;45(5):756-64. doi: https://doi.org/10.1016/j.ejca.2008.11.033
29. Löfgren L, Eloranta S, Krawiec K, et al. Validation of data quality in the Swedish National Register for Breast Cancer. *BMC Public Health* 2019;19(1):495-95. doi: 10.1186/s12889-019-6846-6
30. Henson KE, Elliss-Brookes L, Coupland VH, et al. Data Resource Profile: National Cancer Registration Dataset in England. *Int J Epidemiol* 2020;49(1):16-16h. doi: 10.1093/ije/dyz076 [published Online First: 2019/05/24]
31. Sigurdardottir LG, Jonasson JG, Stefansdottir S, et al. Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness. *Acta oncologica* 2012;51(7):880-89.
32. Leinonen MK, Miettinen J, Heikkinen S, et al. Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. *Eur J Cancer* 2017;77:31-39. doi: 10.1016/j.ejca.2017.02.017 [published Online First: 2017/03/30]
33. Larsen IK, Småstuen M, Johannesen TB, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. *European journal of cancer* 2009;45(7):1218-31.
34. Dimitrova N, Parkin D. Data quality at the Bulgarian National Cancer Registry: An overview of comparability, completenss, validity and timeliness. *Cancer Epidemiology* 2015 doi: 10.1016/j.canep.2015.03.015
35. Ryzhov A, Bray F, Ferlay J, et al. Evaluation of data quality at the National Cancer Registry of Ukraine. *Cancer Epidemiology* 2018;53:156-65.
36. Fung JW, Lim SB, Zheng H, et al. Data quality at the Singapore Cancer Registry: An overview of comparability, completeness, validity and timeliness. *Cancer Epidemiol* 2016;43:76-86. doi: 10.1016/j.canep.2016.06.006 [published Online First: 2016/07/12]

37. Weir H, Sherman R, Yu M, et al. Cancer Incidence in Older Adults in the United States: Characteristics, Specificity, and Completeness of the Data. *Journal of registry management* 2020;47:150-60.
38. Lorez M, Bordoni A, Bouchardy C, et al. Evaluation of completeness of case ascertainment in Swiss cancer registration. *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP)* 2017;26 doi: 10.1097/CEJ.0000000000000380
39. Kearney T, Donnelly C, Kelly JM, et al. Validation of the completeness and accuracy of the Northern Ireland Cancer Registry. *Cancer epidemiology* 2015;39 doi: 10.1016/j.canep.2015.02.005
40. Hackl M, Waldhoer T. Estimation of completeness of case ascertainment of Austrian cancer incidence data using the flow method. *European Journal of Public Health* 2012;23(5):889-93. doi: 10.1093/eurpub/cks125
41. Donnelly C, Cairnduff V, Chen JJ, et al. The completeness and timeliness of cancer registration and the implications for measuring cancer burden. *Cancer Epidemiol* 2017;49:101-07. doi: 10.1016/j.canep.2017.05.007 [published Online First: 2017/06/11]
42. Ladha KS, Eikermann M. Codifying healthcare--big data and the issue of misclassification. *BMC Anesthesiol* 2015;15:179. doi: 10.1186/s12871-015-0165-y [published Online First: 20151215]
43. Raman SR, O'Brien EC, Hammill BG, et al. Evaluating fitness-for-use of electronic health records in pragmatic clinical trials: reported practices and recommendations. *J Am Med Inform Assoc* 2022;29(5):798-804. doi: 10.1093/jamia/ocac004
44. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)* 2016;4(1):1244. doi: 10.13063/2327-9214.1244 [published Online First: 20160911]
45. Huang JY. Representativeness Is Not Representative: Addressing Major Inferential Threats in the UK Biobank and Other Big Data Repositories. *Epidemiology* 2021;32(2):189-93. doi: 10.1097/ede.0000000000001317
46. Programmes NCS. Breast and cervical screening: the first 20 years, 2008:9.
47. Blanks RG, Wallis MG, Alison R, et al. Impact of Digital Mammography on Cancer Detection and Recall Rates: 11.3 Million Screening Episodes in the English National Health Service Breast Cancer Screening Program. *Radiology* 2019;290(3):629-37. doi: 10.1148/radiol.2018181426 [published Online First: 20181211]
48. Weigel S, Heindel W Fau - Heidinger O, Heidinger O Fau - Berkemeyer S, et al. Digital mammography screening: association between detection rate and nuclear grade of ductal carcinoma in situ. (1527-1315 (Electronic))
49. Vigeland E, Klaasen H Fau - Klingen TA, Klingen Ta Fau - Hofvind S, et al. Full-field digital mammography compared to screen film mammography in the prevalent round of a population-based screening programme: the Vestfold County Study. (0938-7994 (Print))
50. Shaaban AM, Hilton B, Clements K, et al. Pathological features of 11,337 patients with primary ductal carcinoma in situ (DCIS) and subsequent events: results from the UK Sloane Project. *Br J Cancer* 2021;124(5):1009-17. doi: 10.1038/s41416-020-01152-5 [published Online First: 20201117]
51. Group TUaASA. NHS Breast Screening Programme and Association of Breast Surgery - An Audit of Screen Detected Breast

Cancers for the Year of Screening April 2012 to March 2013, May 2014.
52. England N. Guidance for radiology and advanced radiographic practice in the NHS Breast Screening Programme, March 2011.
53. gov.uk. English indices of deprivation 2012 [updated 10/12/2020. Available from: https://www.gov.uk/government/collections/english-indices-of-deprivation accessed 18/05/2023 2023.
54. Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology* 2021;21(1):63. doi: 10.1186/s12874-021-01252-7

55. Parkin DM CV, Ferlay J, Galceran J, Storm HH, Whelan SL. Comparability and quality control in cancer registration1994.
56. Roderick J. A. Little DBR. Statistical analysis with missing data. : John Wiley & Sons 2019.
57. England N. About the Health Systems Support Framework [Available from: https://www.england.nhs.uk/hssf/background/ accessed 18/05/2023 2023.
58. Digital N. Breast Screening Programme - national statistics [updated 16/2/2023. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme accessed 18/05/2023 2023.
59. Digital N. Access data on the Personal Demographics Service. www.digital.nhs.uk, 2023.
60. Insitute NC. Surveillance, Epidemiology, and End Results Program [Available from: https://seer.cancer.gov/ accessed 18/05/2023 2023.
61. Reduction in breast cancer mortality from organized service screening with mammography: 1. Further confirmation with extended data. *Cancer Epidemiol Biomarkers Prev* 2006;15(1):45-51. doi: 10.1158/1055-9965.Epi-05-0349
62. Elmore JG, Lee CI. Data quality, data sharing, and moving artificial intelligence forward. *JAMA network open* 2021;4(8):e2119345-e45.

Appendix

A. Changes to the Breast Cancer screening programme and cancer classification over time

Screening programme

Between the start of the programme in 1988 and the end of the study period changes to the programme's operation were introduced.

- The NHS Cancer Plan (Department of Health, September 2000) included the intention to extend screening to cover the age range from 50 to 70. The programme extension was rolled out across the country between 2000 and 2006.
- Initially in most centres two-view (CC and MLO) mammography was only used for prevalent screens with MLO views only in the incident round. Following research in the late 1990s[1] the recommendation was changed and all screening units in England had extended two-view mammography for incident screens by December 2004.
- When the screening programme began there was mixed practice, with most centres only using a single film reader to examine each woman's mammograms and decide whether to recall her for further tests. Over time centres moved to two film readers independently examining each woman's mammograms (called double reading). By December 1996, 76% of units had introduced double reading, and it was standard practice by January 2006. Where the two readers disagreed whether to recall the woman, arbitration by a third reader or group of readers is used. In some centres in some time periods arbitration is used to determine whether to a recall a woman, even if the first two film readers both recommended recall. This is to keep recall rates, and the associated rates of false positive recalls low.
- The Cancer Reform Strategy (Department of Health, Cancer Reform Strategy, December 2007) announced the intention to extend the eligible age range by 6 years to include women aged 47-50th birthday and women in the age group 71-73rd birthday. In 2009 this 'Age Extension' was partially rolled out as a randomised controlled trial (RCT) in 65 screening centres. In these centres 50% of women were randomly assigned to receive extra screening. Of the remaining centres, 4 did not participate in the trial, 9 struggled with the randomisation and instead enrolled all, and 2 were closed and their eligible populations were distributed to others. The trial was piloted in 4 centres before the others joined between January 2010 and December 2014.
- Adoption of digital mammography to replace analogue machines was rolled out between 2009 and September 2015.
- The administrative systems receive their population in the form of batches specified from the population register. Pre-July 2016 these were specified on locally held instances of the National Health Application and Infrastructure Services (NHAIS) via Open Exeter. In 2016 Breast Screening Select was developed, which, although populated from the same source, provided a single national system to support the identification of the eligible population. This has enabled the implementation of restrictions and controls to improve the standardisation of cohort identification and national oversight of practice, reducing variability (see Table B1.1).

Cancer classification

Breast cancer can be classified according to type, stage, and grade to optimise treatment selection, since degree of malignancy is related to morphological appearance of tumours [2].

Type: There are two major histopathological types of breast cancer, carcinoma in situ (mainly DCIS) and invasive carcinoma. This distinction has remained constant throughout the study period.

Stage: Stage considers size, spread, lymph nodes, metastasis, and the statuses of receptors and development in diagnostics can contribute to improved accuracy. A score between 0 and 4 is use for the size of the tumour, whether the tumour has spread to the lymph nodes, and whether the tumour has metastasized. Staging can also consider oestrogen receptor status, progesterone receptor status and Her2 status. During the study period (from late 1999), node staging has likely become more accurate due to a change from axillary sampling/clearance to sentinel lymph node biopsy (SLNB). There is some retrospective evidence that SLNB picks up more, small metastases compared to axillary

clearance (Macaskill 2012) which may have had an impact on staging, though the extent is unknown. There is variation in how SLNB is undertaken resulting in varying sensitivity in finding micrometastases (<0.2mm).

Grade: The appearance of the cancer under the microscope considering differentiation of cells and speed of growth is used to as the basis for a number system of 1 (low grade) to 3 (high grade). The grading of invasive cancers has been uniformly reported since the early 1990s using the Nottingham method which itself remained constant[2]. Prior to that the grading system used scores from 1–4, so grade 4 cancers represent old cases. The Nottingham method has been showing moderate agreement, unchanged over time[3-6]. The system of differentiating DCIS into low, intermediate, and high grade is based on the potential for recurrence or progression to invasive cancer following treatment has been unchanged throughout the study period. Grading for DCIS has remained unchanged. However, coinciding with the change from analogue to digital film mammography a change in DCIS grading has been observed resulting in less low grade and more high-grade disease. Studies have reported only fair overall agreement for grade, with modest agreement in the high and low grade categories and poor agreement in the intermediate grade category[7].

B. Construction of ATHENA-M

1. Data sources – additional information

NBSS

Each centre had its own instance of the system and its own local database. Most widely used was the National Breast Screening System (NBSS) designed in the late 1980s based on a more generic Patient Administrative system known as the 'Oxford' system. Other systems were CAMRASS (South-West London, Surrey, West Sussex), Trent (East Midlands), Kodak (Staffordshire), and HSS (East Anglia and Lancashire). While providing similar types of administrative system, data structures differed slightly. Between 2004 and 2005 a new version of the NBSS system was rolled out as nationwide standard and has undergone few substantial changes since. Data from the previous systems were converted accordingly. Data from units that used the old NBSS system were more compatible with the new NBSS structure.

NBSS data was extracted screening centre by screening centre. The three extracts contained the following information:

- **NBSS-women:** Ethnicity (only populated at a subset of centres), month, year of birth, participation in relevant research trials, and issues with data quality of NHS number identifier
- **NBSS-episode:** Screening date, pseudonymised identifiers for the readers examining the mammograms, their decisions, whether the woman was recalled for further tests, and whether cancer was detected
- **NBSS-feature:** recall characteristics such as side of the body and mammographic characteristics such as mass or microcalcifications

Table B1.1 Systems prior to NBSS

Originally	Developers	Area covered	Became	Date converted to NBSS
Oxford System	Oxford Regional Computer System	2/3 of English programmes	Root system for NBSS	N/A
Trent	AT&T Istel and then McKesson	Current East Mids region	N/A	2004/05
CBSS	Healthcare Software Systems	East Anglia & some areas of the North-West	N/A	Data not available
CAMRASS	BM Computing	South West London	N/A	Data not available

Kodak	Kodak	Mid Staffordshire	N/A	2001/02

BS Select
Self-referral is where women request screening themselves, either because they are over the upper age limit to be invited to screening, or because they did not attend their screening appointment and contacted the breast screening centre to re-book more than 6 months after the original appointment date. General Practitioner referral is where the woman's GP refers her for mammography, either because she has started at a new GP practice and is eligible for screening, or historically symptomatic women were referred via this route. Higher risk screening is for women with a very high risk of developing breast cancer in comparison to the general population. It involves a younger age of initiation and may use tests other than mammography. Some women may later be moved from high-risk screening to the routine triennial call-recall screening. Non-routine early recall appointments are created when the woman is invited for further tests after shorter than the normal recall period.

Cancer registry
Details of all breast cancers C50* and DCIS D05* according to the ICD-10 system, or the pre-1995 equivalents of '174' and '2330', respectively, were included, inclusive of both screen-detected and symptomatically detected breast tumours. Data items include ICD classification, morphology, behaviour, grade, size, number of involved nodes, oestrogen, progesterone and HER2 status, Nottingham Prognostic Index, TNM stage, and whether screen detected. Treatment data items include breast surgery (breast conserving, mastectomy) underarm surgery (axillary clearance, sentinel lymph node biopsy), hormone therapy, radiotherapy, and chemotherapy.

IMD score
To accommodate revisions, the ATHENA-M dataset includes two alternative IMD-based scores. The first data point is the quintile of the income domain of IMD 2015 based on the woman's postcode at the time of her last screening appointment. The second data point is based on the same postcode but uses the version of the IMD current at the time of her last screening appointment: Example to illustrate use IMD 2019 if the last screen was 2014 or later, use IMD 2015 if the last screen was between 2010 and 2013, use IMD 2010 if the last screen was between 2007 and 2009, use IMD 2007 if the last screen was between 2003 and 2006, and use IMD 2004 if the last screen was before 2003. A third variable indicates which year's indices have been applied for each woman.

2. Inclusion criteria

Opting out
Women who opted out of their data being held on the National Disease Registration Service (NDRS) registers were already removed from the Cancer Registry data, so we have no record of breast cancer in these women. Under regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 (SI 2002/ 1438), data on women who have opted out via the national data opt out does not have to be removed from screening work defined as service improvement by Public Health England. We had intended to remove these women even though it is not a legal requirement but were unable to due to issues with identifying them.

3. Data linkage

The major task in the construction of the ATHENA-M dataset was to link NBSS with the other data sources (Figure 2 in main manuscript). Linkage of women between cancer registry and NBSS was based on NHS number, date of birth, postcode, and name, using a scoring system detailed in Tables B3.2-4 below. NBSS episodes with screen detected cancer were linked to cancer registry records if diagnosis date in the Cancer Registry was between 7 days before and 100 days after the screening 'date taken'. The remaining cancers in the Cancer Registry are considered to be symptomatically detected cancers. Linkage to PHE-MBIS death data was on NHS number alone, after tracing invalid NHS numbers. The vital status of women with invalid

non-traceable NHS numbers remains unknown. IMD quintile are based on last address in NBSS, which was also used for linkage. BS Select data is linked through NHS number and date of birth; either matches between any two of the day, month, year parts, or a match on year, with the day and month transposed.

Missing NHS numbers by year

The modern 10 digital NHS number was introduced in 1996. Prior to this date services had more latitude in the generation of local numbers. Data were matched through the tracing service to identify NHS numbers for those clients whose screening predated 1996. If the client had a subsequent interaction with the NHS such as an inpatient stay they would have subsequently had a new NHS number in the system and this would have been matched to their record. However, clients who died, moved outside of England, had their record flagged as sensitive or did not have need to access NHS services e.g., all healthcare was accessed via the independent sector would not be traceable.

The table below lists the counts of valid and invalid NHS numbers, year by year, before exclusions. The relative counts of the invalid ones, recorded as proportion missing below, indicates an overall missingness of 4.0% and also shows that the problem concerns primarily the early years of the screening programme before the modern 10-digit NHS number was introduced in 1996. While 42.9% are missing in the first year of the programme, the missingness drops fast, with 523,969 (28.2%) of women between 1988 and 1996, but only 10,187 (0.1%) between 1997 and 2018. In the last decade (2009 to 2018) only 2,352 women (0.025%) (cumulative invalid NHS number column in Table B3.1 in those years) had an invalid NHS number.

Table B3.1 Missing and invalid NHS numbers by year

Year of last invite	Valid NHS Number	Invalid NHS Number	Total for year	Proportion missing
1988	7438	5581	13019	42.9%
1989	45807	29884	75691	39.5%
1990	118129	74570	192699	38.7%
1991	154167	100985	255152	39.6%
1992	175284	101345	276629	36.6%
1993	197024	91770	288794	31.8%
1994	213693	66299	279992	23.7%
1995	208796	38497	247293	15.6%
1996	216505	15038	231543	6.5%
1997	208240	1079	209319	0.5%
1998	172498	650	173148	0.4%
1999	109896	726	110622	0.7%
2000	82853	687	83540	0.8%
2001	73598	764	74362	1.0%
2002	75061	763	75824	1.0%
2003	116017	690	116707	0.6%
2004	154097	718	154815	0.5%
2005	241524	686	242210	0.3%
2006	265589	522	266111	0.2%
2007	265973	410	266383	0.2%
2008	240228	140	240368	0.1%
2009	204148	248	204396	0.1%
2010	216133	348	216481	0.2%
2011	250937	247	251184	0.1%

2012	280935	204	281139	0.1%
2013	298769	405	299174	0.1%
2014	327830	316	328146	0.1%
2015	1741661	257	1741918	0.01%
2016	2645059	327	2645386	0.01%
2017	2711002	0	2711002	0.0%
2018	707085	0	707085	0.0%
Total	**12725976**	**534156**	**13260132**	**4.0%**

Data linkage between NBSS and cancer registry: scoring system Percent Invalid NHS Number

To link NBSS with cancer registry records, each identifier was given a score from 0 (no match) to 3 (exact match) as summarised in the table below. NHS number and date of birth alone would give a score of 6, this is a standard level of matching. Therefore, we included matches of score 6 if NHS number and date of birth matched exactly and matches of score 7 or more regardless of whether NHS number and date of birth were exact matches. Matches of score 5 or lower and score 6 without exact matches for both NHS number and date of birth have been excluded as unreliable.

Table B3.2 Scoring system for linkage between NBSS and cancer registry

Weighted value	**3**	**2**	**1**	**0**
NHS number	NHSnumberbest equals CAS NHSnumber and neither are null		NHS number exists in one dataset only, indicating a lack of complete data in the other	NHSnumberbest does not equal CAS NHSnumber or both are null
Date of birth	Exact match	2 out of 3 date parts match	Year of birth matches, and one dataset has 01/01/yyyy as date, indicating possibly unknown exact date	Any other non-match or null
Postcode	Complete match, when formatted as 7 characters		First 4 characters match	Any other non-match or null
Name	Exact match of forename and either current or previous surname	Transposition of forename and surname	First 3 characters of forename and first 3 characters of surname match	Any other non-match or null

Matching results NBSS and cancer registry

The table below shows the number of person matches for each combination of scores across the four identifiers, after NHS number tracing. Common reasons for imperfect matches include different spelling of name, same data but in different order (e.g., surname and forename switched), change of name or address (which is plausible because date of screening and date of symptomatic cancer detection may be years or decades apart). Of the women successfully matched, 91% score eight or more.

Table B3.3 Score combinations occurring in linking NBSS and cancer registry

NHS number	Date of birth	Postcode	Name	Score	Person count
3	3	3	3	12	1,327,016
3	2	3	3	11	10,728
3	3	3	2	11	252
3	3	1	3	10	138,399
3	3	3	1	10	946,327
3	2	1	3	9	1,829
3	2	3	1	9	8,765
3	3	0	3	9	207,759
3	3	1	2	9	31
3	3	3	0	9	70,342
0	2	3	3	8	91
3	2	0	3	8	3,209
3	2	3	0	8	771
3	3	0	2	8	72
3	3	1	1	8	107,949
3	3	0	1	7	179,761
3	3	1	0	7	31,268
3	3	0	0	6	65,098

Table B3.4 relates to counts of fully registered breast tumours, diagnosed between 1971 and 2018, for patients linked from the cancer registry to women in NBSS included in ATHENA-M. It gives matches by score and year. Of the breast tumours successfully matched, 94% have a person match score of eight or more.

Table B3.4 Counts of fully-registered breast tumours

Person match score	Tumour count	% of total tumour count
12	376,272	44.93%
11	2,856	0.34%
10	306,226	36.57%
9	73,416	8.77%
8	28,061	3.35%
7	43,438	5.19%
6	7,200	0.86%

Table B3.5 Match scores relating to fully registered breast tumours for patients linked from the cancer registry to women in NBSS by year

Year of diagnosis	Person match score							Total tumours	% of Total
	6	7	8	9	10	11	12		
1971		1	4	5	13	1	21	**45**	0.01%
1972	2	2	4	7	28		25	**68**	0.01%
1973		6	5	10	20	1	31	**73**	0.01%
1974		3	3	7	35		33	**81**	0.01%
1975	1	7	4	15	42	1	41	**111**	0.01%
1976		8	4	7	52	2	45	**118**	0.01%
1977	1	10	1	19	56	2	66	**155**	0.02%
1978	2	9	4	19	58		64	**156**	0.02%
1979	6	13	12	13	96	2	88	**230**	0.03%
1980	2	12	14	29	85	2	104	**248**	0.03%
1981		13	11	41	111	2	141	**319**	0.04%
1982	3	14	12	37	122	4	159	**351**	0.04%
1983	4	19	17	38	160	2	186	**426**	0.05%
1984	2	21	17	53	173	1	177	**444**	0.05%
1985	5	20	14	66	228	4	232	**569**	0.07%
1986	3	24	22	72	213	7	257	**598**	0.07%
1987	4	26	32	59	268	5	291	**685**	0.08%
1988	141	522	359	775	1,979	41	1,772	**5,589**	0.67%
1989	147	680	418	1,028	2,437	67	2,307	**7,084**	0.85%
1990	177	801	507	1,188	3,069	69	3,245	**9,056**	1.08%
1991	200	875	591	1,506	3,978	96	4,071	**11,317**	1.35%
1992	208	1,008	737	1,648	4,552	92	4,612	**12,857**	1.54%
1993	218	1,048	675	1,703	4,771	87	4,854	**13,356**	1.59%
1994	241	1,126	847	1,718	5,152	77	5,373	**14,534**	1.74%
1995	264	1,136	832	1,931	5,425	115	5,755	**15,458**	1.85%
1996	292	1,288	877	2,030	6,028	100	6,396	**17,011**	2.03%
1997	352	1,417	992	2,275	6,966	129	7,356	**19,487**	2.33%
1998	306	1,449	998	2,251	7,836	127	8,321	**21,288**	2.54%
1999	290	1,537	1,102	2,297	8,561	131	9,105	**23,023**	2.75%
2000	292	1,607	1,189	2,289	8,897	135	9,529	**23,938**	2.86%
2001	293	1,655	1,158	2,366	9,372	132	10,107	**25,083**	3.00%
2002	289	1,599	1,163	2,450	9,896	154	10,749	**26,300**	3.14%
2003	298	1,687	1,149	2,609	10,849	128	11,772	**28,492**	3.40%
2004	276	1,795	1,218	2,621	11,113	135	12,758	**29,916**	3.57%
2005	281	1,830	1,187	2,746	11,504	149	13,932	**31,629**	3.78%
2006	295	1,769	1,116	2,780	11,710	151	14,534	**32,355**	3.86%
2007	266	1,665	1,082	2,951	11,699	96	15,187	**32,946**	3.93%

2008	246	1,688	1,078	2,990	12,035	90	17,110	**35,237**	4.21%
2009	241	1,651	963	3,119	11,604	56	17,687	**35,321**	4.22%
2010	259	1,628	949	3,125	12,059	66	18,342	**36,428**	4.35%
2011	229	1,528	855	3,271	11,787	60	19,522	**37,252**	4.45%
2012	191	1,698	1,130	2,989	14,741	64	17,928	**38,741**	4.63%
2013	200	1,684	938	3,272	14,731	51	20,010	**40,886**	4.88%
2014	163	1,756	1,039	3,110	16,491	56	19,894	**42,509**	5.08%
2015	164	1,562	869	2,832	16,318	41	20,035	**41,821**	4.99%
2016	141	1,317	750	2,631	16,067	48	20,349	**41,303**	4.93%
2017	106	1,198	599	2,355	16,232	39	20,677	**41,206**	4.92%
2018	99	1,026	514	2,063	16,607	38	21,022	**41,369**	4.94%
Total tumours	**7,200**	**43,438**	**28,061**	**73,416**	**306,226**	**2,856**	**376,272**	**837,469**	

4. Exclusions

Centres

Three out of 79 centres had to be excluded from the analysis since examination of patterns of screening attendance highlighted systematic issues with data extraction.

Centre A: There were technical issues with running the crystal report allowing only the extraction of NBSS-episode data; it was only possible to extract data for the first 30,000 women screened at that centre.

Centre B: There were large numbers (75%) of women in the Table NBSS-episode not linked to a woman's records within that centre in Table NBSS-women. 64% could not be linked to any centre in Table NBSS-women.

Centre C: 35% of women did not have any associated screening episode.

C. Data quality pillars

Table C1: Number and percent of missing variables from the cancer registry by year group highlighting variables with high missingness Pre 1997 (yellow) and Overall (red)

Variable	Pre 1997	1997 and after	Overall
Diagnosis date	0 (0%)	0 (0%)	0 (0%)
Basis of diagnosis of the tumour	0 (0%)	0 (0%)	0 (0%)
Basis of diagnosis of the tumour text	0 (0%)	0 (0%)	0 (0%)
Site of neoplasm (4-character ICD-10-O2 code)	32,579 (31.03%)	0 (0%)	32,579 (3.95%)
Site of the cancer and text description	0 (0%)	0 (0%)	0 (0%)
Site of the cancer and text description test	0 (0%)	0 (0%)	0 (0%)
Morphology of cancer, original coding system	1 (0%)	0 (0%)	1 (0%)
Morphology of cancer, in the ICD-10-O2 system	32,579 (31.03%)	0 (0%)	32,579 (3.95%)
Behaviour of cancer, in the ICD-10-O2 system	32,579 (31.03%)	0 (0%)	32,579 (3.95%)
Numeric behaviour code of cancer and description	1 (0%)	0 (0%)	1 (0%)
Behaviour code of cancer, text	1 (0%)	0 (0%)	1 (0%)
Grade of tumour	0 (0%)	0 (0%)	0 (0%)
Size of the largest dimension of tumour	64,421 (61.37%)	213,515 (29.66%)	277,936 (33.7%)
Number of nodes excised	85,569 (81.51%)	298,275 (41.44%)	383,844 (46.54%)
Number of nodes involved	90,864 (86.55%)	333,739 (46.36%)	424,603 (51.48%)
Laterality	0 (0%)	0 (0%)	0 (0%)
Multifocal Tumour	89,607 (85.36%)	550,818 (76.52%)	640,425 (77.65%)
Oestrogen receptor status of tumour	96,757 (92.17%)	315,666 (43.85%)	412,423 (50%)
Oestrogen receptor score of tumour	104,843 (99.87%)	610783 (84.85%)	715,626 (86.76%)
Progesterone receptor status of tumour	104,262 (99.32%)	527,924 (73.34%)	632,186 (76.65%)
Progesterone receptor score of tumour	104900 (99.92%)	670,742 (93.18%)	775,642 (94.04%)
Human Epidermal Growth Factor Receptor 2 (HER2) status of tumour	97,867 (93.22%)	372,355 (51.73%)	470,222 (57.01%)
Nottingham Prognostic Index Score	96,116 (91.56%)	362,686 (50.38%)	458,802 (55.63%)
T stage flagged by the registry as ‘best’ T stage	78,821 (75.08%)	269,037 (37.38%)	347,858 (42.17%)
N stage flagged by the registry as ‘best’ N stage	80,561 (76.74%)	291,613 (40.51%)	372,174 (45.12%)
M stage flagged by the registry as ‘best’ M stage	87,290 (83.15%)	418,559 (58.15%)	505,849 (61.33%)

Variable	Pre 1997	1997 and after	Overall
Best 'registry' stage at diagnosis of the tumour	25,047 (23.86%)	159,641 (22.18%)	184,688 (22.39%)

Characteristics of excluded centres

Three centres were excluded from the study upfront due to a variety of irregularities (see section B4). Table C2 provides a detailed list of screening centre performance comparing each of the excluded centres with the other ones, separately for each year in the study period. For most variables there is no obvious evidence that the excluded centres are systematically different, but the IMD in these centres is unusually high making comparison of socio-economic status with the other centres impossible.

Table C2.1: Comparison of excluded centres versus other centres for the whole study period: screening related variables

	Year Group	Other centres	A	B	C
Number of women screened per year 7	1988-1992	29,851	21,717	31,789	16
	1993-1997	56,359	29,830	19,686	42,733
	1998-2002	62,550	23,555	12,131	58,251
	2003-2007	82,520	26,769	10,339	56,725
	2008-2013	99,891	25,072	3,087	99,497
	2014-2018	119,989	20,400	912	98,076
Percent (rank) of women screened recalled for further tests	1988-1992	6.08 (42)	7.19 (22)	8.31 (14)	6.25 (33)
	1993-1997	4.66 (40)	3.27 (73)	5.29 (25)	3.51 (65)
	1998-2002	4.89 (40)	4.94 (43)	4.59 (50)	3.91 (66)
	2003-2007	4.56 (40.5)	5.22 (24)	4.53 (43)	3.61 (65)
	2008-2013	4.01 (39)	3.53 (57)	4.83 (12)	3.65 (52)
	2014-2018	3.92 (39)	2.23 (78)	4.39 (26)	3.66 (52)
Percent (rank) of women screened with a screen detected cancer	1988-1992	0.58 (40)	0.61 (30)	0.51 (58)	0 (80)
	1993-1997	0.49 (42)	0.52 (19)	0.7 (2)	0.56 (15)
	1998-2002	0.58 (42)	0.83 (2)	0.63 (22)	0.57 (53)
	2003-2007	0.77 (41.5)	0.95 (2)	1.13 (1)	0.7 (69)
	2008-2013	0.77 (40)	0.85 (9)	1.26 (1)	0.68 (72)
	2014-2018	0.81 (41)	0.87 (13)	1.32 (1)	0.81 (40)
Percent (rank) of women screened with false positive recalls with benign biopsies	1988-1992	0.64 (40)	0.78 (24)	0.53 (41)	0 (80)
	1993-1997	0.65 (41)	0.38 (60)	0.7 (31)	0.73 (29)
	1998-2002	0.89 (41)	0.94 (37)	0.8 (53)	0.85 (46)
	2003-2007	0.97 (40.5)	0.95 (44)	1.01 (36)	0.8 (57)
	2008-2013	0.96 (39)	0.71 (63)	1.13 (17)	0.73 (59)
	2014-2018	0.99 (38)	0.39 (78)	0.88 (44)	0.84 (49)

Table C2.2: Comparison of excluded centres versus other centres for the whole study period: age

	Year Group	Other centres	A	B	C
AgeCalc	1988-1992	57.0 (53.0, 60.0)	57.0 (53.0, 61.0)	56.0 (53.0, 60.0)	51.0 (49.8, 53.0)
Unknown		4,646	0	0	0
AgeCalc	1993-1997	56.0 (52.0, 60.0)	59.0 (55.0, 62.0)	59.0 (56.0, 62.0)	56.0 (52.0, 60.0)
Unknown		1,529	0	0	9
AgeCalc	1998-2002	56.0 (52.0, 60.0)	60.0 (56.0, 63.0)	61.0 (58.0, 63.0)	56.0 (52.0, 60.0)
Unknown		351	0	0	4
AgeCalc	2003-2007	58.0 (54.0, 62.0)	63.0 (58.0, 66.0)	66.0 (62.0, 68.0)	58.0 (55.0, 62.0)
Unknown		241	0	1	2
AgeCalc	2008-2013	59.0 (54.0, 64.0)	64.0 (59.0, 68.0)	67.0 (60.0, 70.0)	59.0 (54.0, 63.0)
Unknown		68	1	0	2
AgeCalc	2014-2018	59 (53, 65)	64 (59, 68)	65 (63, 70)	60 (57, 65)
Unknown		73	0	0	0

Table C2.3: Comparison of excluded centres versus other centres for the whole study period: IMD

IMDQUINTILE_2015ONLY	Year Group	Other centres	A	B	C
1 - Least deprived	1988-1992	482,923 (19.9%)	171 (0.8%)	2,316 (7.3%)	0 (0.0%)
2		515,550 (21.3%)	248 (1.1%)	2,439 (7.7%)	0 (0.0%)
3		496,899 (20.5%)	210 (1.0%)	1,627 (5.1%)	0 (0.0%)
4		461,879 (19.1%)	95 (0.4%)	810 (2.5%)	3 (18.8%)
5 - Most deprived		421,093 (17.4%)	36 (0.2%)	213 (0.7%)	0 (0.0%)
Missing		45,651 (1.9%)	20,957 (96.5%)	24,384 (76.7%)	13 (81.2%)
1 - Least deprived	1993-1997	1,049,672 (22.0%)	810 (2.7%)	2,137 (10.9%)	1,720 (4.0%)
2		1,110,939 (23.3%)	1,120 (3.8%)	2,150 (10.9%)	2,647 (6.2%)
3		999,161 (20.9%)	923 (3.1%)	1,432 (7.3%)	4,518 (10.6%)
4		858,001 (18.0%)	444 (1.5%)	713 (3.6%)	10,393 (24.3%)
5 - Most deprived		723,968 (15.2%)	169 (0.6%)	186 (0.9%)	12,570 (29.4%)
Missing		30,581 (0.6%)	26,364 (88.4%)	13,068 (66.4%)	10,885 (25.5%)
1 - Least deprived	1998-2002	1,274,827 (23.0%)	2,180 (9.3%)	1,622 (13.4%)	2,368 (4.1%)
2		1,332,074 (24.0%)	2,696 (11.4%)	1,600 (13.2%)	3,349 (5.7%)
3		1,167,344 (21.0%)	2,160 (9.2%)	1,122 (9.2%)	4,978 (8.5%)
4		967,728 (17.4%)	1,157 (4.9%)	560 (4.6%)	9,988 (17.1%)
5 - Most deprived		770,292 (13.9%)	569 (2.4%)	144 (1.2%)	10,816 (18.6%)
Missing		34,345 (0.6%)	14,793 (62.8%)	7,083 (58.4%)	26,752 (45.9%)

1 - Least deprived	2003-2007	1,623,024 (23.4%)	4,040 (15.1%)	1,041 (10.1%)	1,595 (2.8%)
2		1,680,148 (24.3%)	4,534 (16.9%)	1,035 (10.0%)	2,312 (4.1%)
3		1,459,279 (21.1%)	3,400 (12.7%)	686 (6.6%)	3,510 (6.2%)
4		1,185,997 (17.1%)	1,837 (6.9%)	339 (3.3%)	7,032 (12.4%)
5 - Most deprived		930,447 (13.4%)	864 (3.2%)	96 (0.9%)	7,440 (13.1%)
Missing		49,463 (0.7%)	12,094 (45.2%)	7,142 (69.1%)	34,836 (61.4%)
1 - Least deprived	2008-2013	1,934,573 (23.2%)	4,696 (18.7%)	419 (13.6%)	1,664 (1.7%)
2		2,003,228 (24.0%)	4,960 (19.8%)	375 (12.1%)	2,348 (2.4%)
3		1,751,587 (21.0%)	3,986 (15.9%)	278 (9.0%)	3,507 (3.5%)
4		1,439,459 (17.3%)	2,378 (9.5%)	142 (4.6%)	6,090 (6.1%)
5 - Most deprived		1,133,671 (13.6%)	1,158 (4.6%)	39 (1.3%)	6,248 (6.3%)
Missing		75,190 (0.9%)	7,894 (31.5%)	1,834 (59.4%)	79,640 (80.0%)
1 - Least deprived	2014-2018	2,327,340 (23.3%)	5,038 (24.7%)	130 (14.3%)	730 (0.7%)
2		2,353,817 (23.6%)	4,954 (24.3%)	127 (13.9%)	940 (1.0%)
3		2,069,474 (20.8%)	4,068 (19.9%)	98 (10.7%)	1,366 (1.4%)
4		1,731,308 (17.4%)	2,296 (11.3%)	38 (4.2%)	2,249 (2.3%)
5 - Most deprived		1,389,365 (13.9%)	1,090 (5.3%)	17 (1.9%)	1,884 (1.9%)
Missing		100,807 (1.0%)	2,954 (14.5%)	502 (55.0%)	90,907 (92.7%)

Reasons for mammography invitations

Invitations to mammography can be issued for a variety of reasons. The most frequent one is a routine appointment invitation. In addition, there are self-referrals, non-routine (early recalls), GP referrals, and higher risk referrals. Table C3 quantifies this based on data from BS Select for each year in the study period. The table has been populated from two sources, NBSS for routine appointments and BS Select for the other ones. In some instances, the date of the first screening appointment was not available in BS Select. Then, the screening date of the reported screen was used where available and the date the screen was taken otherwise (they are nearly always the same as the screening date with exceptions including recall and a few even less frequent issues). As the table only records year there are some instances where recording appears in the subsequent year, but as this would maximally shift by a year the overall picture, we can gain from this table will not change.

The total number of screening appointment recorded between NBSS (after exclusions) and BS Select is 54,426,307. The vast majority of these took place as part of the routine programme (95.1%), though there has been some variation over time. It drops from an initial 100% in the early years to just below 95% in 1997 and then further to under 93% in the early 2000s until it picks up again stabilising around 95%. This dynamic is large due to increased self-referrals and GP referrals during some time periods. Overall, self-referrals accounted for 3.8% of the total and GP referrals for 0.7%, while all other referral types were rare.

Table C3: First Offered Year by episode type by frequency and percent (referral dataset BS Select)

First Offered Year	**Overall**, N=54,426,307	**Routine**, N=51,759,542	**Self-referral**, N=2,074,471	**Non-routine (early) recall**, N=164,272	**GP referral**, N=419,625	**Higher risk,** N=8,397
1988	34,233	34,224 (100.0%)	3 (0.0%)	0 (0.0%)	6 (0.0%)	0 (0.0%)
1989	252,920	252,872 (100.0%)	15 (0.0%)	24 (0.0%)	9 (0.0%)	0 (0.0%)
1990	733,986	733,721 (100.0%)	163 (0.0%)	71 (0.0%)	31 (0.0%)	0 (0.0%)

First Offered Year	**Overall**, N=54,426,307	**Routine**, N=51,759,542	**Self-referral**, N=2,074,471	**Non-routine (early) recall**, N=164,272	**GP referral**, N=419,625	**Higher risk,** N=8,397
1991	1,112,726	1,112,106 (99.9%)	107 (0.0%)	398 (0.0%)	115 (0.0%)	0 (0.0%)
1992	1,220,784	1,217,046 (99.7%)	187 (0.0%)	1,600 (0.1%)	1,951 (0.2%)	0 (0.0%)
1993	1,273,217	1,270,808 (99.8%)	632 (0.0%)	1,577 (0.1%)	200 (0.0%)	0 (0.0%)
1994	1,240,325	1,233,055 (99.4%)	3,181 (0.3%)	2,712 (0.2%)	1,377 (0.1%)	0 (0.0%)
1995	1,272,802	1,241,403 (97.5%)	15,745 (1.2%)	8,790 (0.7%)	6,864 (0.5%)	0 (0.0%)
1996	1,358,934	1,301,211 (95.8%)	32,803 (2.4%)	11,936 (0.9%)	12,984 (1.0%)	0 (0.0%)
1997	1,422,781	1,350,039 (94.9%)	45,094 (3.2%)	10,706 (0.8%)	16,942 (1.2%)	0 (0.0%)
1998	1,469,031	1,373,901 (93.5%)	63,133 (4.3%)	11,031 (0.8%)	20,966 (1.4%)	0 (0.0%)
1999	1,576,596	1,479,369 (93.8%)	68,987 (4.4%)	6,059 (0.4%)	22,181 (1.4%)	0 (0.0%)
2000	1,593,562	1,479,390 (92.8%)	85,533 (5.4%)	3,382 (0.2%)	25,257 (1.6%)	0 (0.0%)
2001	1,598,552	1,483,338 (92.8%)	93,607 (5.9%)	3,109 (0.2%)	18,498 (1.2%)	0 (0.0%)
2002	1,632,886	1,517,078 (92.9%)	96,650 (5.9%)	2,709 (0.2%)	16,449 (1.0%)	0 (0.0%)
2003	1,726,737	1,612,566 (93.4%)	96,842 (5.6%)	2,368 (0.1%)	14,961 (0.9%)	0 (0.0%)
2004	1,824,895	1,709,622 (93.7%)	99,267 (5.4%)	2,210 (0.1%)	13,796 (0.8%)	0 (0.0%)
2005	1,976,224	1,887,993 (95.5%)	72,697 (3.7%)	2,368 (0.1%)	13,166 (0.7%)	0 (0.0%)
2006	2,082,523	1,993,639 (95.7%)	72,273 (3.5%)	2,712 (0.1%)	13,899 (0.7%)	0 (0.0%)
2007	2,166,653	2,078,793 (95.9%)	72,803 (3.4%)	2,122 (0.1%)	12,935 (0.6%)	0 (0.0%)
2008	2,235,557	2,138,181 (95.6%)	84,840 (3.8%)	1,908 (0.1%)	10,628 (0.5%)	0 (0.0%)
2009	2,283,971	2,173,667 (95.2%)	97,612 (4.3%)	1,540 (0.1%)	11,152 (0.5%)	0 (0.0%)
2010	2,362,968	2,259,370 (95.6%)	92,222 (3.9%)	1,274 (0.1%)	10,090 (0.4%)	12 (0.0%)
2011	2,466,992	2,358,941 (95.6%)	94,688 (3.8%)	1,116 (0.0%)	12,224 (0.5%)	23 (0.0%)
2012	2,575,719	2,474,237 (96.1%)	88,723 (3.4%)	1,127 (0.0%)	11,603 (0.5%)	29 (0.0%)
2013	2,648,553	2,546,400 (96.1%)	89,452 (3.4%)	996 (0.0%)	11,501 (0.4%)	204 (0.0%)
2014	2,764,728	2,623,151 (94.9%)	128,108 (4.6%)	990 (0.0%)	11,650 (0.4%)	829 (0.0%)
2015	2,765,574	2,639,635 (95.4%)	114,591 (4.1%)	928 (0.0%)	8,791 (0.3%)	1,629 (0.1%)
2016	2,859,613	2,745,495 (96.0%)	103,614 (3.6%)	874 (0.0%)	7,455 (0.3%)	2,175 (0.1%)
2017	2,858,466	2,735,886 (95.7%)	112,017 (3.9%)	690 (0.0%)	7,354 (0.3%)	2,519 (0.1%)
2018	732,762	702,405 (95.9%)	27,651 (3.8%)	169 (0.0%)	1,840 (0.3%)	697 (0.1%)
Unknown	301,037	0	121,231	76,776	102,750	280

Non-attendance at mammography invitations

Table C4 shows the distribution of women in the screening dataset by the number of non-attended invitations. The majority of women (53.7%) attended all the appointments they were invited to, while 32.5% did not attend one or two, and only 13.8% % did not attend more than two.

Table C4: Non-attended invitations

	N = 13,094,122

Non-attended invites	Number (%) women
0	7,033,660 (53.7%)
1	2,897,341 (22.1%)
2	1,361,688 (10.4%)
3	722,097 (5.5%)
4	424,816 (3.2%)
5+	654,520 (5.0%)

Table C5 and Figure C6 show the level of attendance at the second appointment only but follows this over the course of the study period. Attendance within expected timeframe was at 58.3% and 56.1% in the first two years of the programme but steadily increased during the 1990s until it plateaued around 78-80% from 1997 onwards and then slightly increased to just above 81.9% in 2007 to stay just above 80%. Note that the table ends in 2014 as information about the attendance at the second screening appointment for later years was not yet available at the time of the data extraction in 2018. For the same reason the last few years of the columns recording attendance as outside expected timeframe, and none are still subject to change; some of the non-attenders will eventually catch up and would move to attendance outside expected timeframe in future data extractions.

Table C5: Attendance at second screening over the time of the study period

	1st screening			2nd screening
Year	**Age Median (IQR) at screening**	**Attendance within expected timeframe**	**Attendance outside expected timeframe**	**No attendance**
1988 N = 24,171	57.00 (53.00, 61.00)	14,212 (58.8%)	3,100 (12.8%)	6,859 (28.4%)
1989 N = 179,587	57.00 (53.00, 61.00)	100,961 (56.2%)	20,775 (11.6%)	57,851 (32.2%)
1990 N = 521,646	57.00 (53.00, 61.00)	331,971 (63.6%)	42,871 (8.2%)	146,804 (28.1%)
1991, N = 757,627	57.00 (53.00, 60.00)	512,225 (67.6%)	58,629 (7.7%)	186,773 (24.7%)
1992 N = 756,089	56.00 (52.00, 60.00)	524,412 (69.4%)	54,710 (7.2%)	176,967 (23.4%)
1993 N = 511,218	54.00 (51.00, 59.00)	364,541 (71.3%)	44,402 (8.7%)	102,275 (20.0%)
1994 N = 298,302	51.00 (50.00, 55.00)	226,552 (75.9%)	34,382 (11.5%)	37,368 (12.5%)
1995 N = 259,285	51.00 (50.00, 52.00)	204,219 (78.8%)	33,000 (12.7%)	22,066 (8.5%)
1996 N = 283,309	51.00 (50.00, 52.00)	227,810 (80.4%)	36,258 (12.8%)	19,241 (6.8%)
1997	51.00 (50.00, 52.00)	239,576 (82.5%)	35,128 (12.1%)	15,591 (5.4%)

	1st screening			2nd screening
Year	**Age Median (IQR) at screening**	**Attendance within expected timeframe**	**Attendance outside expected timeframe**	**No attendance**
N = 290,295				
1998 N = 277,639	51.00 (50.00, 51.00)	229,681 (82.7%)	33,645 (12.1%)	14,313 (5.2%)
1999 N = 287,772	51.00 (50.00, 52.00)	235,151 (81.7%)	39,500 (13.7%)	13,121 (4.6%)
2000 N = 273,364	51.00 (50.00, 52.00)	223,065 (81.6%)	38,921 (14.2%)	11,378 (4.2%)
2001 N = 245,911	51.00 (50.00, 51.00)	199,419 (81.1%)	35,433 (14.4%)	11,059 (4.5%)
2002 N = 232,070	51.00 (50.00, 51.00)	188,994 (81.4%)	32,324 (13.9%)	10,752 (4.6%)
2003 N = 233,152	51.00 (50.00, 51.00)	190,606 (81.8%)	31,414 (13.5%)	11,132 (4.8%)
2004 N = 233,418	51.00 (50.00, 51.00)	191,010 (81.8%)	29,932 (12.8%)	12,476 (5.3%)
2005 N = 231,216	51.00 (50.00, 51.00)	190,386 (82.3%)	27,894 (12.1%)	12,936 (5.6%)
2006 N = 238,605	51.00 (50.00, 51.00)	198,203 (83.1%)	26,920 (11.3%)	13,482 (5.7%)
2007 N = 243,689	51.00 (50.00, 51.00)	207,014 (85.0%)	23,088 (9.5%)	13,587 (5.6%)
2008 N = 256,260	51.00 (50.00, 51.00)	217,215 (84.8%)	24,302 (9.5%)	14,743 (5.8%)
2009 N = 265,164	51.00 (50.00, 51.00)	222,969 (84.1%)	24,861 (9.4%)	17,334 (6.5%)
2010 N = 283,128	50.00 (50.00, 51.00)	237,276 (83.8%)	24,392 (8.6%)	21,460 (7.6%)
2011 N = 324,848	50.00 (49.00, 51.00)	270,355 (83.2%)	27,350 (8.4%)	27,143 (8.4%)
2012 N = 346,664	50.00 (49.00, 51.00)	291,239 (84.0%)	15,550 (4.5%)	39,875 (11.5%)
2013 N = 357,031	50.00 (48.00, 51.00)	299,165 (83.8%)	3,974 (1.1%)	53,892 (15.1%)
2014 N = 324,643	50.00 (48.00, 51.00)	270,762 (83.4%)	1,362 (0.4%)	52,519 (16.2%)

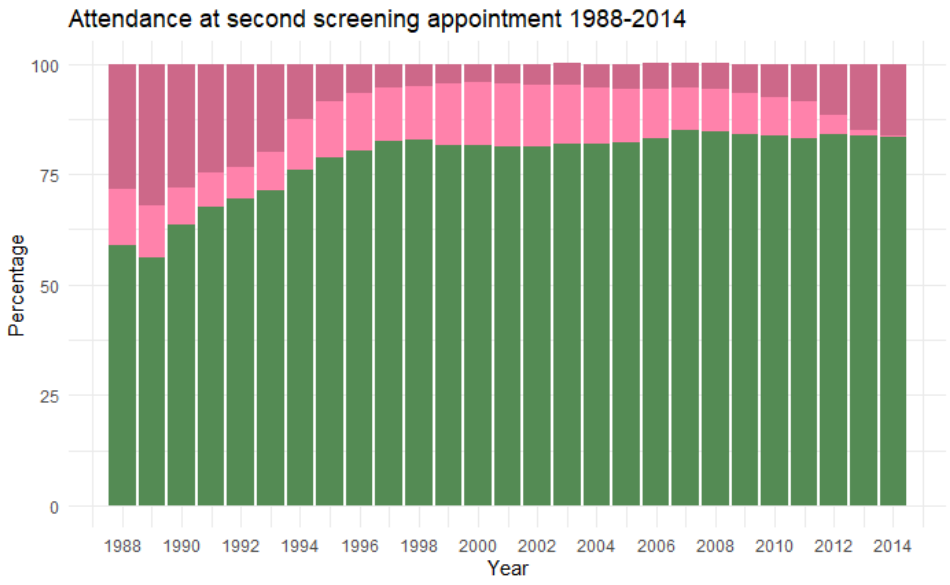

Figure C6. Percentage of women attending their second screening appointment within (green) and outside (light pink) the expected timeframe and not attending it (dark pink) based on available records (relative proportion between outside expected timeframe and no attendance subject to change in later years in future data extraction)

Appendix references:

1. Blanks R, Given-Wilson R, Moss S. Efficiency of cancer detection during routine repeat (incident) mammographic screening: two versus one view mammography. *Journal of Medical Screening* 1998;5(3):141-45.
2. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991;19(5):403-10.
3. Robbins P, Pinder S, De Klerk N, et al. Histological grading of breast carcinomas: a study of interobserver agreement. *Human pathology* 1995;26(8):873-79.
4. Meyer JS, Alvarez C, Milikowski C, et al. Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern Pathology* 2005;18(8):1067-78. doi: 10.1038/modpathol.3800388
5. Ellis IO, Coleman D, Wells C, et al. Impact of a national external quality assessment scheme for breast pathology in the UK. *Journal of Clinical Pathology* 2006;59(2):138-45. doi: 10.1136/jcp.2004.025551
6. Fanshawe TR, Lynch AG, Ellis IO, et al. Assessing agreement between multiple raters with missing rating information, applied to breast cancer tumour grading. *PLoS One* 2008;3(8):e2925.
7. on Breast ECWG, Sloane JP, Amendoeira I, et al. Consistency achieved by 23 European pathologists in categorizing ductal carcinoma in situ of the breast using five classifications. *Human pathology* 1998;29(10):1056-62.