Statistical Analysis Plan for validating MIRAI on the OMI-DB dataset

Celeste Damiani, Adam Brentnall

18th May 2022

Contents

1. Executive summary

- This Statistical Analysis Plan (SAP) describes the planned analyses for the evaluation of the performance of MIRAI, a breast cancer risk model described in [YMS+21]:

  - on a subset of the OPTIMAM Mammography Image Database (OMI-DB) composed by prior (normal) mammograms of screen detected cases and matched controls;
  - on a subset of OMI-DB composed by by prior (normal) mammograms of interval cancers and matched controls.

  The planned analyses identified in this SAP will be included in future manuscripts. Exploratory analysis not necessarily identified in this SAP may be performed to support the planned analysis. Any post-hoc analysis not specified in this SAP will be identified as such in manuscripts for publications and added as amendments to this SAP.

- Data for the evaluation comes from the OPTIMAM Mammography Image Database (OMI-DB). This is an extensive mammography image database containing unprocessed and processed full-field digital mammograms from the UK's National Health Service Breast Screening Program. It also contains expert's determined ground truths and associated clinical data linked to the images. For most malignant cases we have annotations in form of bounding boxes coordinates. From this we extracted three datasets with cases (people that we knew being at risk) and controls (people we knew that three years after the considered mammogram would be still cancer-free), one for each experiment, matched on model of the mammography machine, on the site it was taken and on the age of the patient. We used separate controls for each dataset.

- The primary objective of the statistical analysis is:
  - Assess the strength of the risk score given by the MIRAI model as a predictor of future breast cancer risk in women attending the NHS Breast Screening Program, after adjustment for age, mammography machine, and site where the mammogram was taken.

- The secondary objective of the the statistical analysis is:
  - Assess potential heterogeneity of association by type of cancer (interval vs. screen detected), age and site.

2. General

2.1. SAP scope

Our overall aim is to evaluate the performance of MIRAI in estimating the risk to develop breast cancer. This SAP covers the intended statistical analysis of the performance of MIRAI using previously unseen data from OMI-DB. The specific output of MIRAI to be validated is risk score output for developing cancer after 3 years from the considered mammogram.

3. Summary of OMI-DB and methods of data collection

The OMI-DB is an extensive mammography image database comprised of unprocessed and processed full-field digital mammograms. It collects NHS Breast Screening Programme images from multiple breast screening centres across the UK, with the goal of serving as a large repository of medical images for research and training purposes. It also contains expert's determined ground truths and associated clinical data linked to the images. For most malignant cases we have annotations in form of bounding boxes coordinates.

As part of the data sharing agreement, we obtained a subset of this database composed by data of 18,800 patients, of which (before data cleaning):

- 5500 malignant cases with annotations
- 600 benign cases with annotations
- 800 malignant cases
- 1000 benign cases
- 900 interval cancer cases
- 10000 normal cases

For most imaging events (for example, a screening is an imaging event) we have both processed and unprocessed full-field digital mammograms, and two views of each breast, i.e. medio-lateral oblique (MLO) and cranio-caudal (CC) (plus several other views for cases with abnormalities). The database contains images from different manufacturers, particularly Hologic Inc (models Hologic Lorad Selenia and Selenia Dimensions Mammography Systems), and General Electric (GE) Medical Systems (models Senograph DS and Senographe Essential), Siemens and a smaller amount of Philips. The images were collected from three sites, and not all manufacturers were only used at each site.

4. Study details

4.1. Study objectives

This study is composed by two sub-studies. We will carry out analysis of the output of MIRAI on two datasets:

SCREEN-DETECTED dataset composed by normal screening acquired mammograms prior to screening detected malignant cancers (cases - 3 years prior to diagnosis), matched 1-1 with normal screening acquired mammograms prior to normal screening mammograms (controls) on model of the mammography machine, on the site it was taken and on the age of the patient. Controls are selected to not have any history of malignancy on record.

INTERVAL-CANCER dataset composed by normal screening acquired mammograms prior to interval cancers (cases - ≤ 3 years prior to diagnosis), matched 1-1 with normal screening acquired mammograms prior to normal screening mammograms (controls) on model of the mammography machine, on the site it was taken and on the age of the patient. Controls are selected to not have any history of malignancy on record.

Note that the sets of controls and cases for the two sub-studies are independent. However, since the the SCREEN-DETECTED set was matched first, there was little choice left for some of the cases in the INTERVAL-CANCER study, which had to be more loosely matched on age. For this reason we will consider an alternative dataset with controls that are not independent from the SCREEN-DETECTED dataset, but are perfectly matched on model, site and age.

The overall objective of the study is to externally validate and determine the strength of the MIRAI tool for future breast cancer risk assessment done on normal screening mammograms from the NHS Brest Screening Program, and to evaluate potential heterogeneity of performance for interval vs screen-detected cancers, and other selected subgroups.

All images come from a subset of OMI-DB that the algorithm developers have not previously had access to when developing the model MIRAI.

MIRAI's output is a tuple of risk scores at 1 year, 2 years, ..., 5 years per screening visit (ie. using four mammograms from the left and right breasts, and both views). Our primary analysis will use the score at 3 years because we have the follow-up status of all subject up to 3 years after the considered mammogram. The continuous output will be validated against biopsy-confirmed breast cancer diagnosis at a later date. For the SCREEN-DETECTED study cases were diagnosed a breast cancer at the routine screening exam 3 years later, or not; controls had a screening visit 3 years later but no cancer was detected. For the SCREEN-DETECTED study cases were diagnosed an interval cancer before the following routine screening exam 3 years later; controls were followup to their next screening visit 3 years later where no cancer was detected.

4.2. Analysis objectives

Primary objective Assess the strength of the risk score given by the MIRAI model as a predictor of future breast cancer risk in women attending the NHS Breast Screening Program, after adjustment for the site where the mammogram was taken, mammography machine, and age; and calibration of relative risk.

Secondary objective Assess potential heterogeneity of association by type of cancer (interval vs. screen-detected), age and site; and further methods to assess calibration of relative risk.

4.3. Study design

Design Both substudies are retrospective observational case-control studies.

Setting NHS Breast Cancer Screening units from 2 NHS Trusts in England with mammograms collected since 2010[HBWW$^+$21].

Target population Breast cancer screening population in the UK: women to attend the the NHS Breast Screening Program between the ages of 50 and 70 years.

Endpoints Diagnosis of a breast cancer. Our primary endpoint is diagnosis of invasive ductal carcinoma (IDC) or ductal carcinoma in situ (DCIS). Secondarily we'll look at them separately.

Sample population: selection of patients from OMI-DB

Sub-study 1: priors to screen-detected malignant cancers (SCREEN-DETECTED)

Inclusion criteria: screening studies from women in OMI-DB composed by standard four-view mammography of 'FOR PRESENTATION' type, with normal episode status (status at the time of the exam), normal or malignant episode outcome (status confirmed at the following screening exam), coming from routine referrals or first referrals.

Exclusion criteria: screening studies from women in OMI-DB having had surgical procedures or benign cancers, screening studies captured with Philips images or coming from the 'adde' site (these two characteristics roughly overlapped).

For the detailed process of selection of the screening studies, see Figure 1.
Then we selected the cases for this sub-study as follows:

- Full mammograms (4 images for the standard views) with a malignant prior (MP) 'Episode Outcome'. This means that at the following screening study, usually three years later, the patient was found to have a malignant (M) cancer.
- Images captured by HOLOGIC Lorad Selenia and HOLOGIC Selenia Dimensions machines, as required by MIRAI documentation [Yal21].

We selected only one full mammogram per patient. When studies were comprising of repeated images (same view, same side), we took the most recent one, flagging for possible quality concerns. We proceeded to match cases to controls by site, model of the device and age. 2062 cases are matched exactly on all parameters to 2062 controls. Other 6 are matched exactly by model and site, but up to ± 1 year in age to 6 controls. One more case is matched exactly by model and age, but not by site.
To summarise we have:

- 2069 cases (subjects), with a full set of four images each ('L CC', 'L MLO', 'R CC', 'R MLO')
- 2069 controls (subjects), with a full set of four images each, matched 1:1 on model, site and age, see Appendix A.

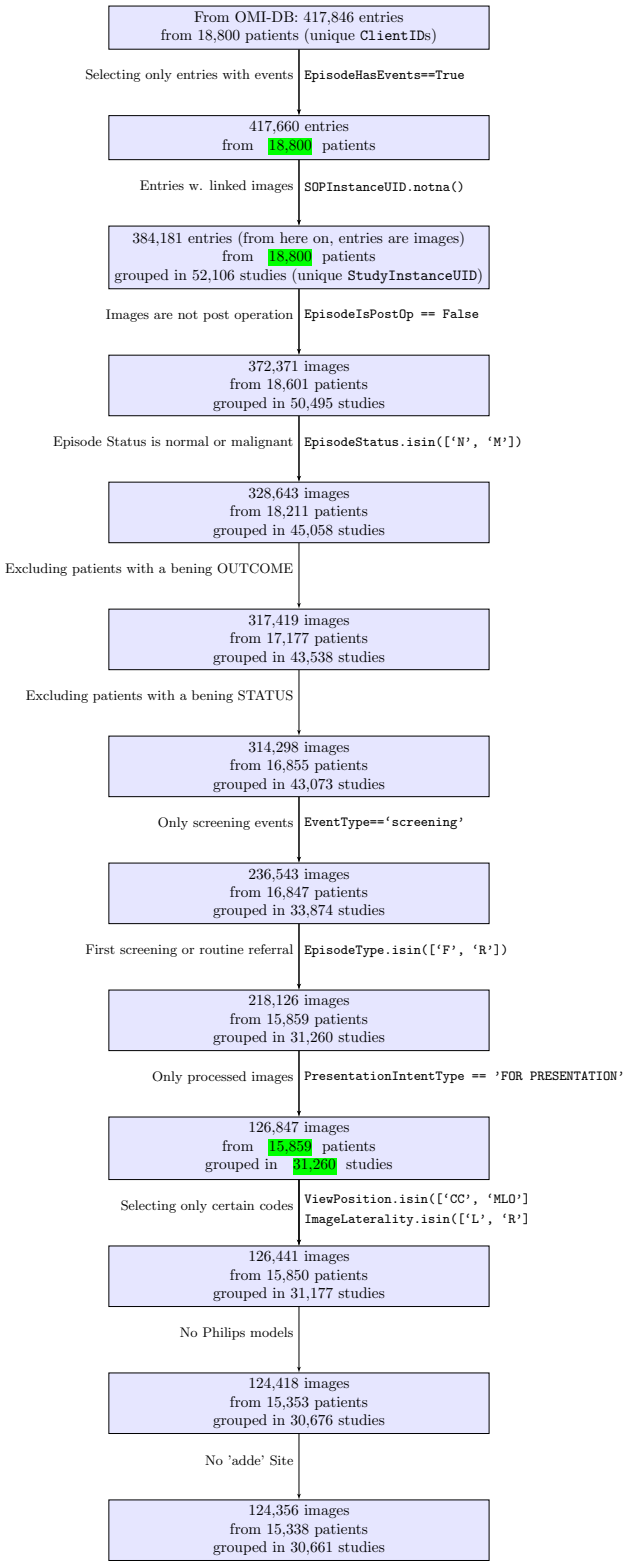

Figure 1: Selecting screening images from our subset of OMI-DB.

Sub-study 2: priors to interval cancers (INTERVAL-CANCER)

Inclusion criteria: screening studies from women in OMI-DB composed by standard four-view mammography of 'FOR PRESENTATION' type, with normal episode status (status at the time of the exam), normal or prior to interval cancer (CIP) episode outcome, coming from routine referrals or first referrals.

Exclusion criteria: screening studies from women in OMI-DB having had surgical procedures or benign cancers, screening studies captured with Philips images or coming from the 'adde' site (these two characteristics roughly overlapped).

The first steps for selecting the images were the same as illustrated in Figure 1.
Then we selected the cases for this sub-study as follows:

- Full mammograms (4 images for the standard views) with a prior to interval cancer (CIP) 'Episode Outcome'.
- Images captured by HOLOGIC Lorad Selenia and HOLOGIC Selenia Dimensions machines, as required by MIRAI documentation [Yal21].

We selected only one full mammogram per patient. We matched cases to controls by site, model of the Hologic device and age. 346 cases are matched exactly on all parameters to 346 controls. Other 58 cases are matched exactly by manufacturer (not model) and site, but up to ± 3 years in age to 58 controls. Other 285 cases are matched exactly by manufacturer (not model), up to ± 5, but not on site.
To summarise we have:

- 687 cases (subjects), with a full set of four images each ('L CC', 'L MLO', 'R CC', 'R MLO')
- 687 controls (subjects), with a full set of four images each, matched 1:1 on model, site and age, see tables in Appendix A.

The controls for this sub-study were selected to be independent from the controls to the SCREEN-DETECTED sub-study, with the SCREEN-DETECTED set matched first. This left little choice for some of the older cases in the interval cancer set. We define then an alternative matched set that includes some of the controls from the prior study, but with closer matching on age. In this alternative dataset for sensitivity analysis we have:

- 709 cases (subjects), with a full set of four images each ('L CC', 'L MLO', 'R CC', 'R MLO')
- 709 controls (subjects), with a full set of four images each, matched 1:1 on model, site and age, see tables in Appendix A.

The number of cases is bigger in this alternative sub-study because in the first one there were cases that did not have a matching control.

Sample size The sample size of 4138 patients (2069 cases + 2069 controls) for the SCREEN-DETECTED study and 1416 patients (687 cases + 687 controls) for the PRIORS-CONTROLS study is based on the maximum size feasible from the available data. Both of these studies have sufficient power to detect the effect of MIRAI if has similar performance to mammographic density. The matched c-index of earlier measures has been estimated to be approximately 0.60[BCFD15]; these studies will have 90% power to show an association between MIRAI and risk if the matched c-index is approximate 0.536 (PRIORS) or 0.562 (interval cancer). Based on previous reports of the performance of MIRAI it is possible that it will yield a stronger association than this for both settings, with the highest expected level of performance for the interval cancer study [YMS+21]. We will have approximately 90% power to test for heterogeneity if the matched c-index is 0.67 for the interval cancers, and 0.60 for the priors, which justifies the utility of testing for heterogeneity in the analysis.

5. Analysis

5.1. General principles

The same analysis will be done separately for both the screen-detected (SCREEN-DETECTED) and interval cancer (INTERVAL-CANCER) studies as primary analysis; both will be combined for certain secondary analysis. In addition to the main analysis below the samples will be described with summary statistics on matching factors by study, and case/control status. Median and inter-quartile range will be used for continuous variables, and categorical counts (%) for categorical variables and those subgroup categories specified below. Exploratory analysis may be done to explore why the algorithm performed differently, if found, including by viewing mammogram images.

5.2. Measures of effect

In both the prior and interval cancer analysis the primary measures of effect with be the odds ratio per standard deviation (in controls) of the natural logarithm of MIRAI absolute 3 years risk, adjusted by matching factors included in the study design; and the concordance index associated with MIRAI 3 years risk, after adjustment for matching factors. Unadjusted measures including area under the receiver operating characterisitic (AUC) will be secondary measures of effect. Heterogeneity will be assessed using tests for interaction.

5.3. Primary analysis

The odds ratio per standard deviation of the natural logarithm of absolute 3 years risk in controls will be estimated using conditional logistic regression, with 95% profile likelihood confidence intervals. The coefficient from the model (ie. log of the odds ratio) will also be presented with a 95%CI, per unit of the log of the 3 years risk. This is a measure of the calibration of the relative risk associated with MIRAI: a perfect calibration would yield a coefficient of 1.0. To measure discrimination of MIRAI, a matched concordance

index will be estimated with 95% confidence intervals from Wilsons method for binomial outcomes [BCFD15].

5.4. Secondary analysis

Interactions will be evaluated separately in each study through the conditional logistic regression model using a likelihood-ratio test, with 95% confidence intervals on the size of effect estimated using profile likelihood. For age, P-values will be based on using age as a continuous variable.

Hetereogeneity of MIRAI by interval cancer / screen-detected cancer will be evaluated by analysing both studies in the same conditional logistic regression model, with a test for interaction by study. In the absence of significant heterogeneity (p-value > 0.05 for the test of interaction), other tests for heterogeneity (age, type of cancer) will be done using both studies combined. Heterogeneity by type of cancer (invasive or in situ) and age will also be tested through the conditional logistic regression.

Logistic regression will be used to evaluate unadjusted effects of all estimates done with conditional logistic regression using profile likelihood 95%CIs; receiver operating characteristic (ROC) and area under the ROC (AUC) statistics will be also obtained using DeLong 95%CIs.

Calibration of the relative risk of MIRAI will be tabulated using deciles of risk in controls. If there is a lack of heterogeneity between interval and screen-detected datasets they will be combined for this, otherwise explored separately. The predicted relative risk will be plotted against observed relative risk following methods in [BvVH$^+$20]. 95%CIs on the observed relative risk by decile would be based on a logistic regression model adjusted for age, site and manufacturer. A smooth estimate of the observed calibration would be obtained using iterative exploration of the most appropriate model using the raw values of 3 years risk, based on model fit. It might include a polynomial in the logistic regression, or other non- or semi-parametric methods.

5.5. Subgroups

The following subgroups will be considered for presentation of effects by age: <55 years, $55-65$ years, 65+ years. These subgroups are chosen partly based on likely menopausal status (women <55 years old more likely to be premenopausal).

Heterogeneity in effects will be assessed for invasive breast cancer vs ductal carcinoma insitu (DCIS).

Secondary analysis of calibration will use deciles of predicted 3 years risk in controls only, as outlined above.

5.6. Missing data

If MIRAI fails for any woman then this will be reported. Primary analysis will be on a complete case basis, and the woman (and her matched case or control) would be excluded from the analysis. We do not anticipate a large proportion of missing data, nor being able

to predict the MIRAI score reliably using other information, and so multiple imputation or other techniques to account for missing data in the analysis will not be done.

5.7. Sensitivity analysis

We will carry out the following sensitivity analyses.

1. Model sensitivity analysis, whereby only Hologic mammograms are included in the analysis. This will be used to test robustness of findings by manufacturer.

2. Matching sensitivity analysis. We could not match exactly on all factors; a sensitivity analysis excluding sets not matched exactly will be carried out to assess if performance metrics improve.

3. The controls from the two datasets were selected to be independent, with the prior set matched first. Unfortunately this left little choice for some of the older cases in the interval cancer set. We will re-run the analysis for interval cancers by using a matched set (defined in advance) that includes some of the controls from the prior study, but with closer matching on age.

References

[BCFD15] Adam R. Brentnall, Jack Cuzick, John Field, and Stephen W. Duffy. A concordance index for matched case-control studies with applications in cancer risk. *Statistics in medicine*, 34:396–405, 2 2015. 8, 9

[BvVH+20] Adam R. Brentnall, Elke M. van Veen, Elaine F. Harkness, Sajjad Rafiq, Helen Byers, Susan M. Astley, Sarah Sampson, Anthony Howell, William G. Newman, Jack Cuzick, and Dafydd Gareth R. Evans. A case–control evaluation of 143 single nucleotide polymorphisms for breast cancer risk stratification with classical factors and mammographic density. *International Journal of Cancer*, 146:2122–2129, 4 2020. 9

[HBWW+21] Mark D. Halling-Brown, Lucy M. Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G. Wallis, Louise S. Wilkinson, Rosalind M. Given-Wilson, Rita McAvinchey, and Kenneth C. Young. Optimam mammography image database: A large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3, 1 2021. 5

[Yal21] Adam Yala. Mirai. `https://https://github.com/yala/Mirai`, 2021. 5, 7

[YMS+21] Adam Yala, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578), January 2021.

Publisher: American Association for the Advancement of Science Section: Research Article. 1, 8

6. Signatures of approval

Version number	Version Date	Summary of changes
1.0		First signed off version

SAP contributors:

Adam Brentnall	Queen Mary University of London	.....
Celeste Damiani	Queen Mary University of London	.....

Appendix A. Tables and figures

	screen-detected		Interval cancer	
	Cases	**Controls**	**Cases**	**Controls**
	$n=2069$	$n=2069$	$n=687$	$n=687$
Age (n, %)				
<55	479 (23%)	479 (23%)	216 (31%)	252 (37%)
55 to <65	1034 (50%)	1038 (50%)	290 (42%)	411 (60%)
65+	556 (27%)	552 (27%)	181 (26%)	24 (3%)
Site (n, %)				
Jarvis	1510 (73%)	1511 (73%)	597 (87%)	344 (50%)
St Georges	559 (27%)	558 (27%)	90 (13%)	343 (50%)
Model (n, %)				
HOLOGIC Lorad Selenia	1987 (96%)	1987 (96%)	646 (94%)	685 (100%)
HOLOGIC Selenia Dimensions	82 (4%)	82 (4%)	41 (6%)	2 (0%)
Cancer type (n, %)				
Invasive	1354 (65%)	-	612 (89%)	-
DCIS	449 (22%)	-	54 (8%)	-
NA	266 (13%)	-	21 (3%)	-

Table 1: Summary statistics for matching and other factors in the both case-control studies, with independent control sets. (Table will be repeated for missing data as supplementary table)

	Interval cancer alternative dataset	
	Cases	**Controls**
	$n=709$	$n=709$
Age (n, %)		
<55	216 (30%)	216 (30%)
55 to <65	290 (41%)	290 (41%)
65+	203 (29%)	203 (29%)
Site (n, %)		
Jarvis	618 (87%)	618 (87%)
St Georges	91 (13%)	91 (22%)
Model (n, %)		
HOLOGIC Lorad Selenia	668 (94%)	668 (94%)
HOLOGIC Selenia Dimensions	41 (6%)	41 (6%)
Cancer type (n, %)		
Invasive	633 (89%)	-
DCIS	55 (8%)	-
NA	21 (3%)	-

Table 2: Summary statistics for matching and other factors in the PRIOR-INTERVAL case-control sub-study with the alternative control sets. (Table will be repeated for missing data as supplementary table)

	aOR (95%CI)	**mC** (95%CI)	**Calibration** (95%CI)	P (het)
Overall	X (A to B)	X (A to B)	X (A to B)	P
Screen-detected	X (A to B)	X (A to B)	X (A to B)	
Interval cancer	X (A to B)	X (A to B)	X (A to B)	
Overall, subgroups				
Age, years				P*
<55	X (A to B)	X (A to B)	X (A to B)	
55 to <65	X (A to B)	X (A to B)	X (A to B)	
65+	X (A to B)	X (A to B)	X (A to B)	
Type of cancer				P*
Invasive	X (A to B)	X (A to B)	X (A to B)	
In situ	X (A to B)	X (A to B)	X (A to B)	
Screen-detected study subgroups				
Age, years				P*
<55	X (A to B)	X (A to B)	X (A to B)	
55 to <65	X (A to B)	X (A to B)	X (A to B)	
65+	X (A to B)	X (A to B)	X (A to B)	
Type of cancer				P*
Invasive	X (A to B)	X (A to B)	X (A to B)	
In situ	X (A to B)	X (A to B)	X (A to B)	
Interval-cancer study subgroups				
Age, years				P*
<55	X (A to B)	X (A to B)	X (A to B)	
55 to <65	X (A to B)	X (A to B)	X (A to B)	
65+	X (A to B)	X (A to B)	X (A to B)	
Type of cancer				P*
Invasive	X (A to B)	X (A to B)	X (A to B)	
In situ	X (A to B)	X (A to B)	X (A to B)	

Table 3: Model performance results (Table repeated using unadjusted AUC and logistic regression ORs as supplementary table, and for sensitivity analysis on matching for the interval-cancer study). *If there is evidence of heterogeneity between screen and interval cancer studies, then further analysis of heterogeneity by age and cancer type will be done separately, otherwise it will be done overall (ie. p-values will not all be reported depending on this). **Abbreviations:** aOR: adjusted odds ratio per standard deviation of the natural logarithm of absolute 3 years risk in controls; mC: matched concordance index. Values to 2dp, p-values to 2sf.

	aOR (95%CI)	**mC** (95%CI)	**Calibration** (95%CI)
Screen-detected study			
Model			
HOLOGIC Lorad Selenia	X (A to B)	X (A to B)	X (A to B)
HOLOGIC Selenia Dimensions	X (A to B)	X (A to B)	X (A to B)
Matching			
Main analysis	X (A to B)	X (A to B)	X (A to B)
Exact match only	X (A to B)	X (A to B)	X (A to B)
Interval cancer study			
Model			
HOLOGIC Lorad Selenia	X (A to B)	X (A to B)	X (A to B)
HOLOGIC Selenia Dimensions	X (A to B)	X (A to B)	X (A to B)
Matching			
Main analysis	X (A to B)	X (A to B)	X (A to B)
Exact match only	X (A to B)	X (A to B)	X (A to B)
Controls			
Main analysis	X (A to B)	X (A to B)	X (A to B)
New controls	X (A to B)	X (A to B)	X (A to B)

Table 4: Sensitivity analyses. Values to 2dp. Relevant part of the Table repeated for the interval-cancer sensitivity on matching.

Decile	Cutpoint MIRAI	Number women	Observed cases	Expected cases	O/E (95%CI)
Overall					
1	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
2	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
3	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
4	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
5	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
6	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
7	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
8	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
9	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
10	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
Screen-detected study					
1	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
2	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
3	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
4	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
5	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
6	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
7	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
8	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
9	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
10	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
Interval cancer study					
1	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
2	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
3	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
4	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
5	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
6	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
7	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
8	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
9	X.XX	N	O	E.E	X.XX (X.XX to X.XX)
10	X.XX	N	O	E.E	X.XX (X.XX to X.XX)

Table 5: Relative risk calibration by decile. Relevant part of the table repeated for sensitivity analysis on matching in the interval cancer data if overall calibration differs.