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

Corresponding author: Roger G Blanks PhD, Cancer Epidemiology Unit, Nuffield Department of Population Health, Oxford University, UK. roger.blanks@ndph.ox.ac.uk tel +44 (0) 1865 289663

Rosalind M Given-Wilson. MBBS, Dept of Radiology, St Georges University Hospital Foundation Trust, UK Rosalind.Given-Wilson@stgeorges.nhs.uk

Susan L Cohen, Public Health England, London UK Sue.Cohen@phe.gov.uk

Julietta Patnick BA. Cancer Epidemiology Unit, Nuffield Department of Population Health, Oxford University, UK julietta.patnick@ndph.ox.ac.uk

Rupert Alison MSc, Cancer Epidemiology Unit, Nuffield Department of Population Health, Oxford University, UK rupert.alison@ndph.ax.ac.uk

Matthew G Wallis MBCHB Cambridge Breast Unit, Cambridge University Hospitals NHS Trust, UK. matthew.wallis@addenbrokes.nhs.uk

Key words

Breast neoplasms, mass screening, mammography, recall rate

Key points

Question: How can we determine optimum recall rates in breast cancer screening?

Findings: In this large observational study we show that increases in recall rates above defined levels are almost exclusively associated with false positive recalls and a very small increase in low/intermediate grade DCIS.

Meaning: High recall rates are not associated with increases in detection of life threatening cancers. The models developed in this paper can be used to help set recall rate ranges that maximise benefit and minimise harm.

Abbreviations

NHSBSP National Health Service Breast Screening Programme

MMV Modelled maximum value

MS Modelled Slope (how rapidly MMV is reached)

P95 Recall rate at which 95% of MMV is reached

DCIS Ductal carcinoma in situ

SEER The Surveillance, Epidemiology and End Results program of the National Cancer Institute

AgeX trial Age extension trial

KC62 Korner return used to collect NHS data (breast screening return is no. 62)

LIG Low/intermediate grade (DCIS)

FPR False positive (non-cancer) recalls

PPV positive predictive value

Abstract

Objective To develop methods to model the relationship between cancer detection and recall rates to inform professional standards.

Methods: Annual screening programme information for each of the 80 English NHSBSP units (totalling 11.3 million screening tests) for the seven screening years from 1st April 2009 to 31st March 2016 and some Dutch screening programme information was used to produce linear and non-linear models. The non-linear models estimated the modelled maximum values (MMV) for cancers detected at different grades and estimated how rapidly the MMV was reached (the modelled 'slope' (MS)). Main outcomes include the detection rate for combined invasive/micro-invasive and high-grade DCIS (IHG) detection rate and the low/intermediate grade DCIS (LIG) detection rate.

Results: At prevalent screens for IHG cancers 99% of the MMV was reached at a recall rate of 7.0%. The LIG detection rate had no discernible plateau, increasing linearly at a rate of 0.12 per 1000 for every 1% increase in recall rate. At incident screens 99% of the MMV for IHG cancer detection was 4.0%. LIG DCIS increased linearly at a rate of 0.18 per 1000 per 1% increase in recall rate.

Conclusions: Our models demonstrate the diminishing returns associated with increasing recall rates. For example, above 7.0% at prevalent screens and 4.0% at incident screens most recalls were either false positive (no cancer detected) or resulted in a small increase in LIG detection only. Unlike all other cancers, which reached a near-plateau at observed recall rates, LIG detection continued to increase with recall rate. The screening programme in England could use the models to set recall rate ranges and other countries could explore similar methodology.

Introduction

Screening aims to detect breast cancers early to maximise the success of treatment and reduce breast cancer mortality [1]. There is controversy about the magnitude of both benefits and harms [2]. The drive to detect as many breast cancers as possible to maximise sensitivity, regardless of biological aggressiveness leads to over diagnosis of cancers that would not otherwise appear or cause problems to an individual woman in her life time and can lead to over treatment [2]. This also leads to high recall rates, which in turn leads to anxiety and societal cost.

The English NHSBSP invites women to three-yearly mammography between the ages of 50 and 70 years. Since 2010 as part of the AgeX trial [14] first invitations now occur in the age range 45-52 years. Double reading is standard. Women recalled undergo 'triple assessment' and can have a clinical examination, further imaging (including mammography and ultrasound) and if the results are suspicious a needle biopsy.

In the UK invasive cancer detection rate targets are informed by the Swedish-Two County randomised controlled trial, which detected nearly all cancers as invasive [3] and are age standardised [4]. European cancer detection rate targets are set as 3 times the underlying incidence rate for first screen and 1.5 for subsequent screens [5]. Detection of non-invasive cancers is more controversial. High rates of DCIS have been associated with high rates of small high grade invasive cancers [6]. A retrospective analysis of 5.2 million women from 2003-2007 showed a significant negative association of screen detected DCIS and rate of interval cancers [7]. However, benefits from detecting low-grade DCIS have been increasingly questioned. SEER data from USA showed surgery added no survival benefit for low risk disease [8] and retrospective analysis of trial and observational data suggests that low grade DCIS had a very low rate of progression. If low grade DCIS does recur as invasive cancer, it generally does so as low grade invasive cancer, where survival and treatment is excellent [9,10].

The recall rate policies of different countries lack a strong evidence base and are driven by the perceived value of a high specificity i.e. the value placed on minimising harms relative to benefit. As a consequently there is wide variation in practice. Targets for recall rates range vary from 2% (previously 1%) in Holland [11] to a recommended upper threshold of 12% in the United States [12]. Europe and the UK National Health Service Breast Screening Programme (NHSBSP) set separate targets for recall rate. At prevalent (first) screens; < 5% in Europe with a minimum standard of <7% (but <7% and <10% respectively in the NHSBSP). At incident (subsequent) screens the European target is <3% with a minimum standard<5% (<5% and <7% in the NHSBSP) [5]. In England, attempts to reduce the recall rates at prevalent screens have not been successful, with 61% of screening units exceeding the 7%

target in 2015/2016 [13]. The same is true in the US where from a sample of 359 radiologists 37.8% had a recall rate above 12% [12].

This study aims to model the relationship between cancer detection and recall rates with a view to inform evidence-based recall rate ranges to balance harms and benefits using data from England supplemented by published data from Holland.

Methods

Cancer detection rates and recall rates are taken from the national (KC62) returns sent to PHE annually and published by NHS Digital [13]. Additional data at low recall rates have been obtained from published data from Dutch national screening programme [11], where the focus is on maintaining a defined low recall rate. We have used prevalent screens because the Dutch programme uses a two- year rather than the English three-year interval, so incident screens are less comparable.

Data are from the 80 English screening units for the seven screening years 2009/2010 to 2015/2016. The KC62 annex provides anonymised information on grade, nodal status and size for each cancer. The analyses are based on restricting the data to prevalent screens at ages 45-52 and incident screens at ages 53-70 to ensure maximum comparability between units. The study has no patient contact, intervention or use of identifiable patient data and is therefore exempted from ethical review in the UK.

English data for prevalent screens has been analysed as four groups with recall rates of <6%, 6-7.49%, 7.5-8.99% and 9+%. Further models using English data only are based on individual data on all 80 units weighted by the number of women screened by each unit to allow adjustment for any potential confounding factors, including age. We estimate that for the period 2009/10 to 2015/16 about 65% of screens in England used digital mammography and the Dutch data has been weighted to give equivalent data. Incident screens have been analysed as four recall rate groups <2.5%, 2.5-2.9%, 3-3.49% and ≥3.5% or as 80 screening units weighted by the number of women screened by each unit.

A linear test of trend across the four English recall rate groups used binomial regression with mean recall rate entered in the model and risk difference specified. Although of limited value as the true relationship is non-linear the finding of no significant evidence of a trend ($p>0.1$) is taken as evidence that the cancers are mostly detectable at the lower recall rates.

To measure the relationship between cancer detection and recall rates the observed data were fitted with two-parameter negative exponential models except for the low/intermediate grade DCIS (LIG) data which were better fitted with a linear model ($y=bx$),

where y is the detection rate, x the recall rate and b the gradient. All models go through the origin (0,0) where no women recalled equates to no cancers detected. The two- parameter negative exponential models ($y=b_1(1-b_2^x)$) give two values. The parameter b_1 is the modelled maximum value (MMV), which is the maximum possible detection rate achieved by just increasing recall rates. The parameter b_2 is a value between 0 and 1.0 we have termed the modelled slope (MS) which gives the rate at which the MMV is reached. Low b_2 values e.g. 0.2 indicate that detection rates rise very quickly with increasing recall rate while a high value b_2 e.g. 0.8 indicates that rates rise more slowly. The recall rate associated with 95% of the MMV (P95) is calculated as ln(1-p)/ln(MS) where p=0.95 and a recall rate of 99% of the MMV (P99) by p=0.99 etc. All statistical analysis was conducted using STATA version 14 or 15 (StataCorp, College Station, Texas).

Results

Between 1 April 2009 and 31 March 2016 there were 11,258,620 screens included in this study, of which 2,295,016 screens were routine prevalent screens at ages 45-52 yrs and 8,963,604 incident screens at ages 53-70 yrs.

Cancer detection and recall rates

Table 1 shows English screening unit data for prevalent screens grouped by recall rate with additional Dutch information in the footnote. Table 1 also shows information on grade of cancers, the significance of the trend across recall rate groups and the modelled maximum value (MMV). Fig 1 shows the modelled association between the prevalent screen cancer detection rate and recall rate ($y=7.71\ (1-0.66^x)$) and the models separately for invasive ($y=5.26\ (1-0.56^x)$) and non-invasive cancers ($y=2.68\ (1-0.81^x)$). The modelled slope (MS) for invasive cancers at 0.56, indicates that the curve rises more rapidly and then reaches a near plateau quicker than for all cancers (MS=0.66) and non/micro-invasive cancers (MS=0.81). The MS value for non/micro-invasive cancers of 0.81 indicates a much slower rise in detection rates increasing beyond the highest observed recall rates. The models are very similar for English only data weighted by number of women screened (see appendix table B2).

There is a substantial difference between high grade DCIS and LIG and therefore a model for all non-invasive cancers is inadequate. There is no trend for grade 3 invasive cancers (p=0.78) nor high-grade DCIS (p=0.83) across recall rates and we conclude that they are generally detected at the lowest recall rates. For LIG the test of trend is highly significant (p<0.001) and a linear model gives a better fit to the data predicting the rates increase by 0.12 per 1000 for each 1% increase in recall rates with no evidence of a maximum value. A MMV is therefore given for all cancers except for LIG.

Table 2 shows similar information to table 1, but for incident screens. Trends are seen for grade 1 & 2 invasive cancers and LIG. Again, for LIG detection a linear model gives a better fit, suggesting that the detection rates increase by 0.18 per 1000 per 1% increase in recall rates with no evidence of a maximum value.

IHG cancer and LIG DCIS detection by recall rate

We have combined invasive, micro-invasive and high-grade DCIS rates, termed the IHG detection rate (shown in the last row of tables 1 and 2). The modelled values are shown in fig 2a for recall rates (%) and fig 2b for false positive recall rate per 1000 (see below). Note the graphs look similar because most women recalled do not have cancer. The model for the prevalent screens IHG detection rate with recall rate is $y=6.41(1-0.52^x)$. It predicts 95% of the MMV at a recall rate of 4.6% and 99% at a recall rate of 7.0%. For incident screens the model is $y=7.53(1-0.32^x)$ and the recall rates needed to detect 95% and 99% of the MMV for IHG are 2.6% and 4.0% respectively. The graphs show that most additional recalls above 7.0% at prevalent screens and 4.0% at incident screens are false positive recalls with only a small increase in LIG detection.

False positive recall rates

We can predict the number of false positive (non-cancer) recalls (FPR) per 1000 women that will occur with increasing recall rate. The FPR rates are shown for prevalent screens in fig 2c and incident screens in fig 2d together with vertical lines indicating the recall rates associated with detection of 95% and 99% of the MMV for IHG cancers. For prevalent screens the FPR = $[10x – (6.4(1-0.52^X) + 0.12x)]$ and for incident screens the FPR = $[10x – (6.5(1-0.32^X) + 0.18x)]$ where x is the recall rate in percent. These formulae are based on the models for IHG cancers and LIG DCIS considered separately as we can only fit a linear model to the LIG DCIS detection rates. At prevalent screens the P95, P99 values for IHG correspond to FPR per 1000 rates of 39.4 and 62.8 respectively giving a positive predictive value of recall (PPV) of 14.3% and 10.3% respectively. For IHG detection the difference between the P99 & P95 value is an increase of 0.26 per 1000 and requires that false positive recalls increase by 23.4 per 1000, which is 90 false positive recalls per additional IHG cancer detected. From P99 to P99.9 the absolute increase in IHG rate is 0.06 per 1000, but we need to increase the false positive rates by 35.1 per 1000 to 97.9 per 1000 (the PPV has dropped to 7.2%), which is 585 false positive recalls per additional IHG cancer detected.

At incident screens the FPR rates for the P95 and P99 values are 18.7 per 1000 and 32.2 per 1000 respectively, giving PPV of recall of 25.5% to 17.9% respectively. This is a difference of 13.5 per 1000 in the false positive rate to detect a difference in IHG rate of 0.30 per 1000 (45 false positive recalls per additional IHG cancer). To increase from P99 to P99.9 incurs 317 false positive recalls per additional IHG cancer. The models therefore predict rapidly

diminishing returns from increasing recall rates, where above the P99 recall rate almost all recalls are false positive, except for a very small increase in LIG. Fig 2b shows graphically that a unit operating at a false positive recall rate of 100 per 1000 detects almost no more IHG cancers than one operating at 40 per 1000. It detects about one extra LIG detected per 1000 women screened.

P99 values by grade of cancer at incident screens

Fig 3 shows modelled data of incident screen cancers by grade against recall rate. Nearly all grade 3 invasive cancers are detected at a recall rate of 2.5%, grade 2 cancers at a recall rate of 3.9% (95%CI 2.7-5.6%) and grade 1 cancers at a recall rate of 5.2% (3.0-9.3%). Grade 3 invasive cancers at incident screens tend to be larger than grade 2 and grade 1 cancers. Further analysis shows that at incident screens the additional grade 1 & 2 invasive cancers from increasing recall rates (see table 2) are ductal rather than lobular or any other histological type. We conclude that the increased detection in invasive cancers between units in England using the higher compared to lower recall rates is mostly a small increase in grade 1 & 2 ductal invasive cancers.

Discussion

We have modelled the relationship between cancer detection rates and recall rates and shown that for prevalent screens 99% of the maximum modelled value (MMV) for IHG cancers is achieved at an estimated recall rate of around 7%. Above 7% there is still an increase in LIG detection, which has a more linear relationship with recall rates and no evidence of a plateau. For incident screens the pattern is similar with 99% of the modelled MMV for IHG cancers achieved at a recall rate of about 4% and no evidence that the LIG rate has a plateau. We have also shown how rapidly the false positive recall rates increase between an IHG rate of 95% to 99% of the MMV. The purpose of breast cancer screening is to reduce breast cancer mortality. There needs, however, to be a balance between this benefit and any harms [15]. Screening programmes need to maximise the detection of higher risk cancers, whilst trying to minimise over diagnosis and false positive recalls to assessment.

Our findings agree with a direct comparison of USA and UK data which showed that recall rates were twice as high in the USA, but the invasive cancers detection rates similar, and the non-invasive detection rates higher [16].

A strength of this study is the large and comprehensive dataset from the English national programme enhanced by data at lower recall rates at prevalent screens from Holland. The effect of very low recall rates on reducing cancer detection has been shown before [17], but not the plateau effect for IHG cancers and the continuing increase in low/intermediate

grade DCIS (0.12 per 1000 for each 1% increase in recall rate at prevalent and 0.18 per 1000 for each 1% increase in recall rate at incident screens).

The current recall rate targets lack a strong scientific basis. Our solution is to fit models to the observed data and use these to examine the association between detection rates and recall rates. No model is perfect, but our interest is in whether it is informative. The models are explicit and can be critically examined and therefore, potentially updated.

In the analyses in this paper our interest is in the relative detection rates achieved by units within the NHSBSP in relation to their recall rates. In England there is rigorous training, guidelines and individual and unit performance monitoring in a double reading environment. There is also a requirement for all readers to read a minimum of 5,000 films per year to remain accredited. There is little or no confounding by background incidence across these units [18].

There are limitations associated with the observational nature of the study and the form of available data. The KC62 returns do not allow us to examine breast density or other potential confounding factors such as population characteristics. All women in the study are asymptomatic women undergoing routine screening and we would not expect much variation in characteristics of women between screening units. We have supplemented our data with that from the Dutch programme at lower recall rates, but models with only English data give similar results. This should not be taken as evidence that the results from our models directly apply to the Netherlands or any other countries screening programmes. National screening programmes vary in screening intervals, age ranges and underlying background incidence as well as using different equipment and programme design. This type of modelling could, however, be undertaken by other countries' screening programmes, but large numbers of screened women are required where outcomes are rare (such as LIG cancers).

We assume that the huge variation between Dutch, UK and US recall rates is not primarily a matter of film reading skill, but more of perceived values in balancing sensitivity with specificity. We make the same assumption between units within England. Once a units assessment clinic workload is determined following its perceived sensitivity/specificity trade-off there is rarely much variation in that units recall rate. Due to these underlying assumptions in the models any change in a unit's recall rates need to be very carefully monitored to ensure that the changes to cancer detection rates are in line with the models.

We have specifically concentrated on modelling the relationship between cancer detection rates and recall rates and we have not tried to review the absolute sensitivity of screening. This is because interval cancer ascertainment is both difficult and less timely. However, we have historical data of programmes with low recall and cancer detection that improve on increasing recall supporting our argument [17] and have previously demonstrated the inverse relationship between interval cancer rates and cancer detection [18]. A recent

paper by Burnside et al using film-screen data from the NHSBSP has shown that there is a small decline in interval cancers with increasing recall rate [19]. They show one extra interval cancer prevented per 180 additional recalls at prevalent screens and one per 80 recalls at incident screens and concur that there should be a minimum recall rate.

About a quarter of English units are currently using incident screen recall rates below 2.6%. Our study suggests that if units moved toward the optimum incident screen recall rate of around 3.1% (range 2.6% to 4%) there would be an overall gain in invasive cancers detected and interval cancer rates may be reduced.

In our study, the modelled maximum value (MMV) is the maximum achievable detection rate from an increase in recall rates with the given technology and double reading. Additional work looking at models for prevalent screens estimated at 100% (rather than 65%) digital gives very similar results. We see no reason to think that the underlying model should produce a different pattern for single reading or different technology. In the most recent data from Oslo, digital Breast Tomosynthesis improved the sensitivity/specificity trade-off i.e. decreased recall rate and increased cancer detection without a reduction in interval cancers [20]. The excess cancers tended to be smaller, lower grade and node negative suggesting new technology might make the slope of the curve (MS) steeper leading to a P99 MMV at an even lower recall rate.

Finally, high quality population screening is about maximising lives saved whilst minimising harm incurred in saving those lives. So a good question is, for example, is a programme with 95% IHG MMV and a three year interval more or less efficient than a programme with an 85% IHG MMV using a two or even one year interval. This is beyond the scope of this current paper.

Conclusions: Above 7.0% at prevalent screens and 4.0% at incident screens most recalls were false positive (no cancer detected) or resulted in a small increase only in LIG DCIS detection. Our models suggest that below the P95 values of 4.6% and 2.6% respectively the IHG detection rate begins to more rapidly decrease. The English screening programme could consider this evidence to inform setting of a target recall rate range and other countries could consider the use of similar methodology.

Acknowledgments

Roger Blanks and Rupert Alison are funded by Public Health England. The work arose from initial discussions at the Clinical Advisory Group for NHSBSP assessment work (members Roger Blanks, Claire Borrelli, Sue Cohen, Alison Duncan, Rosalind Given-Wilson, Jacquie Jenkins, Olive Kearins, Sarah Pinder, Mark Sibbering, Nisha Sharma, Jim Steel, Anne Turnbull, Matthew Wallis).

References

1 Consolidated Standards for NHS Breast Screening Programme April 2017 Public Health England https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/589227/Breast_draft_standards_V1.7.pdf accessed 20/7/2017

2 Marmot MG, Altman DG, Cameron JA, Thompson SG, Wilcox M. The independent UK Review Panel on Breast Cancer screening. The benefits and Harms of breast screening: an independent review BJC 2013;108: 2205-2240

3 Tabar L, Fagerberg G, Duffy S, Day N, Gad A, Grontoft O. Update of the Swedish-Two County programme of mammographic screening for breast cancer. Radiol Clin North Am. 1992;30:187-210

4 Blanks RG, Day NE, Moss SM. Monitoring the performance of breast screening programmes: use of indirect standardisation in evaluating the invasive cancer detection rate. J Med Screening 1996; 3:79-81.

5 European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth Edition. Luxembourg: Office for Official Publication of the European communities, 2006.

6 Evans AJ, Blanks RG Should Breast Screening Programmes Limit their Detection of Ductal Carcinoma In Situ? Clin Rad 2002;57: 1086-1089

7 Duffy SW Dibden A, Michalopoulos D, et al. et al Screen detection of ductal carcinoma in situ and subsequent incidence of invasive interval breast cancers: a retrospective population-based study 2016 Lancet Oncol; 17:109-114. http://dx.doi.org/10.1016/S1470-2045(15)00446-5

8 Sagara Y, Mallory MA, Wong S, I Aydogan F, Desantis S, Barry WT et al. Survival benefit of breast surgery for low-grade ductal carcinoma in Situ JAMA Surg. 2015;150(8):739-745

9 Wallis MG, Clements K, Kearins O, Ball G, Macartney J, Lawrence GM. The effect of DCIS grade on rate, type and time to recurrence after 15 years of follow-up of screen-detected DCIS. Br J Cancer 2012;106: 1611–7. doi:10.1038/bjc.2012.151.

10 Benson JR, Jatoi I, Toi M. Treatment of low risk ductal carcinoma-in-situ: is nothing better than something? Lancet Oncol 2016:17(1): e442-e451.

11 Van Luijt PA, Fracheboud J, Heijnsdijk EAM, den Heeten GJ, de Koning HJ. Nation-wide data on screening performance during the transition to digital mammography: Observation in 6 million screens. European Journal of Cancer (2013) 49, 3517-3525.

12 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

13 NHS Digital, Breast Screening Programme, England - 2015-16 https://digital.nhs.uk/catalogue/PUB23376 accessed 1 Jan 2018

14 http://www.agex.uk/ accessed 1 Jan 2018

15 Raffle AE, Muir-Gray JA. Screening: Evidence and Practice. Oxford University Press. 2007.

16 Smith Bindman R, Chu PW, Miglioretti DL, Sickles E, Blanks R, Ballard-Barbash R, Bobo JK, Lee NC, Wallis MG, Patnick J, Kerlikowske K. Comparison of Screening Mammography in the United States and United Kingdom. JAMA 2003 290: 2129-2138.

17 Wallis MG, Lawrence GM, Brenner RJ. Improving Quality outcomes in a single payer system Lessons learnt from the UK Breast Screening Program. JACR 2008; 5:737-743

18 Given-Wilson R, Blanks RG Moss SM et al. An evaluation of breast cancer screening in South Thames (West) Region of the UK NHS Breast Screening Programme for the first 10 years. The Breast 1999;8: 66-71.

19 Burnside ES Vulcan D, Blanks RG, Duffy SW. The association between screening mammography recall rate and interval cancers in the UK Breast Cancer Service Screening Programme: a Cohort study. Radiology 2018:288(1);47-54.

20 Skaane P, Sebuodegard S Bandos AI, Gur D, Osteras BH, Gullien R, Holvind S. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo Tomosynthesis Screening Trial. Breast Cancer Res Treat 2018 https://doi.org/10.1007/s10549-018-4705-2

Figure and table legends

Table 1 Prevalent screen observed rates women aged 45-52 by recall rate including information on grade of invasive cancer and grade of non-invasive cancer and modelled maximum value (MMV) for detection rate

Table 2 Incident screen observed rates women aged 53-70 by recall rate including information on grade and modelled maximum value (MMV) of detection rate

Fig 1 Datapoints and modelled assocaition between prevalent screen cancer detection rates and recall rates ((All cancers y=7.71 (1-0.66^x)), Invasive cancers (y=5.26(1-0.56^x)) & non-invasive cancers (y=2.68 (1-0.81^x))) using English and Dutch (65% digital) data. MS values of 0.81 for non-invasive cancer indicate a much slower increase in detection with recall rates compared with 0.66 for all cancers and 0.56 for invasive cancers.

Figs 2a-2d Modelled cancer detection rate per 1000 against recall rate (fig 2a) and false positive recall rate (2b) for IHG and LIG at incident and prevalent screens. Modelled false positive recall (FPR), IHG and LIG per 1000 for prevalent screens (fig 2c) with P95 and P99 recall rate values and same for incident screens (fig 2d).

Fig 3 Modelled English data for incident[a] screens by cancer grade indicating P99 recall rate values values for grade 3 and HG DCIS at 2.5%, grade 2 at 3.9% and grade 1 at 5.2%.

Table 1 Prevalent screen observed rates women aged 45-52 by recall rate group including information on grade of invasive cancer and grade of non-invasive cancer,modelled maximum value (MMV) for detection rate and the final two parameter negative exponential model

	Group 1	**Group 2**	**Group 3**	**Group 4**	**Trend[a] p-value**	**Model/ MMV (95%CI)[b]**
Recall rate range	**<6**	**6-7.49**	**7.5-8.99**	**9+**		
Mean recall rate	**5.28**	**6.64**	**8.37**	**9.84**		
Units (N)	**14**	**23**	**21**	**22**		
Screened	378,744	691,293	592,112	632,867		
Mean age (yrs)	50.3	50.4	50.3	50.3		
Invasive (rate per 1000)	1,928 (5.09)	3,520 (5.09)	3,022 (5.10)	3,395 (5.36)	0.04	**$y=5.24\ (1-0.53^x)$** 5.24 (5.06-5.42)
Grade 3 (rate per 1000)	361 (0.95)	648 (0.92)	577 (0.97)	604 (0.95)	0.78	**$y=0.97(1-0.59^x)$** 0.97 (0.87-1.07)
Grade 1 & 2 (rate per 1000)	1567 (4.14)	2872 (4.15)	2445 (4.13)	2789 (4.41)	0.03	**$y=4.26(1-0.50^x)$** 4.26 (4.09-4.43)
Micro-invasive	21 (0.06)	60 (0.09)	37 (0.06)	42 (0.07)	0.74	
Non-invasive (rate per 1000)	678 (1.8)	1382(2.0)	1222 (2.1)	1460 (2.3)	<0.001	**$y=2.50(1-0.79^x)$** 2.50 (2.02-2.99)
% of High grade DCIS	61.2%	54.8%	54.4%	46.6%		
High grade DCIS (rate per 1000)	415 (1.10)	757 (1.10)	665 (1.12)	680 (1.08)	0.83	**$y=1.10(1-0.42^x)$** 1.10 (1.02-1.17)
Low/intermediate grade DCIS (rate per 1000)	263 (0.69)	625 (0.90)	557 (0.94)	780 (1.23)	<0.001	N/A[c] linear model y=0.12x gives better fit.
IHG (inv/micro/HG DCIS)	2364 (6.24)	4337 (6.27)	3724 (6.28)	4115 (6.50)	0.09	**$y=6.41(1-0.52^x)$** 6.41(6.01-6.81)

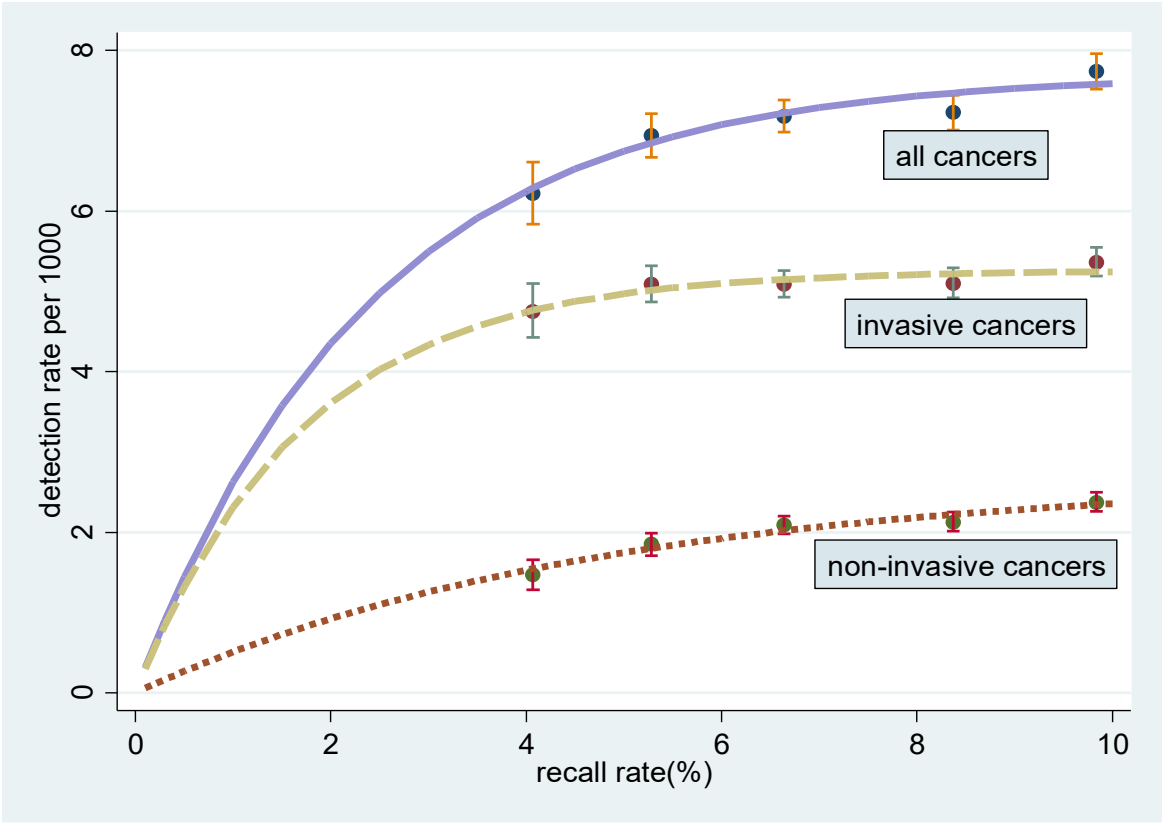
Note Dutch data from van Luijt et al (ref 11) for 65% digital usage at age 49-51 yrs estimated as cancer detection rate 6.22 per 1000, invasive 4.75 per 100 & non-invasive 1.47 per 1000 at a recall rate of 4.07%. [a]Evidence of trend based on binomial regression (binreg command in STATA with mean recall rate entered for each group and risk difference specified)). [b] maximum rate predicted by two-parameter negative exponential (non-linear) model. [c]Note the low/intermediate grade DCIS data is better fitted with a linear model (P<0.001) than a non-linear model (P=0.21) and therefore it is not possible to estimate a maximum detection rate (any maximum is likely to be at much higher recall rates than used by English units).

Table 2 Incident screen observed rates women aged 53-70 by recall rate group including information on grade,modelled maximum value (MMV) of detection rate and the final two parameter negative exponential model

	Group 1	Group 2	Group 3	Group 4	Trend[a] p-value	MMV (95%CI)[b]/Model
Recall rate range	**<2.5**	**2.5-2.99**	**3-3.49**	**3.5+**		
Mean recall rate	**2.21**	**2.79**	**3.22**	**3.81**		
Units (N)	**23**	**19**	**26**	**12**		
Screened	2,584,083	2,040,891	2,972,227	1,366,403		
Mean age (yrs)	61.6	61.5	61.6	61.7		
Invasive (rate per 1000)	15,551 (6.02)	12,566 (6.16)	18,943 (6.37)	8,758 (6.41)	<0.001	**$y=6.46(1-0.30^x)$** 6.46 (6.25-6.67)
Grade 3 (rate per 1000)	3344 (1.29)	2714 (1.33)	3845 (1.29)	1734 (1.27)	0.49	**$y=1.29(1-0.25^x)$** 1.29 (1.23-1.34)
Grade 1 & 2 (rate per 1000)	12,208 (4.72)	9,852 (4.83)	15,098 (5.08)	7,024 (5.14)	<0.001	**$y=5.20(1-0.35^x)$** 5.20(4.98-5.42)
Micro-invasive	139 (0.05)	141 (0.07)	155 (0.05)	119 (0.09)	0.01	
Non-invasive (rate per 1000)	3550 (1.37)	3064 (1.50)	4353 (1.46)	2247 (1.64)	<0.001	**$y=1.63\ (1-0.43^x)$** 1.63 (1.48-1.77)
% of High grade DCIS	66.8%	61.0%	63.6%	59.8%		
High grade DCIS (rate per 1000)	2371 (0.92)	1869 (0.92)	2769 (0.93)	1344 (0.98)	0.07	**$y=0.95(1-0.25^x)$** 0.95(0.89-1.02)
Low/intermediate grade DCIS (rate per 1000)	1179 (0.46)	1195 (0.59)	1585(0.53)	903(0.66)	<0.001	N/A data is considered tp be better fitted with linear model y=0.18x
IHG (inv/micro/HG DCIS)	18061 (6.99)	14576 (7.14)	21867 (7.36)	10221 (7.48)	<0.001	**$y=7.53\ (1-0.32^x)$** 7.53 (7.14-7.93)

[a] Evidence of trend based on binomial regression (binreg command in STATA). [b] maximum rate predicted by two-parameter negative exponential (non-linear) model. [c] Note the low/intermediate grade DCIS data is equally well fitted with a linear model as a non-linear model ($y=0.73(1-0.62^x)$) and therefore it is not possible to estimate a maximum detection rate with any confidence. The two-parameter negative exponential model suggests that 99% of low/intermediate grade DCIS is detected at a recall rate of 5.6% which is well above the observed recall rates and the linear model suggests rates increase at 0.18 per 1000 for every 1% increase in recall rate. Any maximum detection rate is therefore likely to be at much higher recall rates than used by English units.

Fig 1 Datapoints and modelled assocaition between prevalent screen cancer detection rates and recall rates ((All cancers $y=7.71\ (1-0.66^{x})$), Invasive cancers ($y=5.26(1-0.56^{x})$) & non-invasive cancer detection rate ($y=2.68\ (1-0.81^{x})$)) using English and Dutch (65% digital) data



Figs 2a-2d Modelled cancer detection rate per 1000 against recall rate (fig 2a) and false positive recall rate (2b) for IHG and LIG at incident and prevalent screens. Modelled false positive recall (FPR), IHG and LIG per 1000 for prevalent screens (fig 2c) with P95 and P99 recall rate values and same for incident screens (fig 2d)

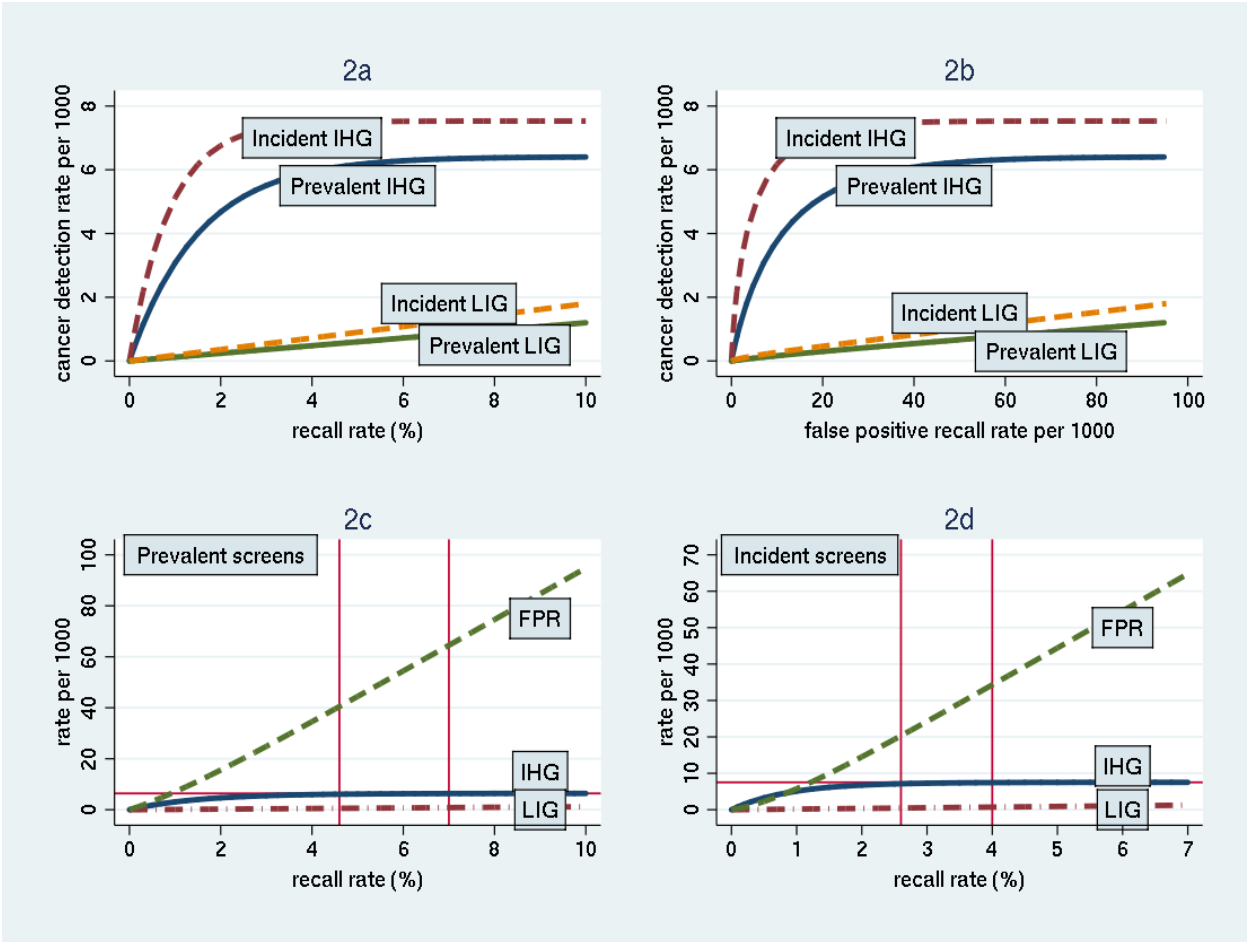
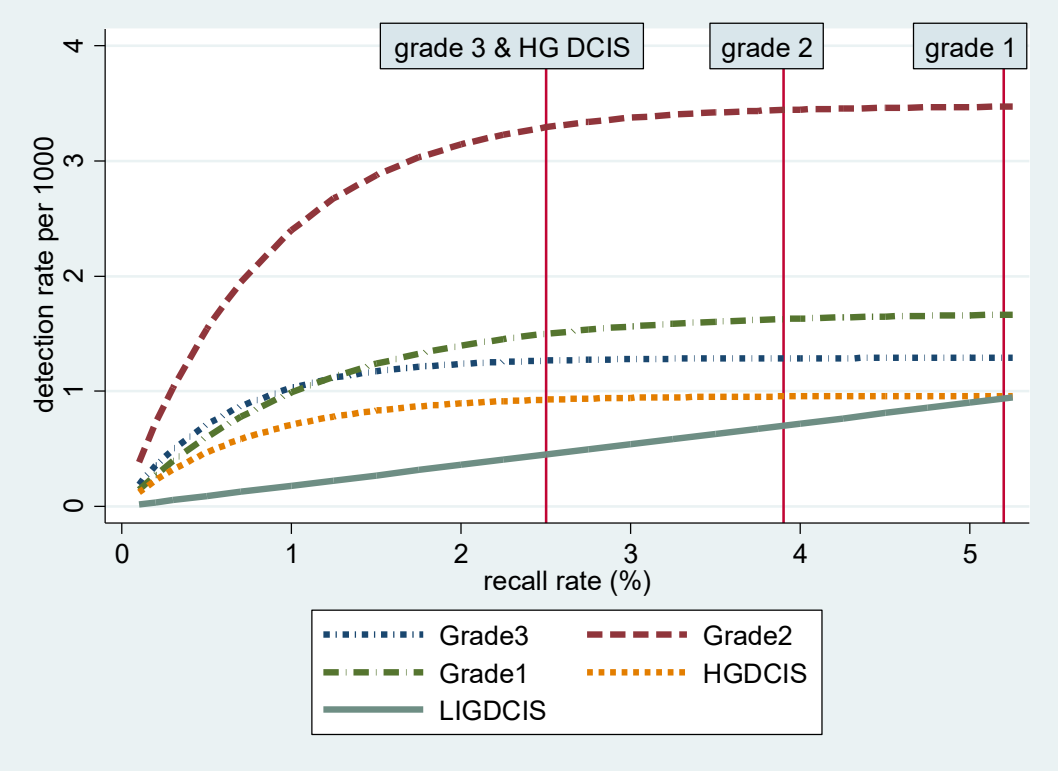


Fig 3 Modelled English data for incident[a] screens by cancer grade indicating P99 recall rate values values for grade 3 and HG DCIS at 2.5%, grade 2 at 3.9% and grade 1 at 5.2%.



[a]Plot not included for prevalent screens because of smaller numbers of cases causing wider confidence limits and therefore difficulty in interpretation. Prevalent screens are best interpreted as IHG and LIG as per fig 2a. Incidence screen rates for each grade of cancer are more statistically robust. Note that for incident screens P99 values are inversley correlated with mean size of invasive (grade 3 being largest and having lowest P99).