

Association between Screening Mammography Recall Rate and Interval Cancers in the UK Breast Cancer Service Screening Program: A Cohort Study

Elizabeth S. Burnside, MD, MPH, MS • Daniel Vulkan, MSc • Roger G. Blanks, PhD • Stephen W. Duffy, BSc, MSc

From the Department of Radiology, University of Wisconsin School of Medicine and Public Health, E3/311 Clinical Science Center, 600 Highland Ave, Madison, WI 53792-3252 (E.S.B.); Centre for Cancer Prevention, Queen Mary University of London, Wolfson Institute of Preventive Medicine, London, England (D.V., S.W.D.); and Nuffield Department of Population Health, University of Oxford, Oxford, England (R.G.B.). Received August 16, 2017; revision requested October 11; revision received December 13; final version accepted December 21. Address correspondence to E.S.B. (e-mail: eburnside@uwhealth.org).

Supported by Clinical and Translational Science Award (CTSA) program through a National Institutes of Health National Cancer Center for Advancing Translational Sciences (NCATS) grant (UL1TR000427), University of Wisconsin Carbone Comprehensive Cancer Center (grant P30CA014520), and the School of Medicine and Public Health, University of Wisconsin-Madison from the Wisconsin Partnership Program. R.G.B. supported by Cancer Research United Kingdom (CRUK). D.V. supported by NHS Cancer Screening Programmes. E.S.B. supported by the National Institutes of Health (grants K24CA194251 and R01CA165229). S.D. took part in this work as part of the program of The Policy Research Unit in Cancer Awareness, Screening, and Early Diagnosis, which receives funding from the Department of Health Policy Research Programme. It is a collaboration between researchers from seven institutions (Queen Mary University of London, UCL, King's College London, London School of Hygiene and Tropical Medicine, Hull York Medical School, Durham University and Peninsula Medical School).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of interest are listed at the end of this article.

Radiology 2018; 288:47–54 • <https://doi.org/10.1148/radiol.2018171539> • Content code: **BR**

Purpose: To determine whether low levels of recall lead to increased interval cancers and the magnitude of this effect.

Materials and Methods: The authors retrospectively analyzed prospectively collected data from the UK National Health Service Breast Screening Programme during a 36-month period (April 1, 2005 to March 31, 2008), with 3-year follow-up in women aged 50–70 years. Data on recall, cancers detected at screening, and interval cancers were available for each of the 84 breast screening units and for each year ($n = 252$). The association between interval cancers and recalls was modeled by using Poisson regression on aggregated data and according to age (5-year intervals) and screening type (prevalent vs incident).

Results: The authors analyzed 5 126 689 screening episodes, demonstrating an average recall to assessment rate (RAR) of 4.56% (range, 1.64%–8.42%; standard deviation, 1.15%), cancer detection rate of 8.1 per 1000 women screened, and interval cancer rate (ICR) of 3.1 per 1000 women screened. Overall, a significant negative association was found between RAR and ICR (Poisson regression coefficient: -0.039 [95% confidence interval: -0.062 , -0.017]; $P = .001$), with approximately one fewer interval cancer for every additional 80–84 recalls. Subgroup analysis revealed similar negative correlations in women aged 50–54 years ($P = .002$), 60–64 years ($P = .01$), and 65–69 years ($P = .008$) as well as in incident screens ($P = .001$) and prevalent screens ($P = .04$). No significant relationship was found in women aged 55–59 years ($P = .46$).

Conclusion: There was a statistically significant negative correlation between RAR and ICR, which suggests the merit of a minimum threshold for RAR.

© RSNA, 2018

Online supplemental material is available for this article.

Breast cancer screening programs routinely specify maximum recall for assessment rates (RARs) after screening mammography to minimize harms like anxiety and interventions associated with false-positive findings. There is consensus that a maximum RAR threshold exists, above which diminishing benefits accrue relative to added harms. However, the concept of a minimal RAR, below which potential benefits are lost, has generated less attention. Rather than define a minimum threshold for the RAR (equivalent to “recall rate” in the United States), programs have used complementary metrics including a minimum threshold for cancer detection rates (CDRs), which ideally stimulates increased RAR, and a minimum threshold for positive predictive value of recall, which encourages both

increased detection rate and decreased recall rate (1–3). In contrast to CDR and positive predictive value, recall rate is a readily available and stable metric; thus, a minimum RAR threshold may be valuable as an early indication of screening performance. However, confirmation of a relationship between the RAR and interval cancer rate (ICR) is a prerequisite to understanding whether a minimal recall rate threshold exists, under which screening mammography benefits decrease.

Study of the relationship between RAR and outcomes like CDR has revealed mixed results. International comparisons show that recall rates, up to approximately 8%, correlate with increasing CDR (4). However, results have been contradictory at higher RARs (5–10). Although the

Abbreviations

CDR = cancer detection rate, DCIS = ductal carcinoma in situ, ICR = interval cancer rate, NHSBSP = National Health Service Breast Screening Programme, RAR = recall to assessment rate

Summary

This analysis confirms the association between increased recalls and decreased interval cancers supporting development of a minimum recall threshold.

Implication for Patient Care

There is consensus that a maximum recall threshold exists, above which diminishing benefits of screening mammography accrue relative to added harms; this study provides evidence that there is also likely a minimal recall threshold, below which some potential benefits of screening mammography are lost.

association between RAR and subsequent ICR is less well studied, it may be more important because it potentially captures the influence of RAR on cancers destined to become clinically manifest. Unlike a high CDR, to which length bias or overdiagnosis may contribute, the incidence of symptomatic interval cancers after a negative screening examination provides a measure reflecting the efficacy of the screening program itself. A retrospective reader study concluded that increasing the recall rate in the Dutch screening program from approximately 1% to 4% would significantly increase CDR and reduce ICR (11).

The RAR depends on many factors (12), such as prevalent versus incident screening round (2), patient population characteristics (13–15), and screening program factors (4,12,16–18). The literature illustrates that two-view mammography led to simultaneous increases in RAR and CDR (17,18) and decreases in ICR (19,20). However, to our knowledge, the association between RAR and ICR has not been quantified in a service screening setting in the interest of supporting a lower threshold for RAR.

To address this gap, we analyzed prospectively collected screening cohorts from 84 screening units in the National Health Service Breast Screening Programme (NHSBSP) to determine whether low levels of recall lead to increased interval cancers and the magnitude of this effect.

Materials and Methods

Because this study did not involve patient contact, intervention, or use of identifiable patient data, it was determined to be exempt from institutional review board approval in the United States and exempt from human subjects ethical review in the United Kingdom. We are not aware of recruitment of our study population to other studies.

Population

We constructed a cohort study by retrospectively analyzing prospectively collected data from the UK's NHSBSP during a 36-month period (April 1, 2005, to March 31, 2008) and outcomes for the 3-year intervals following screening in these years. Thus, for example, for women with negative results at screening in July 2007, we had interval cancer data to June

2010. The study was restricted to women who were between 50 and 70 years of age at the time of screening—the age group of women invited under the program. The number of women is likely slightly lower than the number of mammograms because some women may have received screening before the 36-month round length. However, this occurrence is not quantifiable in our anonymized dataset. The NHSBSP includes England, Wales, Northern Ireland, and Scotland. However, Scotland did not have data available on interval cancers, so it was excluded from our analysis. The NHSBSP invites women, based on age, to mammographic screening every 3 years and screens approximately 2 million women per year. The NHSBSP uses double and batch reading. Our study size was designed to capture a cohort of women undergoing consecutive screening mammography in this large service screening program over a single screening round. All mammograms were performed with the screen-film technique. Use of 3 years of screening plus the following 3 years of interval cancers would give a figure of more than 15 000 interval cancers, conferring sufficient power to detect very small dependencies (21).

Definitions

As in a previous report (22), interval cancers were defined as cancers diagnosed symptomatically in women within 36 months of their last screening examination. If a woman had been recalled but deemed to have negative or benign findings at assessment, a symptomatic cancer diagnosed in the subsequent 36 months was considered an interval cancer. The available datasets classify cancer subtypes as invasive, ductal carcinoma in situ (DCIS), or unknown. The DCIS categorization included DCIS with microinvasion. From the aggregate data, we were able to calculate RAR, CDR, and ICR—in each case as a proportion of the number of women with technically adequate screening mammograms.

Data Sources

We obtained data on recall and screening-detected cancers from the National Health Service Information Centre (now NHS Digital) in England and from counterparts in Wales and Northern Ireland by means of standard returns submitted by breast screening units. We procured the interval cancer data from the 11 regional Quality Assurance Reference Centres, which in turn received cancer notifications from the cancer registries.

Age and type of screening examination (prevalent or incident) are potential confounders in our analysis and prompted subgroup analysis. The KC62 form for each breast screening unit contains aggregated data for screening mammograms (and invitations for screening) for that unit. Each KC62 dataset contains eight separate tables (Table E1 [online]). For the subgroup analysis, we designated KC62 tables A, B, and E as prevalent screens and all other KC62 tables as incident screens. Each KC62 table is further divided into 10 age groups (≤ 44 years, 45–49 years, 50–52 years, 53–54 years, 55–59 years, 60–64 years, 65–69 years, 70 years, 71–74 years, and ≥ 75 years). For each of these subgroups, data include the number of women invited for screening, number of technically adequate screening mammograms,

number of women recalled for further assessment, number of cancers diagnosed, and cancer subtype.

Data recorded for each interval cancer included breast screening unit, patient age at screening (in years), date of screening, nature of screening before diagnosis (prevalent or incident), date of diagnosis, and cancer subtype. For 186 of the 15 867 interval cancers, the nature of the previous screening (ie, prevalent or incident) was not recorded; therefore, those in women screened at age 52 years or younger were assumed to be prevalent and those in women older than 52 years to be incident.

Statistical Analysis

We modeled the association between the ICR and the RAR by using Poisson regression with a random effect for screening unit and by using robust variance estimates (23,24), with number of interval cancers as the outcome variable, offset by the population screened, and RAR as the explanatory variable. In addition to overall analysis, we carried out subgroup analyses according to age and prevalent or incident status. Calculations were carried out with software (Stata, version 13.1; Stata, College Station, Tex) by using the `xtpoisson` command to fit a random effects model with cluster-robust estimates of the variance. These variables were used to calculate confidence intervals on the regression coefficients. $P < .05$ was considered to be indicative of a statistically significant difference.

In our subgroup analysis for age, we used the four 5-year age groups contained within the 50–70 year age range, namely 50–54, 55–59, 60–64, and 65–69 years. We omitted women aged 70 years at screening from this subgroup analysis.

To estimate the trade-off between RAR and ICR, we used the portion of the Poisson regression curve from the 25th percentile to the 75th percentile of the RAR to calculate the number of additional recalls per interval cancer avoided, if any.

Round length may act as a potential effect modifier in our analysis because repeat screening mammography in less than 36 months would decrease the ICR. Although the NHSBSP targets a maximum round length of 3 years, the actual interval between successive screening episodes varies from unit to unit, and some units achieve a round length of less than 36 months. To assess the effect of this round length variability and ensure greater comparability between units, we repeated the analysis using only those interval cancers diagnosed within the first 24 months after screening. To address the issue of dependence between the observations—in other words, the fact that the 252 observations represent three consecutive observations for each breast screening unit—we repeated our analysis on the level of breast screening unit (using 84 observations).

Units in geographical areas with higher underlying breast cancer incidence will tend to have both higher detection rates, which may be reflected in higher recall rates, and higher ICRs. This potential confounding might induce an artificial positive relationship between RAR and ICR. We therefore also carried out the primary Poisson regression adjusting for CDRs at screening. This would also adjust for varying reading sensitivity between units.

We performed formal interaction testing between age and RAR and between prevalent or incident status and RAR, with

age fitted as a four-level factor and formal significance testing with the likelihood ratio test. The subgroup analyses according to age and prevalent or incident status are important not only to quantify trade-offs associated with RAR levels in these different groups, but also in the case of interactions between RAR and these variables.

The use of data for all women participating in screening in three of the four nations of the United Kingdom suggests that issues of representativity and bias should be negligible. With national cancer registration, losses to follow-up should be minimal.

Results

Our analytic dataset, a cohort of 5 126 689 consecutive, technically adequate screening mammograms, is comprised of 4 994 570 mammograms prompted by 6 761 719 invitations and 132 119 mammograms initiated through self-referral or referral by the patient's general practitioner. The 2979 technically inadequate screening mammograms for which women did not report for repeat images were excluded from analysis. There were 84 screening units that read a median of 19 402 mammograms per year (range, 5751–46 349 mammograms). There were 57 191 breast cancers at follow-up, 41 324 were detected at screening and 15 867 were interval cancers. Of the 57 191 cancers, 47 309 (82.7%) were invasive and 9593 were DCIS (16.8%); the subtype was missing for 289 cancers (0.5%).

We found an average RAR of 4.56% (range, 1.64%–8.42%; standard deviation, 1.15%), average CDR of 8.1 per 1000 women screened, and average ICR of 3.1 per 1000 women screened. Prevalent compared with incident round mammograms (Table 1) demonstrated more than double the RAR, a similar CDR, and a lower ICR. Screening mammography outcomes according to age demonstrated that both CDR and ICR were lower in younger age groups as expected based on incidence. Screening-detected cancers demonstrated a higher proportion of DCIS (Table 2).

The primary analysis to estimate the association between ICR and RAR demonstrated a Poisson regression coefficient of -0.039 (95% confidence interval: -0.062 , -0.017), a significant negative association ($P = .001$) (Fig 1a). Adjustment for CDR, an estimate of incidence and a potential confounder, did not substantially change or diminish the significance of the association between ICR and RAR (Table E2 [online]). The Poisson regression coefficient is equivalent to the logarithm of the change in risk; thus, a coefficient of -0.039 indicates a 4% reduction in the risk of an interval cancer per unit increase in the RAR. CDR and RAR demonstrated an association with a Poisson regression coefficient of 0.043 (95% confidence interval: 0.027 , 0.058), a significant positive association ($P < .001$) (Fig 1b).

Subgroup analysis revealed similarly negative correlations between RAR and ICR in women aged 50–54 years (Poisson coefficient: -0.039 , $P = .002$) (Fig 2), 60–64 years (Poisson coefficient: -0.043 , $P = .01$) (Fig 2b), and 65–69 years (Poisson coefficient: -0.052 , $P = .008$) (Fig 2c). No significant correlation was found in women aged 55–59 years (Poisson coefficient: -0.396 , $P = .46$). A negative correlation persisted in incident screens (Poisson coefficient: -0.039 , $P = .001$)

Table 1: Findings at Screening Mammography according to Age and Screening Round

Age	Prevalent Screening Round						Incident Screening Round					
	No. of Women Screened	No. of Women Recalled*	No. of Cancers Detected at Screening [†]	No. of Interval Cancers within 12 Months [‡]	No. of Interval Cancers within 24 Months [‡]	No. of Interval Cancers within 36 Months [‡]	No. of Women Screened	No. of Women Recalled*	No. of Cancers Detected at Screening [†]	No. of Interval Cancers within 12 Months [‡]	No. of Interval Cancers within 24 Months [‡]	No. of Interval Cancers within 36 Months [‡]
50–54 y	767 962	65 797 (8.57)	5731 (7.46)	389 (0.51)	1225 (1.60)	2123 (2.76)	608 011	23 980 (3.94)	2988 (4.91)	296 (0.49)	935 (1.54)	1675 (2.75)
55–59 y	62 985	4896 (7.77)	650 (10.32)	35 (0.56)	125 (1.98)	218 (3.46)	1 375 748	46 762 (3.40)	9065 (6.59)	745 (0.54)	2412 (1.75)	4241 (3.08)
60–64 y	26 925	2123 (7.88)	357 (13.26)	25 (0.93)	72 (2.67)	128 (4.75)	1 233 655	43 769 (3.55)	10 520 (8.53)	619 (0.50)	2187 (1.77)	3947 (3.20)
65–69 y	16 463	1346 (8.18)	314 (19.07)	14 (0.85)	31 (1.88)	50 (3.04)	928 637	37 962 (4.09)	10 327 (11.12)	507 (0.55)	1705 (1.84)	3174 (3.42)
70 y	2 590	198 (7.64)	42 (16.22)	2 (0.77)	4 (1.54)	8 (3.09)	103 713	4831 (4.66)	1330 (12.82)	47 (0.45)	157 (1.51)	303 (2.92)
Overall (50–70 y)	876 925	74 360 (8.48)	7094 (8.09)	465 (0.53)	1457 (1.66)	2527 (2.88)	4 249 764	157 304 (3.70)	34 230 (8.05)	2214 (0.52)	7396 (1.74)	13 340 (3.14)

* Numbers in parentheses are the recall to assessment rates (in percentages).

[†] Numbers in parentheses are the cancer detection rates (per 1000 women screened).

[‡] Numbers in parentheses are the interval cancer rates (per 1000 women screened).

* Numbers in parentheses are the recall to assessment rates (in percentages).

† Numbers in parentheses are the cancer detection rates (per 1000 women screened).

‡ Numbers in parentheses are the interval cancer rates (per 1000 women screened).

(Fig 3a) and prevalent screens (Poisson coefficient: -0.019 , $P = .04$) (Fig 3b). Results of formal interaction tests between age and RAR and prevalent or incident status and RAR were not significant (Tables E3, E4 [online]).

Expressing these results as number of recalls required per interval cancer avoided, we estimated that 80–84 additional recalls are required to avoid one interval cancer in the aggregated dataset (Fig 1a). This trade-off varies in each of the statistically significant subgroups, from a low of 55–58 recalls per interval cancer in the 65–69-year age group (Fig 2c) to 89–96 recalls per interval cancer in the 50–54-year age group (Fig 2a). The incident round subgroup demonstrated 79–83 recalls per interval cancer (Fig 3a). The prevalent round subgroup required 176–186 recalls per interval cancer (Fig 3b).

The distribution of interval cancers during the 36-month round length demonstrates the expected pattern encountered in the UK screening program: between 0–12 months, ICR = 0.52 per 1000 women screened; between 12–24 months, ICR = 1.20 per 1000 women screened; and between 24–36 months, ICR = 1.37 per 1000 women screened. Of note, within the 36-month follow-up time frame, the number of interval cancers increases until approximately 30 months, at which time the number of cases levels off and subsequently decreases (Fig 4). This pattern likely signifies that women begin to undergo repeat screening examinations due to variable round length among breast screening units. We found that the negative correlation between ICR and RAR was preserved when we used interval cancers at 24 months rather than 36 months, with a Poisson regression coefficient of -0.040 (95% confidence interval: -0.067 , -0.013), a significant negative association ($P = .004$). Because no breast screening units have a round length of less than 24 months, this indicates that the variability of round lengths between breast screening units does not alter the conclusions reached.

When we repeated our analysis on the breast screening unit level, using 84 observations to eliminate dependencies, our results were not changed (Table E2 [online]).

Discussion

We showed a consistent and statistically significant association between increasing RAR and decreasing ICR in the NHSBSP population. These findings persist in most age groups and for both prevalent and incident rounds. Although this relationship demonstrates a steeper negative correlation in the upper age groups, the relationship persists in the youngest age group, in which a high proportion of screening studies reflect the prevalent round, but not in the 54–59-year age group, in which the majority of mammograms are the incident screens. In aggregate, 80–84 additional recalls are estimated to be required to avoid one interval cancer, a ratio that varied with regression coefficient at subgroup analysis based on age group and prevalent versus incident screens. A statistically significant positive association between CDR and RAR reinforces the findings of prior literature (5–9).

Table 2: Screen-detected or Interval Cancers according to Screening Round

Screening Round	Screen-detected Cancers				Interval Cancers within 36 Months			
	No. of Cancers	Invasive	DCIS	Unknown	No. of Cancers	Invasive	DCIS	Unknown
Prevalent	7094	5210 (73.4)	1863 (26.3)	21 (0.3)	2527	2370 (93.8)	124 (4.9)	33 (1.3)
Incident	34 230	27 188 (79.4)	6939 (20.3)	103 (0.3)	13 340	12 541 (94.0)	667 (5.0)	132 (1.0)
All	41 324	32 398 (78.4)	8802 (21.3)	124 (0.3)	15 867	14 911 (94.0)	791 (5.0)	165 (1.0)

Note.—Data are for women aged 50–70 years at time of screening. Numbers in parentheses are percentages. DCIS = ductal carcinoma in situ.

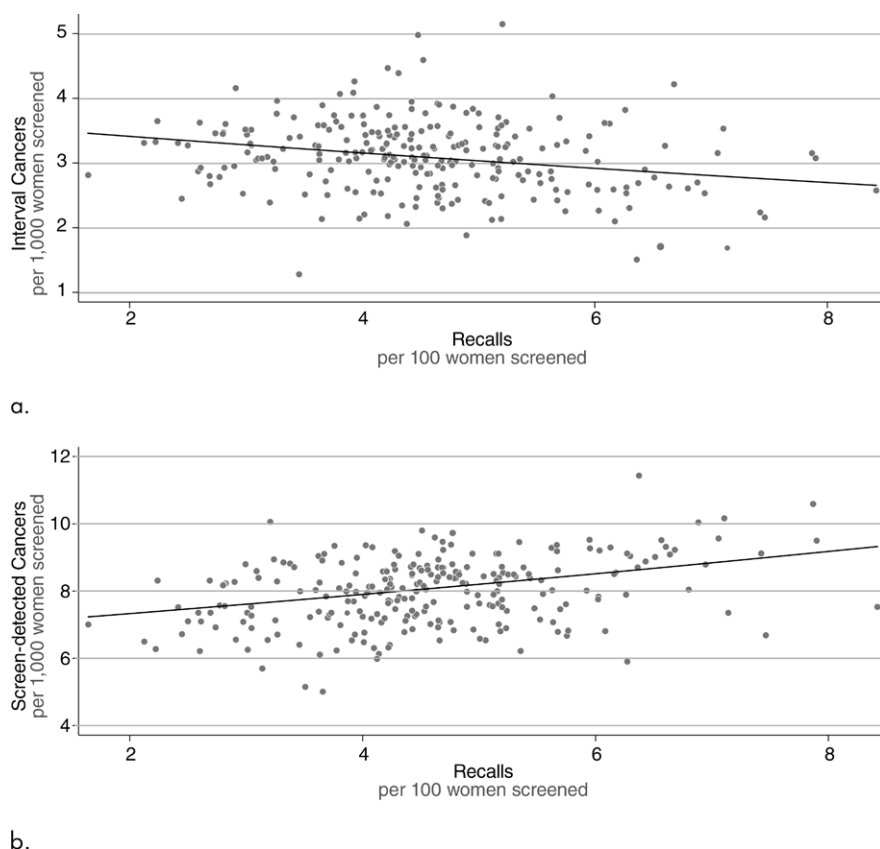


Figure 1: Plots demonstrate association between recall to assessment rate and **(a)** interval cancer rate and **(b)** screening cancer detection rate. Points represent a year of screening mammograms in one of the 84 breast screening units. Line represents Poisson regression line.

The literature (2,7,9), expert panels (25), and guidelines (26–28) all support the importance of establishing a maximum threshold for recall, a point at which false-positive findings outweigh the benefits accrued with early cancer detection. However, there has been fewer investigations into whether a minimal level of recall confers the opposite trade-off, recall too low to deliver the benefit of preclinical diagnosis. The ICR is one of the few metrics available to verify the success or failure of a program to detect cancers in the preclinical phase. In the United Kingdom, a RAR of less than 7% for prevalent screenings and less than 5% for incident screens are deemed to be “achievable” targets, which technically allow low recall rates as long as CDR and positive predictive value remain within acceptable levels (29). RAR

is rapidly available but ICR entails long reporting delays (22). Unlike RAR, the ICR, CDR, and positive predictive value all require large numbers to be reliable (30). Our demonstration of the association between increasing RAR and decreasing ICR supports the possibility of identifying a minimum RAR as a benchmark that is immediately available and statistically stable.

To our knowledge, this is the first investigation into the relationship between RAR and ICR in a large service screening program. However, the literature has examined other dimensions of RAR and its correlation with outcomes. Randomized controlled trials demonstrate a mortality benefit observed with similar RAR levels. The Swedish Two-County Trial showed that recall rates of 5% at the prevalent and 2.5% at the incident round translated into a mortality benefit for screened women (31). Our results reinforce a case-control reader study that demonstrates that increasing recall can decrease ICR (11).

Our analysis reveals that the regression coefficients are steeper, paralleling the decreasing recalls required per interval cancer avoided in older women and incident screens. This “better value” in

terms of in the RAR-ICR trade-off likely reflects several factors, including the underlying rate of disease and the current levels of RAR. Understanding the optimal RAR based on these interacting variables is beyond the scope of this study but is viewed as important future work. Furthermore, many studies demonstrate that RAR is related to population factors including age, breast density, and risk profile (eg, family history, breast biopsy, genetic predisposition, and other high-risk characteristics) (13–15) and to program characteristics including batch reading (16), the number of readers (17), and the number of images (18). Compared with the UK program, which uses a triennial round length and double and batch reading, other programs will encounter unique trade-offs depending on patient population

characteristics and breast cancer screening program implementation details. Important mammography screening differences have been documented between the United Kingdom and the United States (32,33). In the United Kingdom, screening offered with longer round length to narrower age ranges results in fewer false-positive findings but the detection of fewer cancers (DCIS and small invasive cancers) as compared with the United States. The United Kingdom has strict standards for interpretation, including uniform training and accreditation of mammography readers, high minimum screening interpretation volumes (>5000 mammograms per year), annual performance testing, required batch reading, and tightly regulated requirements for CDR and RAR. Double reading is also required, although the actual implementation details (eg, the conduct of arbitration) differ among breast screening units. The UK system overcomes a limited workforce of subspecialty-trained radiologists by including technologists and clinicians as readers. The United States augments its small pool of subspecialty-trained breast imagers with general radiologists. In the United States, although auditing processes are mandated, there are no enforced requirements for CDR and RAR. Minimum volume requirements in the United States are lower than those in the United Kingdom, at 960 screening mammograms every 2 years.

International comparisons including and beyond the United States and the United Kingdom (4,10,33) demonstrate wide variability in RAR between programs, which is unlikely only due to patient population and program differences but also involves the health policy setting (11,12), medical-legal environment (ie, the importance of finding every cancer), and the tolerance of false-positive findings (32,33). Importantly, our methods for determining the trade-offs of RAR levels in terms of ICR transcends these factors and enables any program to determine the trade-offs. The crucial ingredient is accurate collection of interval cancers on a program level, like the system in place in the United Kingdom (29,34,35), ideally in an ongoing way as other program factors evolve over time.

A strength of our study is that we analyzed complete data in a large service screening program that achieves high-quality audit measures by using rigorous established methods (1,2). A limitation of our study is that our data are aggregated and anonymized; thus, we do not have access to detailed

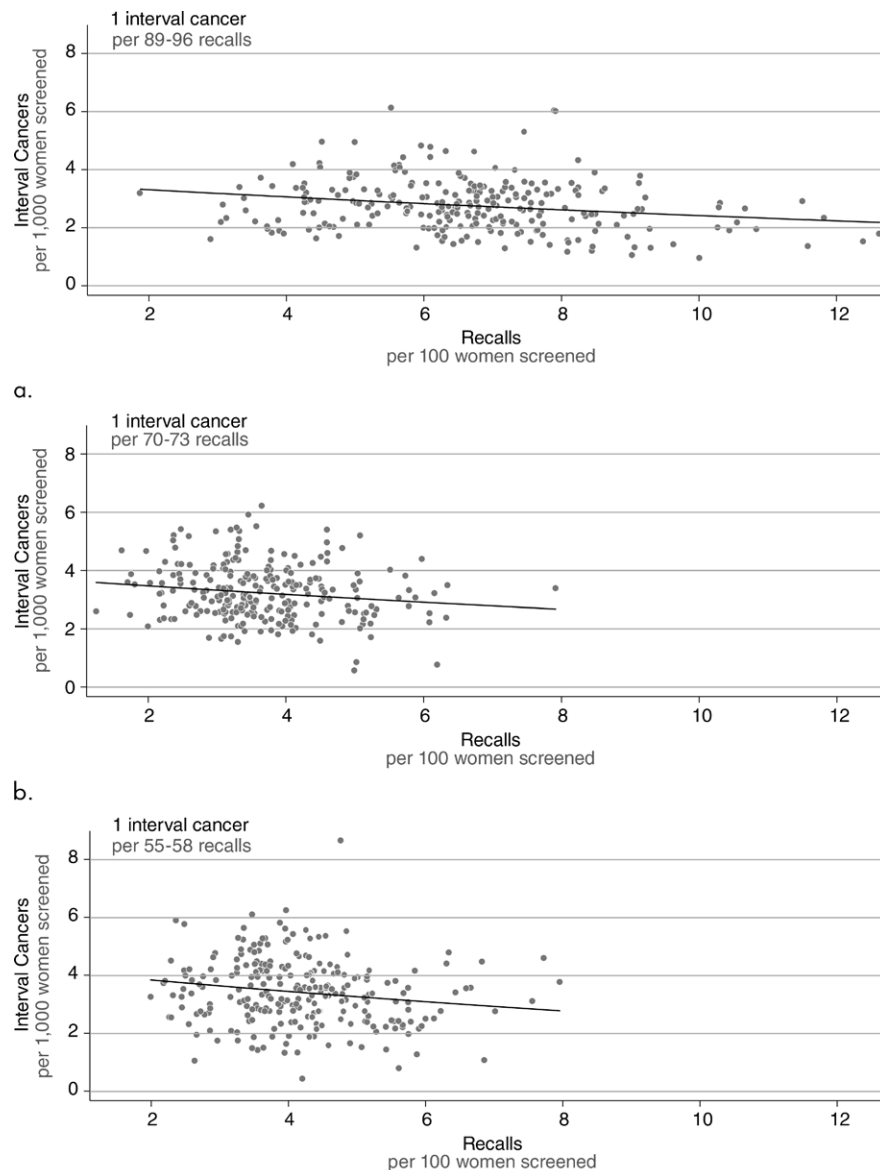


Figure 2: Plots illustrate statistically significant association between recall to assessment rate and interval cancer rate for age-based subgroup analysis of women aged (a) 50–54 years, (b) 60–64 years, and (c) 65–69 years. Points represent a year of screening mammograms in one of the 84 breast screening units. Line represents Poisson regression line. Trade-off between an interval cancer and a recall is noted on each graph.

demographic risk factors beyond age. As a result, we cannot perform subgroup analyses on variables that are known to affect RAR, such as breast density, family history of breast cancer, risk factors, interpretation style (double-reading conventions), radiologist experience, or type of imaging (analog vs digital mammography). Demographic risk factors for the years studied likely emulate those in the literature (36), which may facilitate comparison to other programs. We also do not have access to patient-level information; thus, we were not able to identify repeat patients or precisely specify round length for each woman (however, more recent data on round length suggest that approximately 27% of women screened in 2007–2008 might also have been screened in

2005–2006—that is, screened between 24 and 36 months). In the United Kingdom, radiologists do not report breast density. Therefore, this important variable, which influences both RAR and ICR, is not available for subgroup analysis. Of note, screening in the cohort reported herein took place before 2009, before the change to digital mammography in the United Kingdom. However, given the likely effect of the digital mammography conversion (37), we suggest that the relationship between RAR and ICR would remain after conversion. In the future, interval cancer data will become available from the digital epoch. Round length in the NHS targets a maximum of 36 months, prompted by invitation; therefore, a second mammogram before the 24-month timeframe for any given patient within the National Health Service is unlikely. In a small proportion of women, repeat screening between 24 and 36 months is possible owing to variable breast screening unit round length and may result in dependencies in the data, although this limitation is unlikely to alter our results. We also cannot determine the possibility of screening mammograms outside of the NHSBSP—thereby possibly erroneously labeling a screening-detected cancer (at outside screening) as an interval cancer. Private screening in the United Kingdom is rare, particularly in age ranges covered by the NHSBSP, and thus unlikely to alter our results. Finally, as discussed previously, our specific results regarding trade-offs may not be generalizable to other screening programs that may have different patient populations, program implementation details, or other characteristics that may affect how RAR and outcomes interact.

In conclusion, we established a statistically significant negative association of RAR with ICR in a well-established service screening program, thereby providing important evidence that establishing and enforcing a lower bound threshold for RAR is reasonable and potentially important. Although the specific details of the trade-offs are applicable to the UK NHSBSP alone, our methods provide an opportunity for other programs to determine this trade-off based on population and program characteristics. By establishing a relationship

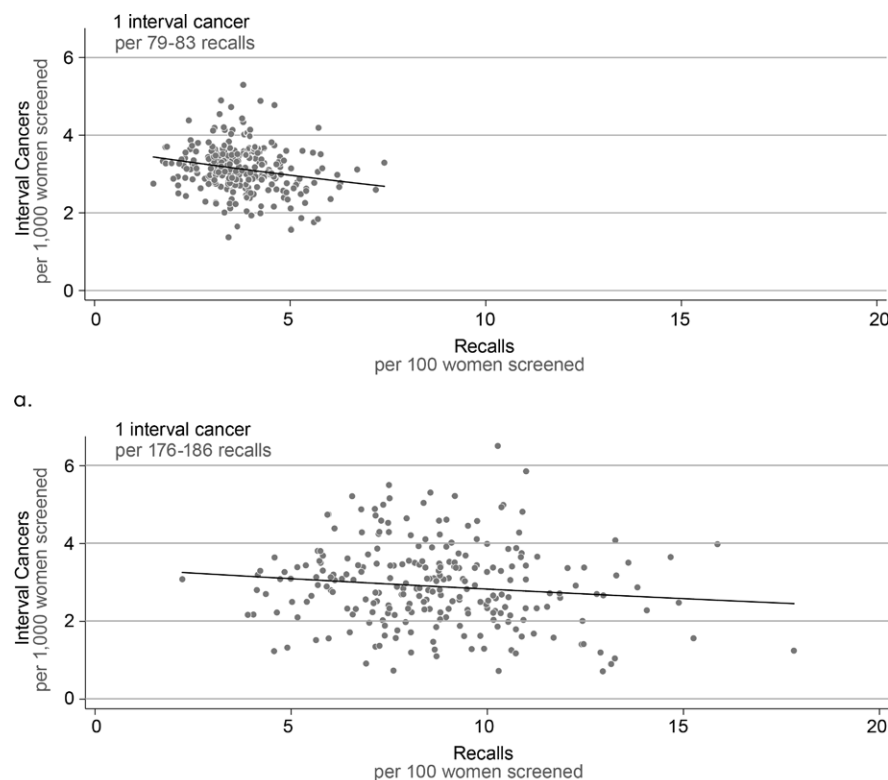


Figure 3: Plots illustrate statistically significant association between recall to assessment rate and interval cancer rate for (a) incident and (b) prevalent subgroup analysis. Points represent a year of screening mammograms in one of the 84 breast screening units. Line represents Poisson regression line. Trade-off between an interval cancer and a recall is noted on each graph.

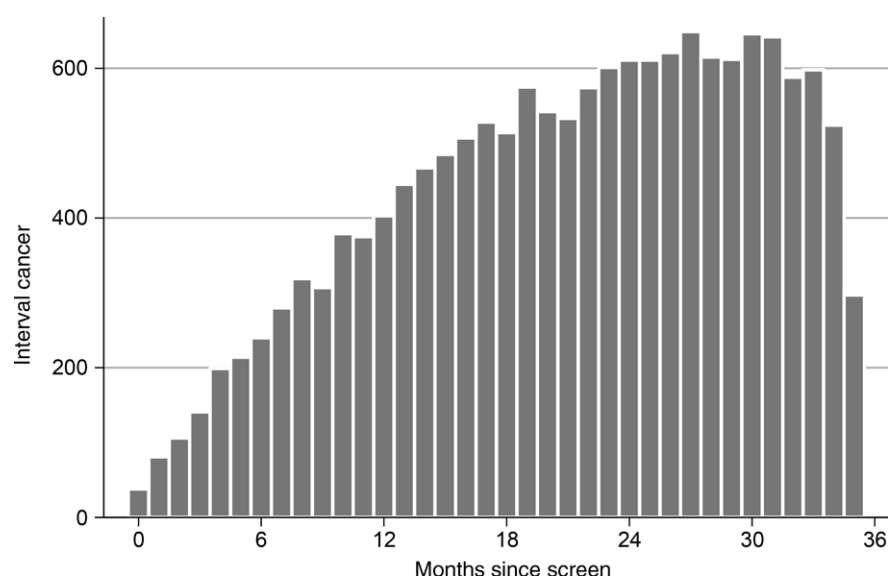


Figure 4: Bar graph shows number of interval cancers according to number of months since negative

between RAR and ICR and providing the methodology to document trade-offs, our results contribute a necessary foundation for determining a minimal RAR threshold that maximizes value for women undergoing breast cancer screening.

Author contributions: Guarantors of integrity of entire study, E.S.B., S.W.D.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, E.S.B., R.G.B.; clinical studies, E.S.B.; experimental studies, E.S.B.; statistical analysis, E.S.B., D.V., S.W.D.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: E.S.B. disclosed no relevant relationships. D.V. disclosed no relevant relationships. R.G.B. disclosed no relevant relationships. S.W.D. disclosed no relevant relationships.

References

- Bennett RL, Blanks RG. Should a standard be defined for the positive predictive value (PPV) of recall in the UK NHS Breast Screening Programme? *Breast* 2007;16(1):55–59.
- Blanks RG, Moss SM, Wallis MG. Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance between individual programmes using PPV-referral diagrams. *J Med Screen* 2001;8(1):24–28.
- Miglioretti DL, Ichikawa L, Smith RA, et al. Criteria for identifying radiologists with acceptable screening mammography interpretive performance on basis of multiple performance measures. *AJR Am J Roentgenol* 2015;204(4):W486–W491.
- Yankaskas BC, Klabunde CN, Ancelle-Park R, et al. International comparison of performance measures for screening mammography: can it be done? *J Med Screen* 2004;11(4):187–193.
- Grabler P, Sighoko D, Wang L, Allgood K, Ansell D. Recall and cancer detection rates for screening mammography: finding the sweet spot. *AJR Am J Roentgenol* 2017;208(1):208–213.
- Gur D, Sumkin JH, Hardesty LA, et al. Recall and detection rates in screening mammography. *Cancer* 2004;100(8):1590–1594.
- Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *AJR Am J Roentgenol* 2001;177(3):543–549.
- Yankaskas BC, Schell MJ, Miglioretti DL. Recall and detection rates in screening mammography. *Cancer* 2004;101(11):2710–2711; author reply 2711–2712.
- Schell MJ, Yankaskas BC, Ballard-Barbash R, et al. Evidence-based target recall rates for screening mammography. *Radiology* 2007;243(3):681–689.
- Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. *J Natl Cancer Inst* 2003;95(18):1384–1393.
- Otten JD, Karssemeijer N, Hendriks JH, et al. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. *J Natl Cancer Inst* 2005;97(10):748–754.
- Mohd Norsuddin N, Reed W, Mello-Thoms C, Lewis SJ. Understanding recall rates in screening mammography: a conceptual framework review of the literature. *Radiography* 2015;21(4):334–341.
- Cook AJ, Elmore JG, Miglioretti DL, et al. Decreased accuracy in interpretation of community-based screening mammography for women with multiple clinical risk factors. *J Clin Epidemiol* 2010;63(4):441–451.
- Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst* 2002;94(18):1373–1380.
- Lehman CD, White E, Peacock S, Drucker MJ, Urban N. Effect of age and breast density on screening mammograms with false-positive findings. *AJR Am J Roentgenol* 1999;173(6):1651–1655.
- Burnside ES, Park JM, Fine JP, Sisney GA. The use of batch reading to improve the performance of screening mammography. *AJR Am J Roentgenol* 2005;185(3):790–796.
- Blanks RG, Wallis MG, Moss SM. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. *J Med Screen* 1998;5(4):195–201.
- Blanks RG, Given-Wilson RM, Moss SM. Efficiency of cancer detection during routine repeat (incident) mammographic screening: two versus one view mammography. *J Med Screen* 1998;5(3):141–145.
- Dibden A, Offman J, Parmar D, et al. Reduction in interval cancer rates following the introduction of two-view mammography in the UK breast screening programme. *Br J Cancer* 2014;110(3):560–564.
- Seigneurin A, Exbrayat C, Labarère J, Colonna M. Comparison of interval breast cancer rates for two-versus single-view screening mammography: a population-based study. *Breast* 2009;18(5):284–288.
- Signorini DF. Sample size for Poisson regression. *Biometrika* 1991;78(2):446–450.
- Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. *Br J Cancer* 2011;104(4):571–577.
- Breslow N, Day N. Statistical methods in cancer research, vol 2. The design and analysis of cohort studies. Lyon, France: International Agency for Cancer Research, 1987.
- Breslow NE, Clayton DG. Approximate Inference in generalized linear mixed models. *J Am Stat Assoc* 1993;88(421):9–25.
- Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. *Radiology* 2010;255(2):354–361.
- Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis: fourth edition—summary document. *Ann Oncol* 2008;19(4):614–622.
- American College of Radiology. Breast imaging reporting and data system (BI-RADS). 5th ed. Reston, Va: American College of Radiology, 2014.
- Sardanelli F, Aase HS, Álvarez M, et al. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. *Eur Radiol* 2017;27(7):2737–2743.
- NHS Breast Screening Programme: consolidated standards. <https://www.gov.uk/government/publications/breast-screening-consolidated-programme-standards>. Published 2017. Accessed April 14, 2017.
- Burnside ES, Lin Y, Munoz del Rio A, et al. Addressing the challenge of assessing physician-level screening performance: mammography as an example. *PLoS One* 2014;9(2):e89418.
- Tabár L, Fagerberg G, Duffy SW, Day NE, Grøntoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiol Clin North Am* 1992;30(1):187–210.
- Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. *JAMA* 2003;290(16):2129–2137.
- Williams J, Garvican L, Tosteson AN, Goodman DC, Onega T. Breast cancer screening in England and the United States: a comparison of provision and utilisation. *Int J Public Health* 2015;60(8):881–890.
- Offman J, Duffy SW. Breast screening: interval cancer data, April 2003 to March 2005. <https://www.gov.uk/government/publications/breast-screening-interval-cancer-data-april-2003-to-march-2005>. Published 2012. Accessed April 14, 2017.
- Wilson R, Liston J. Quality assurance guidelines for breast cancer screening radiology: NHSBSP publication no 59. <https://www.gov.uk/government/publications/breast-screening-quality-assurance-standards-in-radiology>. Published 2011. Accessed April 14, 2017.
- Douglas E, Waller J, Duffy SW, Wardle J. Socioeconomic inequalities in breast and cervical screening coverage in England: are we closing the gap? *J Med Screen* 2016;23(2):98–103.
- van Luijt PA, Fracheboud J, Heijnsdijk EA, den Heeten GJ, de Koning HJ; National Evaluation Team for Breast Cancer Screening in Netherlands Study Group (NETB). Nation-wide data on screening performance during the transition to digital mammography: observations in 6 million screens. *Eur J Cancer* 2013;49(16):3517–3525.