The British Journal of Radiology, 85 (2012), e1226–e1232

Sensitivity and specificity of mammographic screening as practised in Vermont and Norway

[1,2]S HOFVIND, PhD, [3]B M GELLER, EdD, [4]J SKELLY, MSc and [4]P M VACEK, PhD

[1]*Cancer Registry of Norway, Oslo, Norway,* [2]*Faculty of Health Science, Oslo University College, Oslo, Norway,* [3]*Office of Health Promotion Research, University of Vermont, Burlington, VT, USA, and* [4]*Department of Medical Biostatistics, University of Vermont, Burlington, VT, USA*

Objective: The aim of this study was to examine the sensitivity and specificity of screening mammography as performed in Vermont, USA, and Norway.
Methods: Incident screening data from 1997 to 2003 for female patients aged 50–69 years from the Vermont Breast Cancer Surveillance System (116 996 subsequent screening examinations) and the Norwegian Breast Cancer Screening Program (360 872 subsequent screening examinations) were compared. Sensitivity and specificity estimates for the initial (based on screening mammogram only) and final (screening mammogram plus any further diagnostic imaging) interpretations were directly adjusted for age using 5-year age intervals for the combined Vermont and Norway population, and computed for 1 and 2 years of follow-up, which ended at the time of the next screening mammogram.
Results: For the 1-year follow-up, sensitivities for initial assessments were 82.0%, 88.2% and 92.5% for 1-, 2- and >2-year screening intervals, respectively, in Vermont ($p=0.022$). For final assessments, the values were 73.6%, 83.3% and 81.2% ($p=0.047$), respectively. For Norway, sensitivities for initial assessments were 91.0% and 91.3% ($p=0.529$) for 2- and >2-year intervals, and 90.7% and 91.3%, respectively, for final assessments ($p=0.630$). Specificity was lower in Vermont than in Norway for each screening interval and for all screening intervals combined, for both initial (90.6% *vs* 97.8% for all intervals; $p<0.001$) and final (98.8% *vs* 99.5% for all intervals; $p<0.001$) assessments.
Conclusion: Our study showed higher sensitivity and specificity in a biennial screening programme with an independent double reading than in a predominantly annual screening program with a single reading.
Advances in knowledge: This study demonstrates that higher recall rates and lower specificity are not always associated with higher sensitivity of screening mammography. Differences in the screening processes in Norway and Vermont suggest potential areas for improvement in the latter.

Received 6 November 2011
Revised 13 March 2012
Accepted 16 April 2012

DOI: 10.1259/bjr/15168178

In a previous study in which selected early outcome measures of mammographic screening in Vermont, USA, and Norway were compared, higher recall and interval cancer rates were shown for Vermont than for Norway. The rate of screen-detected cancers did not differ [1]. The findings were consistent with other international studies [2–4]. Different radiological reading procedures have been suggested as a possible reason for the findings [1, 2, 4].

Breast cancer screening involves a series of events that begins with the screening examination (bilateral two-view mammography), and may continue with a recall for diagnostic work-up. The diagnostic work-up may lead to a recommendation for a biopsy, which determines whether the suspect lesion is benign or malignant. In both Vermont and Norway, the decision to recall a female patient is based on the assessment of her initial screening mammogram. In the USA, single reading is the usual practice, while in Norway an independent double reading with consensus is performed, in accordance with the European guidelines [5]. In a single reading, a radiologist decides whether the female patient should be recalled for diagnostic work-up, while in an independent double reading with consensus, two radiologists discuss the findings and a consensus is reached as to whether to recall the patient. In both processes, a final assessment is reached after additional breast imaging (including ultrasound) to determine whether to recommend a biopsy.

We surmise that the different procedures for initial assessment will affect the sensitivity and specificity of both the initial and the final assessments. However, this can be difficult to ascertain when comparing countries that also have differing screening intervals. To better understand how differences in the interpretation procedures of screening mammography may influence cancer detection, we have taken a detailed look at the sensitivity and specificity of initial and final assessments in our

Address correspondence to: Professor Solveig Hofvind, Cancer Registry of Norway, N-0304 Oslo, Oslo and Akershus University College of Applied Sciences, Faculty of Health Science, N-0130 Oslo, Norway. E-mail: solveig.hofvind@kreftregisteret.no
This study was funded by the National Cancer Institute U01 CA070013 (P M Vacek, J Skelly, B M Geller).

previously studied cohort of female patients aged 50–69 years who underwent screening mammography in Vermont or Norway during 1997–2003. The aim of this study was to determine and compare the sensitivity and specificity of the initial and final assessments of mammographic screening as practised in Vermont and Norway.

Methods and materials

This study is based on data from the Vermont Breast Cancer Surveillance System (VBCSS) and the Norwegian Breast Cancer Screening Program (NBCSP). Both the VBCSS and the NBCSP collect data describing the assessments of the screening mammograms and any work-up imaging, patient characteristics and cancer outcomes. Use of the Vermont data was approved with an alteration of consent by the Institutional Review Board for the Protection of Human Subjects at the University of Vermont and is compliant with the Health Insurance Portability and Accountability Act. Collecting information from the screening programme in Norway was covered by the regulations on the collection and processing of personal health data at the Cancer Registry (Cancer Registry Regulations) [6]; because we received only aggregated data, no ethical committee or Data Inspectorate approval was necessary.

This study was based on the same data set used in a previous paper describing early outcome measures for screening mammography in Vermont and Norway [1]. Females aged 50–69 years with no history of breast cancer at a screening examination during 1997–2003 were included in the study. Only screening mammograms that took place subsequent to a previous screening were used. The screenings were classified by intervals (time since the previous screening) of 1 year (range, 10–19 months), 2 years (range, 20–27 months) or >2 years (>27 months). In the NBCSP, female patients were invited by letter to attend every 2 years, so there were no 1-year screenings. If a patient did not attend after her initial invitation and a reminder was sent, the next invitation was sent 2 years afterwards, so the interval to her next screening was approximately 4 years.

Study population

In Vermont, 45 050 female patients contributed 141 284 subsequent screening examinations. 116 996 screenings (83%) were performed at a 1-year interval, 13 982 (10%) at a 2-year interval and 10 306 (7%) at a >2-year interval. The average number of screenings for each patient was 3.1. Data from 744 subsequent screenings were available for initial assessment only; 727 patients either did not return for diagnostic work-up or received further care outside Vermont, while 17 had data from a diagnostic work-up but no final assessment was recorded. The number of screenings with a final assessment was therefore 140 540.

In Norway, 194 430 female patients contributed to 360 872 subsequent screening examinations, of whom 350 202 (97%) attended regularly, in response to an invitation approximately 2 years after the last scheduled appointment. A total of 10 670 (3%) female patients did not respond after one or several invitations and had an irregular interval of >2 years. The patients had an average of 1.9 screenings during the study period.

The characteristics of the study population have been described previously [1]. Briefly, approximately 95% of the patients in the Vermont and Norway study population were white. Vermont patients were younger at screening examination and a larger proportion had a college-level education; reached menarche at the age of 13 years or younger; and were younger than 20 years at their first birth compared with the Norwegian patients. The proportions of female patients in Vermont and Norway who reported ever using hormonal therapy were similar. The same study [1] found a significantly higher incidence of breast cancer (screen-detected and interval cancer) in the Vermont patients undergoing screening (4.0 per 1000 female-years) than in the Norway patients (3.4 per 1000 female-years).

Screening mammography in Vermont

Both the VBCSS and the NBCSP are described in detail elsewhere [1, 7, 8]. The VBCSS is funded by the National Cancer Institute in the USA as part of the Breast Cancer Surveillance Consortium [9]. Since 1994, the VBCSS has collected patient risk factors, breast imaging and breast pathology data for all breast imaging and breast biopsies performed in Vermont [10]. The data are combined with the breast cancer cases identified by the Vermont Cancer Registry and the New Hampshire Tumor Registry for nearly complete cancer follow-up.

All mammography facilities in Vermont are accredited by the US Food and Drug Administration, and operate under the rules and regulations of the Mammography Quality Standards Act [11]. Most Vermont female patients follow the recommendations of the US Preventive Services Task Force or the American Cancer Society, which recommend regular screening every 1 or 2 years from the age of 40 or 50 years [12, 13].

The standard procedure for radiological interpretation in Vermont is single reading. However, some facilities carried out double reading during the study time period [14], while others used computer aided detection (CAD), but these represent very few of the total number of mammograms (<20% were either double read or involved CAD use). The initial and final assessments of the screening mammograms were based on the Breast Imaging Reporting and Data Systems (BI-RADS) of the American College of Radiology [15].

Screening mammography in Norway

The NBCSP is a governmentally organised, governmentally funded, population-based, nationwide screening programme run by the Cancer Registry of Norway [7]. All information regarding screening examinations and interpretations is electronically transferred to the database at the Cancer Registry. The procedures used in the recall examination are reported online. Breast cancers are identified through the Cancer Registry, which is 99% complete for solid tumours [16]. The target population is

approximately 540 000 female patients aged 50–69 years who are biennially invited for screening. The patients receive a pre-scheduled time and place for the screening examination.

The NBCSP performs independent double readings with consensus. The procedure involves two radiologists who read the initial screening mammograms independently according to a five-point interpretation scale, reflecting the probability of cancer. The final decision to recall a patient is made in a consensus meeting in which a third radiologist may be asked to help render a recall decision. The NBCSP does not recommend short-term follow-up.

Sensitivity and specificity

Sensitivity and specificity were calculated for both the initial and the final assessments. The initial assessment refers to the interpretation of the screening views only (mediolateral–oblique and craniocaudal), while the final assessment refers to the interpretation based on the screening views as well as any diagnostic work-up that may have been done, including additional mammographic imaging and/or ultrasound. A biopsy was not considered to be a part of the diagnostic work-up in this study. This definition was adopted to enable the sensitivity and specificity of the initial interpretation of the screening mammograms in Norway and Vermont to be compared despite procedural differences; while a biopsy is performed as a part of the diagnostic work-up and performed at the same appointment as additional imaging and ultrasound in Norway, the procedure is usually performed at a separate visit in Vermont.

The criteria for classifying a screening mammogram as negative or positive are standardised in the USA and Europe [5, 15, 17]. For Vermont, BI-RADS scores of 1, 2 and 3 on an initial screening mammogram were considered negative, while scores of 0, 4 and 5 were considered positive. For Norway, the initial assessment was defined as negative if the patient was not recalled for diagnostic work-up and positive if she was called back. A final assessment was considered negative in Vermont if after the additional breast imaging it was interpreted as BI-RADS Category 1, 2 or 3, and positive if it was interpreted as BI-RADS Category 4 or 5. BI-RADS 3 classification included a repeated mammogram 6 months later which was negative. If information about the final assessment was missing, the case was excluded. In the NBCSP, a biopsy is normally considered as a part of the diagnostic work-up. For this study, a negative final assessment was defined as a diagnostic work-up without biopsy. The final assessment was defined as positive if a biopsy was needed to make a conclusion of the finding. Virtually all female patients participating in the screening programme in Norway show up for diagnostic work-up, which is provided free of charge. These definitions were adopted to maximise comparability; referral for additional imaging and subsequent biopsy indicates a positive interpretation of the mammogram in Norway. In the USA, patients with BI-RADS assessments of 4 or 5 for the initial interpretation are invariably referred for diagnostic imaging, while final assessments are referred for biopsy. We did not use these subsequent procedures to define a positive assessment in Vermont because, unlike in Norway, not all patients return for the recommended diagnostic work-up.

A positive initial or final assessment was defined as true positive (TP) if it was followed by a biopsy that confirmed breast cancer and as false positive (FP) if no breast cancer was diagnosed within a specified follow-up time. A negative initial or final assessment was defined as false negative (FN) if a breast cancer was subsequently diagnosed within a specified follow-up time and as true negative (TN) if it was not. Sensitivity (TP/TP+FN) and specificity (TN/TN+FP) were calculated using both 1- and 2-year follow-ups for cancer detection. In both cases, follow-up was ended at the time of either a breast cancer diagnosis or the next screen if it occurred before the specified follow-up time. The latter was done because it is impossible to know when a cancer detected at the next screening would have been diagnosed had the screening not occurred, as well as to avoid overlapping follow-up periods for sequential screenings. However, this biased the Vermont results for the 2-year follow-up by overestimating sensitivity.

Statistical analysis

All sensitivity and specificity estimates were directly adjusted for age using 5-year age intervals and adjusted to the age distribution for the combined Vermont and Norway population. The age at the time of the screening mammogram was used for the age classification. Because the usual screening interval differed in the two countries, sensitivity and specificity were computed based on both 1 and 2 years of follow-up, ending at the time of the next screening mammogram. Power calculations were based on the number of cancers identified during 1 year of follow-up. The results indicated that at $p=0.05$ there was an 80% chance of detecting a difference of 5% or more in sensitivity of the initial screening interpretation for a 2-year screening interval, a difference of 11% for a >2-year screening interval and a difference of 4% overall. Statistical power was the same for detecting differences in the final assessment and was somewhat higher when 2 years of follow-up were used to identify cancers. Associations between mammography outcome (TP, FN, TN or FP) and age group were examined using χ^2 tests. For each outcome, logistic regression was used to determine whether there was an interaction between age group and country. Logistic regression was also used to assess differences in sensitivity and specificity between Vermont and Norway and between screening intervals after adjustment for age. A p-value ≤ 0.05 was considered statistically significant. The analyses were conducted using SAS v. 8 (SAS Institute, Inc., Cary, NC).

Results

For both initial and final assessments, the distribution of TN, FP, TP and FN were significantly related to age group in both Vermont and Norway ($p<0.001$; Table 1). There was no significant interaction between age group and country, indicating that the effect of age on outcome did not differ between the two countries.

Table 1. The distribution of true-negative (TN), false-positive (FP), true-positive (TP) and false-negative (FN) findings for initial and final assessments for 1 year of follow-up, in opportunistic screening in Vermont and organised screening in Norway, 1997–2003, by 5-year age groups

	Vermont					Norway				
Age at screening (years)	Screens (*n*)	TN (%)	FP (%)	TP (%)	FN (%)	Screens (*n*)	TN (%)	FP (%)	TP (%)	FN (%)
	Initial assessment									
50–54	38 012	89.6	10.0	0.28	0.07	77 022	97.1	2.5	0.41	0.06
55–59	43 098	89.8	9.6	0.46	0.09	120 898	97.1	2.3	0.52	0.04
60–64	32 894	90.4	9.0	0.53	0.08	91 309	97.4	2.0	0.55	0.05
65–69	27 280	91.0	8.4	0.49	0.10	71 643	97.5	1.9	0.56	0.05
Total	*141 284*	*90.1*	*9.4*	*0.43*	*0.08*	*360 872*	*97.2*	*2.2*	*0.51*	*0.05*
	Final assessment									
50–54	37 784	98.4	1.2	0.25	0.09	77 022	99.0	0.59	0.41	0.06
55–59	42 873	98.3	1.2	0.41	0.13	120 898	98.9	0.56	0.52	0.05
60–64	32 721	98.3	1.1	0.46	0.13	91 309	98.9	0.50	0.55	0.05
65–69	27 162	98.2	1.2	0.43	0.15	71 643	98.9	0.45	0.56	0.05
Total	*14 054*[a]	*98.3*	*1.2*	*0.38*	*0.12*	*360 872*	*99.5*	*0.53*	*0.51*	*0.05*

[a]Information about the final assessment was missing for 744 screenings.

For Vermont, age-adjusted sensitivity for initial assessments was 82.0% for the 1-year interval and 1-year follow-up, 88.2% for the 2-year interval and 92.5% for the >2-year interval ($p=0.022$; Table 2). For final assessments the sensitivity was 73.6% for the 1-year interval, 83.3% for the 2-year interval and 81.2% for the >2-year interval ($p=0.047$). 1-year screening intervals are not performed in the NBCSP, but the values for initial assessment were 91.0% for the 2-year interval and 91.3% for the >2-year interval ($p=0.529$). For the final assessments, the sensitivities were 90.7% and 91.3% for the 2- and >2-year screening intervals, respectively.

The sensitivity of the initial assessments was statistically significantly lower in Vermont than in Norway for all screening intervals combined (83.8% *vs* 91.0%, respectively; $p<0.001$), but did not differ for the 2-year screening interval (88.2% *vs* 91.0%, respectively; $p=0.328$) or for the >2-year screening interval (92.5% *vs* 91.3%, respectively; $p=0.731$). For the final assessments, sensitivity was also lower in Vermont than in Norway for all screening intervals combined (75.6% *vs* 90.6%, respectively; $p<0.001$), as well as for the 2-year interval (83.3% *vs* 90.7%, respectively; $p=0.012$). Specificity was statistically significantly lower in Vermont than in Norway for each screening interval, as well as for all screening intervals combined, both for initial (90.6% *vs* 97.8% for Vermont and Norway, respectively, for all intervals; $p<0.001$) and final assessments (98.8% *vs* 99.5% for Vermont and Norway, respectively, for all intervals; $p<0.001$).

Sensitivity in both Vermont and Norway was lower when it was calculated based on cancers that occurred within 2 years following the screening if they were not preceded by another screening (Table 3). Sensitivity of the initial and final assessments was associated with the length of screening interval, in both Vermont and Norway. Sensitivity of the initial assessments was higher in Vermont than in Norway for the >2-year screening interval (89.2% *vs* 79.0%, respectively; $p=0.050$), but did not differ for the 2-year screening interval ($p=0.161$) or for all screening intervals combined ($p=0.367$). For the final assessments, sensitivity did not differ significantly between Vermont and Norway for either the 2-year screening interval (74.3% *vs* 74.4%, respectively; $p=0.804$) or the >2-year screening interval (75.2% *vs* 79.1%, respectively; $p=0.794$), but was significantly lower in Vermont than in Norway for all intervals combined (67.6% *vs* 74.6%, respectively; $p<0.001$). Screening interval did not influence specificity in either Vermont or Norway. The specificity of both the initial and the final

Table 2. Age-adjusted sensitivity and specificity for initial and final assessments for 1 year of follow-up, by screening interval

Screening interval	Vermont				Norway[a]			
	Cancer (*n*)	Sensitivity (%)	No cancer (*n*)	Specificity (%)	Cancer (*n*)	Sensitivity (%)	No cancer (*n*)	Specificity (%)
	Initial assessment							
1 year	545	82.0	116 451	90.8	–	–	–	–
2 years	76	88.2	13 906	89.9	1915	91.0	348 287	97.8[b]
>2 years	107	92.5	10 199	89.4	111	91.3	10 559	97.2[b]
All intervals	*728*	*83.8*	*140 556*	*90.6*	*2026*	*91.0*[b]	*358 846*	*97.8*[b]
	Final assessment							
1 year	533	73.6	115 872	98.9	–	–	–	–
2 years	73	83.3	13 824	98.6	1915	90.7[b]	348 287	99.5[b]
>2 years	104	81.2	10 134	98.5	111	91.3	10 559	99.3[b]
All intervals	*710*	*75.6*	*139 830*	*98.8*	*2026*	*90.6*[b]	*358 846*	*99.5*[b]

Information about the final assessment was missing for 744 screenings.
[a]1-year screening intervals do not take place in Norway.
[b]Significantly different from Vermont: $p<0.05$ for a two-sided test.

Table 3. Age-adjusted sensitivity and specificity for initial and final assessments for 2 years of follow-up, by screening interval

Screening interval	Vermont				Norway[a]			
	Cancer (*n*)	Sensitivity (%)	No cancer (*n*)	Specificity (%)	Cancer (*n*)	Sensitivity (%)	No cancer (*n*)	Specificity (%)
	Initial assessment							
1 year	623	74.9	116 373	90.8	–	–	–	–
2 years	87	82.5	13 895	89.9	2340	75.7	347 862	97.8[b]
>2 years	119	89.2	10 187	89.4	129	79.0[b]	10 541	97.2[b]
All intervals	*829*	*77.3*	*140 455*	*90.6*	*2469*	*75.8*	*358 403*	*97.8*[b]
	Final assessment							
1 year	610	65.4	115 795	98.9	–	–	–	–
2 years	84	74.3	13 813	98.6	2340	74.4	347 862	99.5[b]
>2 years	115	75.2	10 123	98.5	129	79.1	10 541	99.3[b]
All intervals	*809*	*67.6*	*139 731*	*98.8*	*2469*	*74.6*[b]	*358 403*	*99.5*[b]

Information about the final assessment was missing for 744 screenings.
[a]1-year screening intervals do not take place in Norway.
[b]Significantly different from Vermont: $p<0.05$ for a two-sided test.

assessments with 2 years of follow-up was statistically significantly lower in Vermont than in Norway for the 2-year and >2-year intervals, and for all screening intervals combined.

Discussion

This study indicates that sensitivity and specificity of mammographic screening are influenced by screening interval and length of follow-up. For the 1-year follow-up, the sensitivity of initial and final assessments was lower in Vermont than in Norway for all intervals combined, but for the 2-year follow-up, the sensitivity for initial assessments did not differ. Specificity was lower for Vermont for all screening intervals, for both 1 and 2 years of follow-up.

Sensitivity was higher in Norway than in Vermont for the 1-year follow-up across all screening intervals for both initial and final assessments, although the difference was statistically significant only for the 2-year screening interval and all intervals combined for the final assessment. The lower sensitivity in Vermont may be because most female patients in Vermont return after 1 year for a rescreen and the shorter screening interval is known to reduce sensitivity [17]. With 2-year screening intervals, cancers should be easier to detect because they are larger and more clearly defined than at 1 year. However, the mean and median tumour size for screen-detected cancers in Vermont and Norway do not differ [1]. The optimal screening interval has been investigated and a study by White et al [18] found no evidence that American female patients aged ≥50 years who undergo biennial mammography screening have an increased risk of late-stage breast cancer compared with patients who undergo annual screening. However, results from other studies are inconsistent regarding the relative advantages of 1-year *vs* 2-year screening intervals [19, 20], and a study from Finland estimated the effectiveness of intensive screening with poor attendance to be the same as that of infrequent screening with a high attendance rate [21]. It is obvious that finding an optimal screening interval in an organised screening setting is a complex issue, owing to individual characteristics and social conditions.

For the 2-year follow-up for cancer, Vermont had a similar or higher sensitivity to Norway for the initial assessment. Sensitivity for the final assessments for the 2-year screening interval was almost the same as in Norway. Although this implies that both Vermont and Norway achieved similar sensitivity despite differences in their processes for interpreting mammograms, this may not be the case. When a female patient returns for another screening, follow-up on her previous mammogram ceases and there is no further opportunity for it to be classified as an FN. Because most Vermont patients return after 1 year, there are limited data for 2 years of follow-up. Results from our previous study comparing cancer detection rates in Vermont and Norway showed that interval cancer rates are higher in Vermont at all time points during 2 years of follow-up [1] . This implies that sensitivity in Vermont would be lower than in Norway if all mammograms had 2 years of follow-up. The lower specificity in Vermont than in Norway did not improve sensitivity in the final assessment. Rather, the overall sensitivity of the final assessments was significantly lower in Vermont than in Norway, primarily owing to the lower sensitivity with 1-year screening intervals. This difference in sensitivity was more pronounced when only 1 year of follow-up was used because most interval cancers in Norway are detected during the second year following screening.

The high number of FN screening examinations in Vermont produced a statistically lower specificity than in Norway for initial assessments for all screening intervals. Even after the additional imaging was completed and a final assessment was rendered, Vermont still had a significantly lower proportion of TN examinations than Norway. This leads to a higher proportion of recommendations for biopsy and a higher proportion of biopsies for benign lesions in Vermont than in Norway.

The lower recall rate and higher specificity in Norway may be due to two programmatic differences. In Norway all screening mammograms are independently double read, resolving disagreements through consensus. Studies have shown that independent double reading may reduce or increase recall rate depending on how disputes are resolved [14, 22, 23]. Resolution with a consensus meeting tends to reduce recall rates because both radiologists need to agree that the finding is

important enough to request additional imaging. Another reason for the higher specificity in Norway may be because the desirable goal for recall is <3% for subsequent screenings, while in the USA the BI-RADS desirable goal is 5–10% for recall and >90% for specificity [5, 15]. Currently, Norway does not have a defined goal for specificity for initial and final assessments. The lower sensitivity of the final *vs* the initial assessments in Vermont suggests that the Vermont radiologists err on the side of caution by sending patients for additional imaging, which may be because of concerns about missing cancers or about medical malpractice. This inflates the sensitivity of the initial assessment. The final assessment is therefore a better measure of the radiologists' accuracy.

Several studies have also shown that double reading increases the sensitivity of mammography screening by 5–15%, depending on the method used to resolve disputes and the skill of the reading radiologists [5]. If double reading were used in Vermont it might not improve the final assessment sensitivity because the sensitivity drops from the initial to the final assessment, the point in the process where double reading is not used. Anecdotally, however, radiologists who double read with consensus say that the process is educational and therefore improves their accuracy. Also, if there were fewer recalled cases the proportion of cancers among the recalled cases would presumably increase and the number of cancers that are dismissed after additional imaging (FPs) may be reduced. The European guidelines recommend independent double reading with consensus, or arbitration from a third radiologist [5], which is how the initial assessment is performed in Norway.

Both the initial and final assessments provide valuable insight into how to improve mammography accuracy. At the initial screening views, cancers may be missed because of inaccurate visual perception or misinterpretation of a finding [10], and a study from Norway found that 24% of the screen-detected cases were interpreted as positive by only one of the two readers [24]. Generally, after additional imaging, cancers are missed because of misinterpretation. Therefore, calculating outcome audits with both the initial and the final assessment helps to identify which part of the screening process may need improvement. This is the recommendation of both the American College of Radiology's Breast Imaging Reporting and Data Systems Atlas [15] and the European guidelines [5].

Because there are many differences between the way screening is carried out in Vermont and Norway, in addition to single and double reading, it is difficult to deduce what may be responsible for Norway's better interpretative performance. In Norway, it is recommended that radiologists read 5000 screening mammograms per year, while the Mammography Quality Standards Act in the USA requires radiologists to read only 960 mammograms every 2 years. It is known that only 40% of the radiologists in Norway reach that standard [20], but the radiologists there are salaried, while most US radiologists are reimbursed per procedure [25]. The USA has a high rate of malpractice suits for missing breast cancer [26], while this is likely to be low in Norway. Independent double readings have been shown to improve performance [27, 28], while results from studies about reading volume and sensitivity are inconsistent [29, 30]. As far as we know, there are no studies regarding type of payment and accuracy.

There were several challenges in making a comparison between the two different healthcare delivery systems. However, our study was able to address most of these challenges. To make valid comparisons of sensitivity and specificity in this study, we used the same definition of TP, FN, TN and FP for both sites. In addition, because both sensitivity and specificity are influenced by whether a prevalent or subsequent mammogram is being evaluated and by a personal history of breast cancer, we defined our study population as only subsequent screenings in female patients with no history of breast cancer. All the estimates are age-adjusted and calculated for screens performed at equal intervals and with the same follow-up time.

There are several limitations to our study. Mammographic density decreases sensitivity [31], but because Norway routinely collects breast density measures in only recalled female patients, we were unable to adjust for this factor. However, we have no reason to believe that the mammographic densities of comparably aged patients in Norway and Vermont are different.

Our comparisons are limited by the different screening intervals in the two countries: most of the Vermont patients return annually for screening mammograms, whereas none of the Norwegian patients is invited for annual mammograms. Also, the defined follow-up periods are somewhat artificial for both countries, because patients return for screening at different intervals. The European guidelines have a slightly different definition for sensitivity than the one we used in this study, so we caution European investigators to be aware of this difference when comparing our results with their own [5]. The European guidelines include biopsy procedures as a part of the final assessment ([5], page 52), so they will have a slightly higher number of FN examinations if the biopsy procedure is an FN. This is a rare occurrence, so this should not decrease the sensitivity very much. Lastly, the multiple comparisons that have been made between the two data sets in this study increased the probability of observing a difference between the programmes by chance.

In summary, the goal of mammographic screening is to recall as few female patients as possible while missing as few cancers as possible, yielding a high specificity and a high sensitivity. Accuracy of mammography depends on visually perceiving an abnormality on the screening mammogram and interpreting that finding as either potentially malignant or not. In our study, the Norwegian programme was more likely to discriminate between cancer and no-cancer findings from the initial screening examination, and after two independent readings and a consensus conference. Future research could focus on whether it is the individual radiologist's skills, which may be enhanced by increased volume and training, or the process of independent double reading with consensus that leads to improved accuracy.

Conflict of interest

S Hofvind was employed by the Cancer Registry of Norway.

References

1. Hofvind S, Vacek P, Skelly J, Weaver D, Geller B. Comparing screening mammography for early breast cancer detection in Vermont and Norway. J Natl Cancer Inst 2008;100:1082–91.
2. Smith-Bindman R, Ballard-Barbash R, Miglioretti DL, Patnick J, Kerlikowske K. Comparing the performance of mammography screening in the USA and the UK. J Med Screen 2005;12:50–4.
3. Yankaskas BC, Klabunde CN, Ancelle-Park R, Renner G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? J Med Screen 2004;11:187–93.
4. Hofvind S, Yankaskas BC, Bulliard JL, Klabunde CN, Fracheboud J. Comparing interval breast cancer rates in Norway and North Carolina: results and challenges. J Med Screen 2009;16:131–9.
5. Perry N, Broeders MJ, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. European Communities. 2006 [accessed June 2011]. Available from: http://ec.europa.eu/health/ph_projects/2002/cancer/fp_cancer_2002_ext_guid_01.pdf
6. Regulations on the collection and processing of personal health data in the Cancer Registry of Norway (Cancer Registry Regulations). December 2001 [accessed June 2011]. Available from: www.ub.uio.no/ujur/ulovdata/for-20011221-1477-eng.doc
7. Hofvind S, Geller B, Vacek P, Thoresen S, Skaane P. Using the European Guidelines to evaluate the Norwegian Breast Cancer Screening Program. Eur J Epidemiol 2007;22:447–55.
8. Geller B, Worden J, Ashley J, Oppenheimer R, Weaver D. Multipurpose statewide breast cancer surveillance system: The Vermont experience. J Registry Mgmt 1996;23:168–74.
9. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol 1997;169:1001–8.
10. Geller BM, Barlow WE, Ballard-Barbash R, Ernster VL, Yankaskas BC, Sickles EA, et al. Use of the American College of Radiology BI-RADS to report on the mammographic evaluation of women with signs and symptoms of breast disease. Radiology 2002;222:536–42.
11. Food and Drug Administration. Quality Mammography Standards. Final Rules-21 CRF Parts 16 and 900 (docket No. 95N-0192) RIN 0910-AA24 edition. Washington, DC: Department of Health and Human Services; 1997.
12. US Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2009;151:716–26.
13. Smith RA, Cokkinides V, Brooks D, Saslow D, Brawley OW. Cancer screening in the United States, 2009: a review of current American Cancer Society guidelines and issues in cancer screening. CA Cancer J Clin 2010;60:99–119.
14. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol 2003;180:1461–7.
15. D'Orsi C, Bassett L, Berg W, Feig SA, Jackson JA, Kopans D, et al. Breast imaging and reporting data system—mammography: Illustrated BI-RADS. 4th ed. Reston, VA: American College of Radiology; 2003.
16. Larsen IK, Småstuen M, Johannesen TB, Langmark F, Parkin DM, Bray F, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. Eur J Cancer 2009;45:1218–31.
17. Rosenberg RD, Yankaskas BC, Hunt WC, Ballard-Barbash R, Urban N, Ernster VL, et al. Effect of variations in operational definitions on performance estimates for screening mammography. Acad Radiol 2000;7:1058–68.
18. White E, Miglioretti DL, Yankaskas BC, Geller BM, Rosenberg RD, Kerlikowske K, et al. Biennial versus annual mammography and the risk of late-stage breast cancer. J Natl Cancer Inst 2004;96:1832–9.
19. Miltenburg GA, Peeters PH, Fracheboud J, Collette HJ. Seventeen-year evaluation of breast cancer screening: the DOM project, The Netherlands. Diagnostisch Onderzoek (investigation) Mammacarcinom. Br J Cancer 1998;78:962–5.
20. Kaas R, Hart AA, Besnard AP, Peterse JL, Rutgers EJ. Impact of mammographic interval on stage and survival after the diagnosis of contralateral breast cancer. Br J Surg 2001;88:123–7.
21. Wu JC, Hakama M, Anttila A, Yen AM, Malila N, Sarkeala T, et al. Estimation of natural history parameters of breast cancer based on non-randomized organized screening data: subsidiary analysis of effects of inter-screening interval, sensitivity, and attendance rate on reduction of advanced cancer. Breast Cancer Res Treat 2010;122:553–66.
22. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population based mammography screening program. Radiology 1994;191:241–4.
23. Ciatto S, Del Turco MR, Morrone D, Catarzi S, Ambrogetti D, Cariddi A, et al. Independent double reading of screening mammograms. J Med Screen 1995;2:99–101.
24. Hofvind S, Geller BM, Rosenberg RD, Skaane P. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology 2009;253:652–60.
25. Farria DM, Schmidt ME, Monsees BS, Smith RA, Hildebolt C, Yoffie R, et al. Professional and economic factors affecting access to mammography: a crisis today, or tomorrow? Results from a national survey. Cancer 2005;104:491–8.
26. Elmore JG, Taplin SH, Barlow WE, Cutter GR, D'Orsi CJ, Hendrick RE, et al. Does litigation influence medical practice? The influence of community radiologists' medical malpractice perceptions and experience on screening mammography. Radiology 2005;236:37–46.
27. 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;5:195–201.
28. Duijm LE, Groenewoud JH, Hendriks JH, de Koning HJ. Independent double reading of screening mammograms in The Netherlands: effect of arbitration following reader disagreements. Radiology 2004;231:564–70.
29. Buist DS, Anderson ML, Haneuse SJ, Sickles EA, Smith RA, Carney PA, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology 2011;259:72–84.
30. Smith-Bindman R, Chu P, Miglioretti DL, Quale C, Rosenberg RD, Cutter G, et al. Physicianpredictors of mammographic accuracy. J Natl Cancer Inst 2005;97:358–67.
31. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst 2000;92:1081–7.