Eur Radiol
DOI 10.1007/s00330-015-4074-8

BREAST

Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain

Laia Domingo[1,2] · Solveig Hofvind[3,4] · Rebecca A. Hubbard[5] · Marta Román[3,6] · David Benkeser[7] · Maria Sala[1,2] · Xavier Castells[1,2]

Received: 12 June 2015 / Revised: 19 September 2015 / Accepted: 15 October 2015

Abstract
Objective To compare accuracy measures for mammographic screening in Norway, Spain, and the US.
Methods Information from women aged 50–69 years who underwent mammographic screening 1996–2009 in the US (898,418 women), Norway (527,464), and Spain (517,317) was included. Screen-detected cancer, interval cancer, and the false-positive rates, sensitivity, specificity, positive predictive value (PPV) for recalls (PPV-1), PPV for biopsies (PPV-2), 1/PPV-1 and 1/PPV-2 were computed for each country. Analyses were stratified by age, screening history, time since last screening, calendar year, and mammography modality.
Results The rate of screen-detected cancers was 4.5, 5.5, and 4.0 per 1000 screening exams in the US, Norway, and Spain respectively. The highest sensitivity and lowest specificity were reported in the US (83.1 % and 91.3 %, respectively), followed by Spain (79.0 % and 96.2 %) and Norway (75.5 % and 97.1 %). In Norway, Spain and the US, PPV-1 was 16.4 %, 9.8 %, and 4.9 %, and PPV-2 was 39.4 %, 38.9 %, and 25.9 %, respectively. The number of women needed to recall to detect one cancer was 20.3, 6.1, and 10.2 in the US, Norway, and Spain, respectively.
Conclusions Differences were found across countries, suggesting that opportunistic screening may translate into higher sensitivity at the cost of lower specificity and PPV.
Key Points
- *Positive predictive value is higher in population-based screening programmes in Spain and Norway.*
- *Opportunistic mammography screening in the US has lower positive predictive value.*
- *Screening settings in the US translate into higher sensitivity and lower specificity.*
- *The clinical burden may be higher for women screened opportunistically.*

Keywords Mammographic screening · Positive predictive value · Sensitivity · Specificity · Variability

Abbreviations

PPV-1 (positive predictive value-1)	The number of screen-detected breast cancers divided by the number of recalls due to positive mammographic findings.
PPV-2 (positive predictive value-2)	The number of screen-detected breast cancers divided by the number of recall examinations

✉ Xavier Castells
xcastells@hospitaldelmar.cat

1 Department of Epidemiology and Evaluation, IMIM (Hospital del Mar Medical Research Institute), Pg Marítim 25-29, 08003 Barcelona, Spain

2 Research Network on Health Services in Chronic Diseases (REDISSEC), Madrid, Spain

3 Department of Screening, Cancer Registry of Norway, P.O. 5313, Majorstua 0304, Oslo, Norway

4 Faculty of Health Science, Oslo and Akershus University College of Applied Sciences, P.O. 4, St. Olavs Plass, 0130 Oslo, Norway

5 Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, 604 Blockley Hall, 423 Guardian Dr, Philadelphia, PA 19104, USA

6 National Advisory Unit for Women's Health, Oslo University Hospital, Rikshospitalet, P.O. 4950, Nydalen 0424, Oslo, Norway

7 Department of Biostatistics, School of Public Health, University of Washington, F-600, Health Sciences Building, Seattle, WA 98195, USA

Published online: 11 November 2015

	including invasive procedures or (in the BCSC) recommendation for invasive procedures.
BCSC	Breast Cancer Surveillance Consortium
SFM	Screen-film mammography.
FFDM	Full-field digital mammography.
BI-RADS	Breast Imaging Reporting and Data System
CAD	Computer-aided detection.
CFPR	Cumulative false-positive risk
SEER	Surveillance, epidemiology, and end results
FNAC	Fine-needle aspiration cytology
DCIS	Ductal carcinoma in situ
CI	Confidence interval

Introduction

Over the last 20 years, mammographic screening has become widespread in most developed countries, with the aim of reducing breast cancer mortality through early detection of the disease. However, the organisation and delivery vary across geographic regions in ways that may influence its effectiveness. Most countries in Europe offer population-based screening programmes following the recommendations of the European guidelines with defined screening intervals and target populations related to age [1]. In the U.S, although several organisations recommend routine screening [2, 3], actual screening practices vary by personal and medical provider preferences. Access to care also varies, for instance, by insurance status. Most commonly, screening is opportunistic in response to recommendations made during a routine medical consultation or on the basis of a possible increased risk of developing breast cancer [2].

Comparing accuracy measures across different organisational models for mammographic screening provides valuable information that can be used to guide clinical practice and breast cancer screening policy. Only a few studies have compared such measures across organisational models for breast cancer screening and most have focused on the sensitivity, specificity, and cancer rates [4–8]. Positive predictive value (PPV), the proportion of women either recalled or who undergo a biopsy and are subsequently diagnosed with cancer, is reported less often.

This study utilises data collected in Norway and Spain as part of the service screening programmes and in the US by the Breast Cancer Surveillance Consortium (BCSC), a large ongoing study of mammographic screening performance in community practice. We estimated accuracy measures for mammographic screening in the three countries with the aim of comparing the sensitivity and specificity as well as PPV.

Materials and methods

Information from 898,418 screened women in the US (1996–2008), 527,464 in Norway (1996–2007), and 517,317 in Spain (1996–2009) was included in the study. All women were aged 50 to 69 years at screening. These women contributed a total of 5,713,594 screening exams.

Screening organisation

No organised mammography screening programme exists in the US (Table 1). Screening is performed opportunistically, typically according to the guidelines of organisations such as the American Cancer Society [3] and the US Preventive Services Task Force [2]. Under the Patient Protection and Affordable Care Act, all health insurance plans must cover screening mammography with no patient cost sharing [9]. All facilities performing screening mammography are accredited by the US FDA and follow the regulations set forth by the Mammography Quality Standards Act [10]. Recommendations generally call for initiation of screening at age 40 or 50 and continuation until at least age 74 [2]. The recommended screening interval also varies, with some organisations calling for annual and others for biennial screening. In consultation with their medical providers, women choose to receive screening mammography according to personal preference. Screening mammography typically consists of two-view (mediolateral oblique and craniocaudal views) bilateral examinations. Radiologists' assessments and recommendations are based on the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS®) [11] and are typically read by an individual radiologist, sometimes with the use of computer-aided detection (CAD). Full-field digital mammography (FFDM) began diffusing into community practice after FDA approval of the first FFDM machines in 2000. As of December 2009, 60 % of accredited mammography facilities in the US were using FFDM [12].

Both Norway and Spain adhere to the European Guidelines for Quality Assurance in Mammographic Screening [1] , which recommend biennial invitation to mammography screening for women aged 50 to 69 years.

The Norwegian Breast Cancer Screening Programme is administered by the Cancer Registry of Norway, which is also responsible for the surveillance and quality assurance of the programme. The Programme started in 1996 and became nationwide in 2005. The participation rate was 76.2 % [13]. Women are invited to two-view bilateral mammography by a personal invitation letter, regardless of cancer history. Women are screened at stationary and mobile units. The programme performs independent reading with consensus/arbitration. A score of 1–5 is given for each breast, by each radiologist, where 1 indicates a negative screening examination and 5 a high likelihood of malignancy. All cases with a score

Table 1 Description of the main characteristics of breast cancer screening organization in the US, Norway, and Spain

	US	Norway	Spain
Organisation	Opportunistic screening	Population-based screening programme	Population-based screening programme
Target population	40 years and older	50-69 years	50-69 years
Screening interval	1-2 years	2 years	2 years
Reading method	Mostly single reading Some use of CAD	Independent reading with consensus or arbitration	Double reading with consensus or arbitration
Population coverage	Vary by insurance status and personal preference	100 %	100 %

CAD Computer-aided detection

of 2 or higher given by one or both readers are discussed at a consensus meeting where the final decision of whether to recall the woman is made. Recall examinations take place at one of the 16 screening centres. FFDM was implemented gradually in Norway, from 2000v2011. As of the end of 2008, 48 % of the screening mammograms were performed with FFDM [14].

The Spanish Breast Cancer Screening Programme started in 1990 and was nationwide by 2006. The overall participation rate was 74.0 % [15]. In Spain, breast cancer screening is government funded. Women are actively invited to participate in population-based mammography screening by an invitation letter. The standard procedure for radiological performance in Spain is two-view mammography with double reading. The BI-RADS® [11] scale or equivalent is used to rate the probability of cancer. Women with positive mammographic findings, scored as 3, 4, 5, or 0, are recalled for further assessments to confirm or rule out malignancy at reference hospitals of each screening area. From 2004 onwards, FFDM was gradually introduced in Spain. As of December 2009, digital mammograms represented 25.7 % of screening tests.

Data sources

The study is based on data from the BCSC in the US, the Cancer Registry in Norway, and the updated database of the Cumulative False Positive Risk (CFPR) study in Spain [16].

The BCSC is a consortium of breast imaging registries throughout the US linked to population-based cancer registries. These registries collect information from community mammography facilities on mammography examinations and patient risk factors. The study included data on screening examinations in 1996–2008 captured by seven regional registries from diverse geographic locations that have previously been used to describe the distribution of screening mammography accuracy in the US [17]. Subsequent breast cancer diagnoses were obtained by linking BCSC data to pathology databases, regional Surveillance, Epidemiology, and End Results (SEER) programmes, and state tumour registries. Data were pooled at a central Statistical Coordinating Center. Data for this study were obtained from the BCSC Research Resource [18].

Screening data from Norway include examinations from women screened throughout the country between 1996 and 2007. Data from screening in Spain were drawn from an anonymised database that gathers information from eight screening areas. The database was originally created in 2006 for the CFPR Study [16] and was subsequently updated [19, 20]. The study includes data from women screened between 1996 and 2009.

All BCSC registries and the BCSC Statistical Coordinating Center received Institutional Review Board approval for active or passive consenting processes or a waiver of consent to enrol participants, link data, and perform analysis. All procedures were Health Insurance Portability and Accountability Act compliant, and registries and the Coordinating Center received a Federal Certificate of Confidentiality and other protections for the identities of women, physicians, and facilities. Data collection in Norway followed the regulations of the Cancer Registry of Norway and no ethical committee approval was necessary since all data received were aggregated. Data collection in Spain was performed following a study protocol approved by the institutional review boards at all participating screening areas.

Definitions

For US women, a screening mammogram was defined as bilateral mammograms with screening indication performed on women without a personal history of breast cancer or breast augmentation who had not received mammography within the prior 9 months. In Norway and Spain, all mammograms performed on women attending the population-based screening programme were considered screening mammograms.

A recall was defined as abnormal findings on the screening mammogram, leading to a recall for further assessment. Based on the findings of the imaging workup, women were referred back to screening or for an invasive procedure [fine-needle

aspiration cytology (FNAC), core needle biopsy, or an open biopsy]. Short-term follow-up at 6 months after the screening examination is sometimes recommended in the US but is not recommended in Spain and in Norway where further assessment takes place and concludes within 4 months of the screening examination. For the BCSC cohort, a recall was defined as a BI-RADS assessment of 0, 4, or 5 [11].

For all three countries a false-positive recall was defined as a recall for further assessment where no breast cancer was confirmed, regardless of the procedures performed. A false-positive screening result may also include an invasive procedure with benign morphology, referred to as a false positive with invasive procedures. A screen-detected cancer was defined as ductal carcinoma in situ (DCIS) or invasive breast cancer diagnosed as a result of further assessment due to abnormal findings on the screening mammograms.

In the BCSC data, a positive screening result was defined as false positive if no cancer was diagnosed within 12 months of the screening examination and prior to the next screening mammogram. All cancers diagnosed within 12 months of a positive screening mammogram and prior to the next screening mammogram were considered screen-detected. An interval cancer was defined as a breast cancer detected within 12 months after a negative screening mammogram and prior to the next screening mammogram.

In Norway and Spain, false-positives and screen-detected cancers were defined based on cancers diagnosed as a result of further assessment conducted following the screening mammogram. An interval cancer was defined as a breast cancer diagnosed within 730 days after a negative screening examination, with or without an invasive procedure, and before the next screening examination.

Sensitivity was defined as the number of screen-detected cancers divided by the number of screen-detected cancers plus interval cancers, while specificity was defined as the number of true-negative screening examinations divided by the number of true-negatives tests plus false positives.

Rates were defined as the number of cases per 1000 screening examinations. PPV-1 was defined as the number of screen-detected breast cancers divided by the number of recalls due to positive mammographic findings. PPV-2 refers to recalled examinations including invasive procedures or (in the BCSC) recommendation for invasive procedures. The number of women needed to be recalled and to undergo an invasive procedure to detect one breast cancer was estimated by taking the inverse of PPV-1 (1/PPV-1) and PPV-2 (1/PPV-2), respectively

Statistical analysis

We included all screening mammograms performed on eligible women during the study period, including multiple screening mammograms for some women. We used generalised estimating equations (GEE) to account for within-woman correlation by means of the robust Huber-White (sandwich) variance estimator [21]. The z-test was used to examine differences in accuracy measures between countries. *P*-values< 0.05 were considered statistically significant.

Estimates of sensitivity and specificity were stratified by several factors: first or subsequent screen, calendar year of the screening mammogram, age at screening, and screening modality [screen-film mammography (SFM) or FFDM]. Cancer detection rates, false-positive rates, PPV-1, and PPV-2 were stratified by the time since the last screening mammogram (<18 months, 18 to 30 months, >30 months). The 95 % confidence intervals (95 % CIs) were calculated.

Analyses were conducted using R v.3.0.0 (US), STATA v.12 (Spain), SPSS v.12.0 (Spain and Norway), and SAS (Norway).

Results

The study included information about 5,713,594 screening examinations from 1,943,199 women screened in 1996–2009 at age 50 to 69 years. Overall, 26,430 cancers were screen-detected and 6,756 emerged as interval cancers.

Tables 2 and 3 show overall measures of screening accuracy in the three countries. The highest rate of screen-detected cancers was found in Norway, followed by the US and Spain [5.5 cancers per 1,000 screening mammograms, 4.5‰ and 4.0‰, respectively ($p<0.001$)], which is equivalent to 181.5, 223.0, and 247.4 screening examinations needed to detected one cancer, respectively. The highest rate of DCIS was observed in the US, followed by Norway and Spain [1.1 ‰, 0.9‰, and 0.7‰, respectively ($p<0.001$)]. The highest sensitivity was reported in the US, followed by Spain and Norway (83.1 %, 79.0 %, and 75.5 %). Conversely, the highest specificity was found in Norway, followed by Spain and the US (97.1 %, 96.2 %, and 91.3 %). PPV-1 was 16.4 % in Norway, 9.8 % in Spain, and 4.9 % in the US, which implies that 6.1 women were required to undergo further workup to detect one cancer in Norway, 10.2 in Spain, and 20.3 in the US. PPV-2 was 39.4 % in Norway, 38.9 % in Spain, and 25.9 % in the US.

Stratification revealed differences between the countries for both sensitivity and specificity (Table 4) that were consistent with the overall measures observed in Table 3. In all strata, sensitivity was higher in the US than in Norway and Spain, and specificity was lower in the US than in Norway and Spain. However, there were some notable patterns in the differences within specific strata. Specifically, differences in specificity between countries were larger at the first compared to subsequent screenings. The smallest differences for sensitivity, but not for specificity, were detected in women aged 60–69 (83.8 %, 82.1 %, and 79.2 % in the US, Spain, and

Table 2 Number and rate (per 1000 screening examinations) of screen-detected, interval cancer and false-positive screening examinations (per 100 screening examinations) in mammographic screening performed in the US, Norway, and Spain

	US (1996–2008) *n*=2,656,834 *n* (rate)	Norway (1996–2007) 1,470,854 *n* (rate)	Spain (1996–2009) 1,585,906 *n* (rate)
Screen-detected cancers (*n*, rate per 1000 screening examinations)			
All malignant lesions	11,916 (4.5‰)	8105 (5.5‰)	6409 (4.0‰)
Invasive	9028 (3.4‰)	6714 (4.6‰)	5147 (3.2‰)
DCIS	2888 (1.1‰)	1391 (0.9‰)	1077 (0.7‰)
Unknown	0	0	185 (0.1‰)
Interval cancers (*n*, rate per 1000 screening examinations)			
All malignancies	2429 (0.9‰)	2623 (1.8‰)	1704 (1.1‰)
Invasive	2191 (0.8‰)	2485 (1.7‰)	(NA)
DCIS	238 (0.1‰)	138 (0.1‰)	(NA)
False-positive screening examinations (*n*, rate per 100 screening examinations)			
Additional imaging	230,016 (8.7 %)	42,426 (2.9 %)	59,414 (3.7 %)
Invasive procedures	24,627 (0.9 %)[a]	12,476 (0.8 %)	10,229 (0.6 %)

[a] US data do not capture all biopsies: the number of women who were recommended to receive a biopsy, defined as BI-RADS assessment at the end of all imaging workup of 4 or 5, or BI-RADS assessment of 0 or 3 accompanied by recommendation for biopsy, FNAC, or surgical consult is thus reported

Norway, respectively). Differences in sensitivity among the US, Norway, and Spain were greater with FFDM (85.8 %, 73.5 %, and 76.4 %, respectively). The only exception to this pattern was found in the first years of the study period, where the sensitivity in Spain was higher than in the US.

In both in Norway and Spain, the largest percentage of subsequent examinations was performed 18–30 months after the prior examination (95.8 % and 93.9 %, respectively) whereas in the US 68.5 % of screening tests were performed within 18 months of the prior test (Table 5). For all countries, the longer the time since the prior screening test, the higher the rate of screen-detected cancers, invasive cancers, false-positives, and PPV-1 was. PPV-1 for mammograms performed 18–30 months after the last screening was 5.2 %, 19.4 %, and 11.6 % in the US, Norway, and Spain, respectively.

Discussion

We compared accuracy measures for mammographic screening performed in community practice in the US and through population-based screening programmes in two European countries. The highest specificity and PPV were found in the European population-based screening programmes, whereas the highest sensitivity was found in the US. The results suggest that the opportunistic approach with annual mammography requires more interventions to detect one cancer compared with biennial screening in organised programmes.

Opportunistic screening is known to be more interventionist than population-based approaches, which translates into a higher number of recalls and false-positive results, as reported in prior studies comparing screening performance indicators between the US and Europe [4, 7]. Different explanations for the higher recall rates in the US have been proposed. First, the

Table 3 Sensitivity, specificity, positive predictive value of recalls (PPV-1) and invasive procedures (PPV-2) in mammographic screening performed in the US, Norway, and Spain

	US (1996–2008) *n* =2,656,834 % (95 % CI)	Norway (1996–2007) 1,470,854 % (95 % CI)	Spain (1996–2009) 1,585,906 % (95 % CI)
Sensitivity	83.1 (82.4−83.7)	75.5 (74.7−76.4)	79.0 (78.1−79.9)
Specificity	91.3 (91.2−91.3)	97.1 (97.1−97.1)	96.2 (96.2−96.2)
PPV-1	4.9 (4.8−5.0)	16.4 (16.0−16.7)	9.8 (9.8−9.8)
1/PPV-1	20.3 (20.0−20.7)	6.1 (6.0−6.2)	10.2 (10.2−10.3)
PPV-2	25.9[a] (25.4−26.4)	39.4 (38.7−40.1)	38.9 (38.9−38.9)
1/PPV-2	3.9 (3.8−3.9)	2.5 (2.5−2.6)	2.6 (2.6−2.6)

[a] US data do not capture all biopsies: the number of women who were recommended to receive a biopsy, defined as BI-RADS assessment at the end of all imaging workup of 4 or 5, or BI-RADS assessment of 0 or 3 accompanied by recommendation for biopsy, FNAC, or surgical consult is thus reported

Table 4 Sensitivity and specificity by type of screening history (first or subsequent), year, age at screening examination, and type of mammography (screen-film mammography, SFM, or full-field digital mammography, FFDM)

	US			Norway			Spain		
	Screen examinations (n)	Sensitivity (%, 95 % CI)	Specificity (%, 95 % CI)	Screen examinations (*n*)	Sensitivity (%, 95 % CI)	Specificity (%, 95 % CI)	Screen examinations (*n*)	Sensitivity (%, 95 % CI)	Specificity (%, 95 % CI)
Screening history									
First	48,822	91.9 (89.2- 94.1)	83.3 (83.0-83.6)	500,476	77.4 (76.1-78.7)	95.7 (95.6-95.8)	517,317	83.2 (81.9-84.5)	94.1 (94.0-94.3)
Subsequent	2,555,186	82.4 (81.8-83.1)	91.6 (91.6-91.7)	970,378	74.4 (73.3-75.5)	97.8 (97.8-97.8)	1,068,589	76.4 (75.3-77.6)	97.2 (97.2-97.2)
Year of screening examination									
1996-97	313,627	78.7 (76.7- 80.5)	91.8 (91.7- 91.9)	111,531	76.9 (74.3-79.6)	96.3 (96.2-96.4)	37,826	84.3 (79.3-89.4)	95.1 (95.1-95.1)
1998-99	403,672	78.7 (77.0-80.4)	91.2 (91.2- 91.3)	129,090	73.2 (70.3-76.1)	96.9 (96.8-97.0)	68,294	83.8 (79.9-87.8)	96.3 (96.3-96.3)
2000-01	421,336	83.2 (81.7-84.7)	91.3 (91.2-91.4)	196,307	74.4 (72.2-76.6)	97.2 (97.1-97.2)	143,140	82.5 (79.7-85.4)	96.2 (96.2-96.3)
2002-03	492,878	85.0 (83.5-86.3)	91.0 (90.9-91.1)	307,587	76.7 (74.9-78.4)	96.9 (96.8-97.0)	208,267	78.5 (76.0-80.9)	96.5 (95.5-96.5)
2004-05	453,885	85.3 (83.7-86.7)	91.2 (91.1-91.3)	357,485	76.4 (74.7-78.1)	97.3 (97.2-97.3)	268,244	76.8 (74.5-79.1)	96.9 (96.8-96.9)
2006-07	388,167	84.8 (83.3-86.3)	91.3 (91.2-91.4)	368,854	74.7 (73.1-76.4)	97.4 (97.3-97.4)	323,949	77.7 (75.7-79.7)	96.4 (96.4-96.4)
2008-09	183,269	86.6 (84.4-88.5)	91.4 (91.2-91.5)	NA	NA	NA	536,186	79.2 (77.7-80.6)	95.8 (95.8-95.8)
Age at screening examination									
50-59	1,603,333	82.5 (81.6-83.3)	90.8 (90.7-90.8)	863,622	72.4 (71.2-73.5)	96.7 (96.6-96.7)	936,038	74.9 (73.7-76.2)	95.6 (95.6-95.6)
60-69	1,053,501	83.8 (82.8-84.6)	92.1 (92.0-92.1)	607,232	79.2 (78.1-80.4)	97.7 (97.7-97.7)	649,868	82.1 (80.9-83.3)	97.1 (97.1-97.1)
Type of mammography									
SFM	2,320,543	82.7 (82.0-83.3)	91.3 (91.3-91.4)	1,283,482	75.9 (75.0-76.8)	97.1 (97.0-97.1)	1,211,029	78.1 (77.0-79.1)	96.7 (96.6-96.7)
FFDM	336,291	85.8 (84.0-87.3)	90.9 (90.8-91.0)	187,372	73.5 (71.3-75.6)	97.2 (97.2-97.3)	374,877	76.4 (75.2-77.7)	94.6 (94.6-94.6)

Table 5 Number and rates (per 1000 screening examinations) of screen-detected cancer, false-positive screening examinations (*n*, per 100 screening examinations) and positive predictive value of recalls (PPV-1) and invasive procedures (PPV-2) by time since last screening examination

	US			Norway		Spain	
	<18 months (*n*=1,635,247)	18-30 months (*n*=496,738)	>30 months (*n*=255,126)	18-30 months (*n*=903,709)	>30 months (*n*=39,681)	18-30 months (*n*=1,003,210)	>30 months (*n*=65,379)
Screen-detected cancers (n, rate per 1000 screening examinations)							
All malignancies	5927 (3.6‰)	2363 (4.8‰)	1692 (6.6‰)	4486 (5.0‰)	311 (7.8‰)	3501 (3.5‰)	350 (5.4‰)
Invasive	4383 (2.7‰)	1814 (3.7‰)	1332 (5.2‰)	3723 (4.1‰)	265 (6.7‰)	2798 (2.8‰)	274 (4.2‰)
DCIS	1544 (0.9‰)	549 (1.1‰)	360 (1.4‰)	763 (0.8‰)	46 (1.2‰)	550 (0.5‰)	59 (0.9‰)
Unknown	0	0	0	0	0	153 (0.2‰)	17 (0.3‰)
False positives (*n*, rate per 100 screening examinations)							
Additional images	124,896 (7.6 %)	42,644 (8.6 %)	26,784 (10.5 %)	19,068 (2.1 %)	1,181 (3.0 %)	26,731 (2.7 %)	2345 (3.6 %)
Invasive procedures	12,702 (0.8 %)	4220 (0.8 %)	3268 (1.3 %)	4805 (0.5 %)	387 (1.0 %)	3615 (0.4 %)	316 (0.5 %)
Positive predictive value (%, 95 % CI)							
PPV-1	4.5 (4.4-4-6)	5.2 (5.0 - 5.4)	5.9 (5.6 - 6.2)	19.4 (18.9-19.9)	21.4 (19.4-23.6)	11.6 (11.6-11.6)	13.0 (13.0-13.0)
PPV-2	25.2 (24.5-25.8)	29.5 (28.4-30.7)	27.7 (26.5-29.0)	48.3 (47.3-49.3)	44.6 (47.3-49.3)	49.7 (49.7-49.7)	52.6 (52.5-52.5)

use of a single reading and the lower radiologist interpretive volume required in the US for accreditation may partially explain the differences [12]. A minimum interpretive volume of 5000 mammograms per radiologist per year is recommended in Europe [1]. These programmatic characteristics, however, have not been associated with a decrease in the sensitivity or cancer detection rate [22]. A second explanation is that the threat of lawsuits for malpractice in the US might induce radiologists to order further tests and procedures aiming to decrease the number of missed cancers [23, 24]. Finally, differences exist with respect to the targets for recall rates: European guidelines recommend <3 % of mammograms should result in recalls [1] while the BI-RADS recommendations in the US are 5-10 % [11]. In spite of organisational similarities between Norway and Spain, we observed differences in PPV-1 as a result of the higher detection rates and lower false-positive rates in Norway than in Spain. The smaller cancer detection rate observed in Spain can be partly attributed to the lower background breast cancer incidence in this country in comparison with Norway and the US [25]. Different definitions of recall for invasive procedures among the US, Norway, and Spain may also partly explain the lower PPV-2 values reported in the US.

The higher sensitivity in the US in comparison to Norway and Spain can be partially explained by the different screening periodicity. It could also be affected by other factors such as the test sensitivity. However, identifying the specific contributing factors is beyond the scope of this study. Women in the US were less likely to develop an interval cancer between screening examinations, which were mostly annual, which was directly reflected in sensitivity. Unfortunately, because of differences in screening practices between the US and the European screening programmes, we were not able to compare the sensitivity for women screened every 2 years in the three countries. Nevertheless, prior work comparing the sensitivity between the US (based on data from Vermont and North Carolina) and Norway indicated that the sensitivity for 2-year screening intervals in these US regions was almost the same as that in Norway [4, 5, 8].

The trend of higher sensitivity and lower specificity in the US vs. Spain and Norway persisted across all strata investigated. Differences observed in first screenings, both for the percentage of mammograms and for sensitivity and specificity, could be related to differences in recommendations for screening initiation. In the US, some organisations and providers recommend that women begin screening at age 40 [3]. As a result, relatively few women have their initial screening at age 50 or older. The comparison of characteristics of first screening examinations is thus confounded by age at first screening.

When comparing cancer rates among women who attended screening with an 18-30-month interval, the values from the US and Norway—countries with similar background breast cancer incidence [25] —were similar. However, PPVs continued to differ. This reflects the fact that, while cancer detection is mainly dependent on background incidence, PPV is more sensitive to variations in radiological practice and the organisation of the screening programme.

Our study has some limitations. First, we have taken a descriptive approach based on aggregate data, and therefore we did not control for potential confounders like individual age. However, in an attempt to make data more comparable across countries, we restricted the study population to women aged 50–69. Second, despite using consensual definitions of

the screening terms, some unavoidable differences remained such as the definition for interval cancer in the US and Europe, which directly affects the reported screening sensitivity. In the US, estimates of sensitivity and/or interval cancers based on 2-year follow-up after the screening mammogram would be biased because the majority of women return for another screening examination at a 1-year interval. However, the results represent variability in the radiological performance between countries, which is the main objective of the study. Third, some overestimation of sensitivity estimates in Spain cannot be discounted since there is a lack of a nationwide cancer registry, which may result in some missed interval cancers. However, the mechanisms for identifying interval cancers have improved over time [26], which is reflected in a decrease in sensitivity estimates.

In summary, the opportunistic approach to screening in the US is more interventionist, resulting in more frequent follow-up evaluations and shorter screening intervals than the European population-based approaches. This translates into a somewhat higher sensitivity of screening mammography in the US but at the cost of higher clinical burden on the women. Population-based approaches stress the balance between sensitivity and specificity, aiming to decrease the clinical burden—and the related harms and costs—to participating women.

Acknowledgments The scientific guarantor of this publication is Xavier Castells. The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article. This work was partially supported by grants from Instituto de Salud Carlos III FEDER (PI11/01296), BELE Study, the National Institutes of Health (T32CA09168), and the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). The collection of US cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We thank the participating women, mammography facilities, and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/. One of the authors (Marta Román) has significant statistical expertise. All BCSC registries and the BCSC Statistical Coordinating Center received Institutional Review Board approval for active or passive consenting processes or a waiver of consent to enrol participants, link data, and perform analysis. All procedures were Health Insurance Portability and Accountability Act compliant, and registries and the Coordinating Center received a Federal Certificate of Confidentiality and other protections for the identities of women, physicians, and facilities. Data collection in Norway followed the regulations of the Cancer Registry of Norway and no ethical committee approval was necessary since all data received were aggregated. Data collection in Spain was performed following a study protocol approved by the institutional review boards at all participating screening areas.

Methodology: retrospective, observational, multicentre study.

References

1. Perry N (2006) In: Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L (eds) European guidelines for quality assurance in breast cancer screening and diagnosis, Fourthth edn. Office for Official Publications of the European Communities, Luxembourg
2. Screening for breast cancer: US Preventive Services Task Force recommendation statement (2009). Ann Intern Med 151:716–236
3. Smith RA, Saslow D, Sawyer KA et al (2003) American Cancer Society guidelines for breast cancer screening: update 2003. CA Cancer J Clin 53:141–69
4. Hofvind S, Vacek PM, Skelly J, Weaver DL, Geller BM (2008) Comparing screening mammography for early breast cancer detection in Vermont and Norway. J Natl Cancer Inst 100:1082–1091
5. Hofvind S, Geller BM, Skelly J, Vacek PM (2012) Sensitivity and specificity of mammographic screening as practised in Vermont and Norway. Br J Radiol 85:e1226–e1232
6. Lynge E, Ponti A, James T et al (2014) Variation in detection of ductal carcinoma in situ during screening mammography: a survey within the International Cancer Screening Network. Eur J Cancer 50:185–192
7. Smith-Bindman R, Ballard-Barbash R, Miglioretti DL, Patnick J, Kerlikowske K (2005) Comparing the performance of mammography screening in the US and the UK. J Med Screen 12:50–54
8. Hofvind S, Yankaskas BC, Bulliard JL, Klabunde CN, Fracheboud J (2009) Comparing interval breast cancer rates in Norway and North Carolina: results and challenges. J Med Screen 16:131–139
9. Public Law 111–148, Patient Protection and Affordable Care Act. 23:3 (2010)
10. Department of Health and Human Services (1997) Quality mammography standards. Final Rules. Washington
11. American College of Radiology (ACR) (2003) Breast imaging reporting and data system Atlas (BI-RADS®Atlas). Reston
12. US Drug and Food Administration website. The Mammography Quality Standards Act of 1998 (as amended by MQSRA of 1998 and 2004). http://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/DocumentArchives/ucm128078.htm (Last Accessed: 12 Apr 2015)
13. Hofvind S, Geller B, Vacek PM, Thoresen S, Skaane P (2007) Using the European guidelines to evaluate the Norwegian Breast Cancer Screening Program. Eur J Epidemiol 22:447–55
14. Hofvind S, Skaane P, Elmore JG, Sebuødegård S, Hoff SR, Lee CI (2014) Mammographic performance in a population-based screening program: before, during, and after the transition from screen-film to full-field digital mammography. Radiology 272:52–62
15. Spanish Cancer Screening Programmes network website. http://www.programascancerdemama.org/. (Last Accessed 17 Sept 2015)
16. Román R, Sala M, Salas D, Ascunce N, Zubizarreta R, Castells X (2012) Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening. Ann Oncol 23:104–11
17. Rosenberg RD, Yankaskas BC, Abraham LA et al (2006) Performance benchmarks for screening mammography. Radiology 241:55–66

18. The Breast Cancer Surveillance Consortium website. http://breastscreening.cancer.gov/. (Last Accessed 12 Apr 2014)
19. Castells X, Domingo L, Corominas JM et al (2015) Breast cancer risk after diagnosis by screening mammography of nonproliferative or proliferative benign breast disease: a study from a population-based screening program. Breast Cancer Res Treat 149:237–44
20. Domingo L, Salas D, Zubizarreta R et al (2014) Tumor phenotype and breast density in distinct categories of interval cancer: results of population-based mammography screening in Spain. Breast Cancer Res 16:R3
21. Diggle P, Heagerty P, Liang K (2002) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford
22. Theberge I, Chang SL, Vandal N et al (2014) Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. J Natl Cancer Inst 106:djt461
23. Whang JS, Baker SR, Patel R, Luk L, Castro A (2013) The causes of medical malpractice suits against radiologists in the United States. Radiology 266:548–54
24. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF (2003) International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst 95:1384–93
25. GLOBOCAN (2012) Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. International Agency for Research on Cancer (IARC) http://globocan.iarc.fr/
26. Sala M, Domingo L, Macià F, Comas M, Burón A, Castells X (2015) Does digital mammography suppose an advance in early diagnosis? Trends in performance indicators 6 years after digitalization. Eur Radiol 25:850–9