Breast cancer risk is increased in the years following false-positive breast cancer screening

Mathijs C. Goossens[a,c], Isabel De Brabander[b], Jacques De Greve[a], Evelien Vaes[b], Chantal Van Ongeval[d], Koen Van Herck[c,e] and Eliane Kellen[c,d]

A small number of studies have investigated breast cancer (BC) risk among women with a history of false-positive recall (FPR) in BC screening, but none of them has used time-to-event analysis while at the same time quantifying the effect of false-negative diagnostic assessment (FNDA). FNDA occurs when screening detects BC, but this BC is missed on diagnostic assessment (DA). As a result of FNDA, screenings that detected cancer are incorrectly classified as FPR. Our study linked data recorded in the Flemish BC screening program (women aged 50–69 years) to data from the national cancer registry. We used Cox proportional hazards models on a retrospective cohort of 298 738 women to assess the association between FPR and subsequent BC, while adjusting for potential confounders. The mean follow-up was 6.9 years. Compared with women without recall, women with a history of FPR were at an increased risk of developing BC [hazard ratio = 2.10 (95% confidence interval: 1.92–2.31)]. However, 22% of BC after FPR was due to FNDA. The hazard ratio dropped to 1.69 (95% confidence interval: 1.52–1.87) when FNDA was excluded. Women with FPR have a subsequently increased BC risk compared with women without recall. The risk is higher for women who have a FPR BI-RADS 4 or 5 compared with FPR BI-RADS 3. There is room for improvement of diagnostic assessment: 41% of the excess risk is explained by FNDA after baseline screening. *European Journal of Cancer Prevention* 26:396–403 Copyright © 2017 The Author(s). Published by Wolters Kluwer Health, Inc.

European Journal of Cancer Prevention 2017, **26**:396–403

Keywords: breast neoplasms, false-positive recall, mammographic screening, risk

[a]Vrije Universiteit Brussel, [b]Belgian Cancer Registry, [c]Centrum voor Kankeropsporing (Center for Cancer Detection), [d]University Hospital Leuven and [e]Ghent University

Correspondence to Mathijs C. Goossens, MSc, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium
Tel: +00 32 2477 5402; fax: +00 32 2477 6210;
e-mail: mathieu.goossens@uzbrussel.be

Received 20 May 2016 **Revised** 29 July 2016

Introduction

Widespread implementation of mammographic screening for breast cancer (BC) has the potential of decreasing BC mortality but also involves a risk of adverse effects such as overdiagnosis and false-positive recall (FPR) (Marmot *et al.*, 2013).

FPR means a woman was recalled due to suspicious findings on the screening mammogram, but no evidence of BC was found at diagnostic assessment (DA). However, DA is not infallible: when cancers that were seen at screening are missed at DA, this is called false-negative diagnostic assessment (FNDA). When FNDA occurs, screenings that detected BC are incorrectly classified as FPR. In other words, recall can be either true-positive (screen-detected cancer) or false-positive (no screen-detected cancer), but occasionally a FPR turns out to be a true-positive recall that was misclassified as false-positive because DA failed to find the cancer that was seen at screening (Duijm *et al.*, 2004; von Euler-Chelpin *et al.*, 2014).

In Europe, women aged 50–69 years have a cumulative risk of between 8 and 32% of having at least one FPR over the course of 10 screening rounds (Castells *et al.*, 2006; Hofvind *et al.*, 2012). This variation can be explained by differences in screening organization, protocol characteristics, recall rates, and a woman's own risk profile (Christiansen *et al.*, 2000; Castells *et al.*, 2006). The Recall rate is the number of women recalled for assessment as a proportion of all women who had a screening examination. According to the European guidelines, recall rates should be below 7% in first round screenings (preferably below 5%) and below 5% in subsequent round screenings (preferably below 3%). FPR has several disadvantages, which include patients experiencing anxiety while waiting for the result, decreasing reattendance rates in the next screening round, increasing financial burden on the healthcare system, and increasing workload for healthcare staff (Bangsboll-Andersen *et al.*, 2008; Alamo-Junquera *et al.*, 2011; Maxwell *et al.*, 2013; Goossens *et al.*, 2014). Women who have had an FPR are also at an increased risk for BC compared with women who were not recalled. The

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (*www.eurjcancerprev.com*).

0959-8278 Copyright © 2017 The Author(s). Published by Wolters Kluwer Health, Inc. DOI: 10.1097/CEJ.0000000000000313

relative risk for BC after FPR has been estimated to be between 1.67 [95% confidence interval (CI): 1.45–1.88] and 2.69 (95% CI: 2.28–3.16) and may be highest for women who underwent fine-needle aspiration cytology (FNAC) or core-needle biopsy (CNB) during assessment (von Euler-Chelpin *et al.*, 2012; Castells *et al.*, 2013; Henderson *et al.*, 2015).

However, not all of these studies correct for FNDA and it is unclear as to what proportion of the risk is attributable to FNDA. Deciding whether a recall led to a screen-detected cancer (SDC) or whether it was an FPR requires knowledge of the conclusion of DA; this is why the Flemish BC screening program systematically contacts the woman's physician after each recall to request the DA conclusion. For screenings in 2005 and 2006, this provided a clear conclusion (SDC or no cancer) for about 77% of recalls; the remaining 23% either consists of nonresponse or of indecisive conclusions (unknown results, refusal to perform DA, etc.). This method alone is not sufficient for a study such as this one as it does not provide information on the BC status of women who were not recalled, or for whom the DA conclusion is not available, and it does not follow-up women in the years after screening. Another way to decide whether a recall led to an SDC is to use cancer registry data. The advantage of this option is that these data are routinely available and the degree of completeness and accuracy can be estimated. They are also available for women without recall, and allow women to be followed through the years after screening. However, although cancer registry data typically provide a date of diagnosis, they do not usually include information from DA, such as whether it is an SDC. A frequently used solution is to decide on a cutoff period for the time between recall and the date of diagnosis within which a BC is assumed to have been found at the DA following screening and is therefore an SDC. Some studies that investigated the risk for BC after FPR have put this cutoff at 12 months after recall (Peeters *et al.*, 1988; Henderson *et al.*, 2015), whereas some studies do not explain in detail how they determine SDC status (von Euler-Chelpin *et al.*, 2012; Castells *et al.*, 2013). Although this is a pragmatic approach, using a 12-month cutoff means that some BC classified as SDC were in reality not found at the DA after screening.

The type of statistical analysis of the studies investigating the risk for BC after FPR is also up for debate: the outcome variable in BC studies can either be seen as dichotomous (cancer/no cancer), or as the time elapsed until BC. In the first case, logistic or Poisson regression is a good choice, whereas in the second case a time-to-event analysis would be warranted. Even though both analysis options are defendable, studies with a long follow-up period will have patients who are censored before the end of the study. With logistic regression these patients would have to be excluded from the study or be assumed not to have had the event, but time-to-event analysis is able to deal with censored data. In our view, this makes time-to-event analysis the preferred method of analysis (Hosmer and Lemeshow, 2000). To our knowledge, only one study has investigated long-term BC risk after FPR using time-to-event analysis (such as Cox proportional hazards models), but this study did not investigate the role of FNDA (Henderson *et al.*, 2015).

Further research on FPR as a risk factor for BC is important because an adapted follow-up regimen might be indicated if the risk for BC is considerably increased for these women. In the current study we apply time-to-event analysis on a retrospective cohort of screened women to estimate their BC risk, while taking FNDA into account.

Patients and methods

Setting

The Belgian region of Flanders has about 1.4 million female inhabitants between 50 and 69 years of age and has a BC screening program compliant with relevant European guidelines (Perry *et al.*, 2008). Within the BC screening program, every 2 years all eligible women in the age group 50–69 years receive an invitation letter with a set appointment for a BC screening mammogram. Women at an increased risk for breast cancer are not excluded from the mailing list, but are advised to discuss with their physician whether the organized MBCS program is suitable for them, or whether they would benefit more from other types of prevention. Flanders also has opportunistic screening, which is by prescription only. Opportunistic screening, unlike the BC screening program, does not include organized quality control (e.g. double reading), its data are not stored in one central database, and it is not free of charge. Women with BC found through opportunistic screening have been excluded from the BC screening program mailing list since 2016. The percentage of women screened in Flanders during the period 2006–2007 was 21% in opportunistic screening and 44% through the BC screening program, giving a total coverage of 65% (Intermutualistisch Agentschap, 2010).

Screening in the BC screening program always consists of a two-view mammogram (medio-lateral-oblique and cranio-caudal) of each breast, without ultrasound or clinical breast examination. All examinations are read independently by two certified screening radiologists, both of whom use a scoring system to describe whether they recommend recall. This system resembles the Breast Imaging Reporting and Data System (BI-RADS): 0 = screening mammogram is of insufficient quality to make a conclusion, 1 = no abnormality, 2 = benign lesion, 3 = probably benign lesion, 4 = probably malignant lesion, and 5 = highly suspicious for malignancy. If the two readers have discrepant assessments, a third reader is consulted. Score 0 will lead to the women being recalled for a new mammogram without any conclusion on the

presence of cancer. Recall for assessment will always be recommended for scores 3, 4 and 5, whereas no recall will be requested for scores 1 and 2.

Irrespective of whether recall score 3, 4 or 5 is used, readers can advise DA either to take place as soon as possible or to be performed 6 or 12 months after screening without any immediate assessment (short-interval follow-up). For recall score 3, immediate additional imaging was recommended for 97.3% of recalls; the remaining 2.7% were recommended short-interval follow-up. For recall scores 4 or 5 the corresponding percentages were 99.4 and 0.6%. A recommendation for the DA type is sent to the woman's physician along with the screening result and can include noninvasive procedures (MRI, ultrasound, and additional mammography) and/or invasive (fine-needle aspiration cytology, core-needle biopsy and open biopsy). About 96% of women receive their results within 3 weeks of screening and more than 90% of DA is performed within 1 month after recall (Martens *et al.*, 2015).

Besides screening interpretation, readers also estimate breast density, classified according to the percentage of fibroglandular tissue (A $\leq$ 25%, B = 26 – 50%, C = 51–75%, D > 75%).

Ethics statement

When registering for mammographic screening, all women are asked to provide written consent for their data to be used in research related to the quality of the screening program. The Sectoral Committee of Social Security and Health (the national privacy commission) approved the use of a unique patient identifier to crosslink screening data to oncological data from the national population-based cancer registry (Belgian Cancer Registry, BCR) for women who signed such informed consent.

Study population

Supplementary Fig. S1 (Supplemental digital content 1, *http://links.lww.com/EJCP/A122*) shows the study flow diagram. We built a retrospective cohort of all women who participated in the Flemish BC screening program between January 2005 and December 2006. Exclusion criteria were as follows: lack of a signed informed consent form and BC diagnosis preceding the baseline screening. The remaining cohort ($n = 292\,731$) was then split into three groups: no recall; recall score 3; and recall score 4 or 5. Women with SDCs or women who were recalled but did not go for diagnostic assessment within 12 months of recall were excluded from groups of recalled women; this led to three groups in the study [Supplementary Fig. S1 (Supplemental digital content 1, *http://links.lww.com/EJCP/A122*)]:

(1) group without recall,
(2) FPR 3 group (FPR after a recall for probably benign lesion),
(3) FPR 4/5 group (FPR after a recall for probably malignant or highly suspicious lesion).

Study follow-up

The follow-up period started at baseline screening (T_0) for all participants in the study. At the time of this study, BCR data on cancer incidence were complete until the end of the year 2012; the follow-up of this study therefore ends on 31 December 2012. Loss to follow-up was defined as emigration out of Belgium before the end of follow-up. Individual data on emigration and vital status were obtained through linkage with the Crossroads Bank for Social Security (CBSS).

Primary outcome

The primary outcome of this study was time to incident BC. BC was defined as invasive carcinoma or ductal in-situ carcinoma of the breast (C50 and D05, respectively, of ICD-O, third edition, version 10).

Definition of false-positive recall and screen-detected cancer

Women with FPR were found by removing all SDCs (the true positives) from the group of recalled women. This means that defining SDCs is in fact the first step to defining FPR. In our study, BCs were classified as SDCs if they were diagnosed within 3 months of the date of DA (which is available through the BCR). This method uses cancer registry data and includes a risk of underestimating SDC; in the discussion we will therefore compare the results of this method with the results based on the conclusions from DA obtained from the women's physicians.

Definition of false-negative diagnostic assessment

Location data are routinely registered in the screening database and the BCR, and include up to 10 different segments in each breast. Table 1 shows how location data together with the timing of diagnosis were used to classify BCs after FPR as either FNDA or new cancers (BC unrelated to the T_0 recall). The cancers in the group without recall at T_0 were not classified.

Statistical analysis

When appropriate, sample characteristics were compared between the three groups using the χ^2-test, Fisher's exact test, or one-way analysis of variance. Incidence rates were calculated as the number of BC per 100 000 person-years. We calculated cumulative BC incidence estimates by constructing time-to-event curves using Kaplan–Meier estimates for each group (no recall, FPR score 3, and FPR score 4/5 recall) and used the log-rank test to calculate corresponding *P*-values.

Table 1 Using location data and the timing of diagnosis to classify breast cancers after false-positive recall as either false-negative diagnostic assessment or new cancers

	Was BC found in the segment for which recall was recommended at T_0?	
	No	Yes
When was BC diagnosed?		
Before the first screening that followed T_0 recall	New BC	FNDA
At the first screening that followed T_0 recall	New BC	FNDA
At the second or later screening after T_0 recall, with the intermediate screenings being positive	New BC	FNDA
At the second or later screening after T_0 recall, with the intermediate screenings being negative	New BC	New BC

BC, breast cancer; FNDA, false-negative diagnostic assessment at T_0.

Table 2 Baseline characteristics and follow-up of study participants

	No recall	FPR score 3	FPR score 4/5	*P*-value[a]
n	281 247	10 597	840	
Mean age in years (SD)	58.2 (5.7)	57.3 (5.9)	58.1 (5.8)	< 0.001
Screening round				< 0.001
Initial	95 953 (34.1)	5452 (51.4)	395 (47.0)	
Successive	185 294 (65.9)	5145 (48.6)	445 (53.0)	
Breast density category[b]				< 0.001
A (≤ 25%)	45 268 (16.1)	946 (9.0)	79 (9.4)	
B (26–50%)	150 372 (53.5)	5829 (55.2)	412 (49.3)	
C (51–75%)	70 082 (25.0)	3276 (31.0)	302 (36.1)	
D (>75%)	15 186 (5.4)	508 (4.8)	43 (5.1)	
Missing	339 (–)	38 (–)	4 (–)	
Mean follow-up in years (SD)	6.9 (0.8)	6.8 (0.8)	6.8 (1.0)	
Person-years	1 927 608	72 145	5701	
Breast cancers	5708	384	59	
Incidence rate (95% CI)[c]	296.1 (288.5–303.9)	532.3 (480.5–588.1)	1034.9 (788.7–1333.0)	
Classification				
New cancer	5,708 (100.0)	261 (78.1)	33 (73.3)	
FNDA	0 (0.0)	73 (21.9)	12 (26.7)	
Unclassifiable	0 (–)	50 (–)	14 (–)	

Flanders, Belgium, 2005–2012.
Data are *n* (%) unless noted differently.
FNDA, false-negative diagnostic assessment at T_0; FPR, false-positive recall at T_0.
[a]*P*-value based on χ^2-test or one-way analysis of variance (for age).
[b]Categories based on percentage fibroglandular tissue.
[c]Per 100 000 person-years.

Cox proportional hazards models were used to calculate hazard ratios (HRs). Confounding variables were candidates for purposeful selection if they had *P*-values of 0.20 or less in univariate Cox proportional hazards models; age (years), screening round (initial vs. successive) and breast density (categories visible in Table 2) were tested. Forward model selection was carried out using the likelihood ratio test, with significance set at *P* of 0.01 or less. Multivariate HRs were thus adjusted for confounding variables. A sensitivity analysis was performed that presumed all unclassifiable BC were FNDA.

The risk that a BC that was seen at screening would be missed at assessment was estimated as follows:

$$\text{Risk} = \frac{\text{number of FNDA}}{\text{number of FNDA} + \text{number of SDC}}$$

The FNDA rate was calculated as the percentage of FNDA among all women recalled to assessment.

The χ^2-test was used to compare differences in tumor behavior, size, grade, and nodal status between the BC that were detected at DA (SDC), and those that were missed at DA (FNDA).

Analyses and data storage were conducted using the software package Stata, version 13 (StataCorp., College Station, Texas, USA); all statistical tests were two sided.

Results

Sample size and follow-up

After exclusions, the group without recall included 281 247 women (1 927 608 person-years), the FPR 3 group included 10 597 women (72 145 person-years), and the FPR 4/5 group included 840 women (5701 person-years) [Supplementary Fig. S1 (Supplemental digital content 1, *http://links.lww.com/EJCP/A122*), Table 2]. The mean follow-up duration in years (± SD) was 6.9 (± 0.8), 6.8 (± 0.8), and 6.8 (± 1.0) in the no recall group, FPR 3 group, and FPR 4/5 group, respectively. Baseline characteristics of these groups are presented in Table 2; the

mean age in years (± SD) was 58.2 (± 5.7), 57.3 (± 5.9), and 58.1 (± 5.8) in the no recall group, the FPR 3 group, and the FPR 4/5 group, respectively. Women in an initial round were more likely to have a false-positive recall, as were women with higher mammographic breast density.

Primary outcome and categorization of breast cancer

BC incidence is presented in Table 2; a total of 6151 BC were found, of which 443 (7.2%) occurred in the FPR groups. Of those 443 BC, the classification status could not be determined for 64 BC (14.5%). Of the 379 remaining BC, 294 (77.6%) were new, whereas 85 (22.4%) were FNDA.

Risk estimates

Figure 1 presents time-to-event curves for the primary endpoint (BC incidence). Dotted lines reflect the curves that take all BC into account, whereas the solid lines exclude FNDA.

When all BC were included, the 5-year cumulative BC incidence was significantly different ($P < 0.001$) among the groups: 1.4% in the no recall group, 2.7% in the FPR 3 group, and 5.7% in the FPR 4/5 group. After excluding FNDA, the differences remained significant, but 5-year cumulative BC incidence in the FPR groups decreased to 2.1% (FPR 3) and 4.3% (FPR 4/5).

The final multivariate Cox model included the following variables as confounders: age, screening round, and breast density. Figure 2 shows the HRs; the no-recall group always serves as a reference. Overall, the HR of any type of FPR was 1.90 (95% CI: 1.72–2.09) when all BC were included and decreased to 1.53 (95% CI: 1.38–1.70) after excluding FNDA. This decrease from 1.90 to 1.53 represents a 41% drop.

The FPR 4/5 group had higher HRs compared with the FPR 3 group. Before excluding FNDA, HRs were 3.43 (95% CI: 2.65–4.43) and 1.77 (95% CI: 1.60–1.97), respectively, which decreased to 2.73 (95% CI: 2.05–3.64) and 1.44 (95% CI: 1.28–1.61). In sensitivity analysis, which excluded the unclassifiable BC as well, the risk estimates were 1.92 (95% CI: 1.36–2.70) and 1.21 (95% CI: 1.06–1.37), respectively.

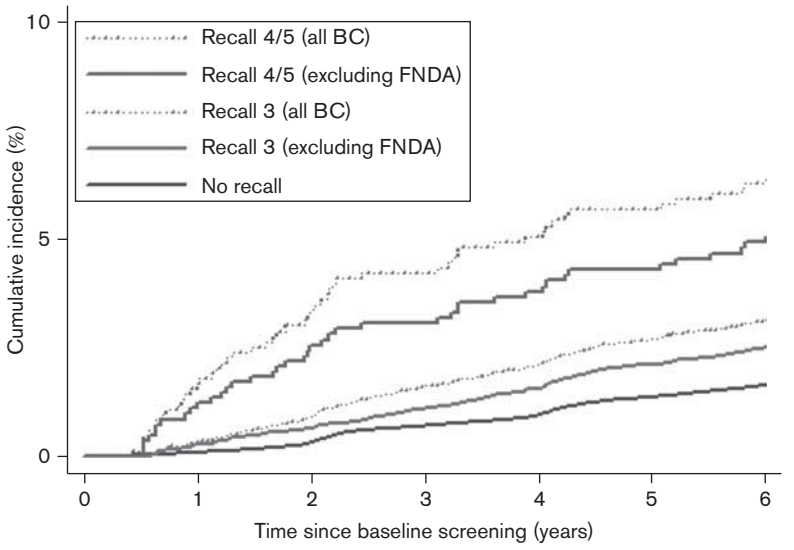
Cancers missed at diagnostic assessment

There was a 4.3% risk that a BC that was seen at screening would subsequently be missed at DA. This corresponds to an FNDA rate of 0.64% among women recalled to assessment. There were no significant differences in tumor characteristics (Table 3).

Discussion

To our knowledge, this is the second study that uses time-to-event analysis to investigate BC risk after FPR,

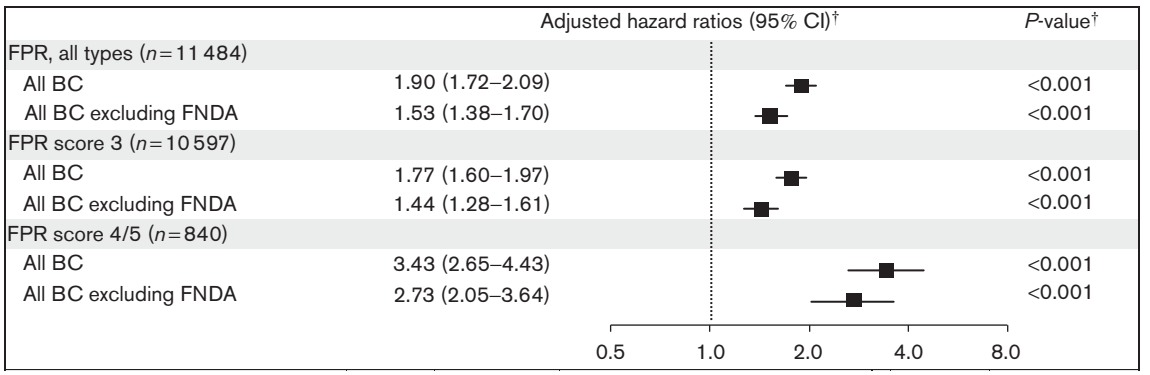
Fig. 1



Number at risk							
Recall 4/5 (all BC)	840	822	804	793	772	761	748
Recall 4/5 (excluding FNDA)	828	814	799	791	771	761	748
Recall 3 (all BC)	10 597	10 531	10 437	10 325	10 229	10 117	9991
Recall 3 (excluding FNDA)	10 524	10 465	10 391	10 306	10 215	10 108	9986
No recall	281 247	280 398	278 802	276 619	274 762	272 312	270 118

Time-to-event curves for the primary endpoint (BC). Flanders, Belgium, 2005–2012. BC, breast cancer; FNDA, false-negative diagnostic assessment at T_0.

Fig. 2



Adjusted hazard ratio (HR) from Cox proportional hazards models for breast cancer screening result. Flanders, Belgium, 2005–2012. †HR with adjustment for age, breast density, and screening round. Reference category was always the no-recall group (not visible). BC, breast cancer; CI, confidence interval; FNDA, false-negative diagnostic assessment at T_0; FPR, false-positive recall at T_0.

Table 3 Tumor characteristics of breast cancers found as screen-detected cancer or false-negative diagnostic assessment

	SDC	FNDA	*P*-value[a]
Total number of breast cancers	1873 (100.0)	85 (100.0)	
Behavior			0.206
DCIS	338 (18.1)	20 (23.5)	
Invasive	1529 (81.9)	65 (76.5)	
Missing	6 (–)	0 (–)	
Tumor size[b]			0.716
≤10 mm	455 (33.3)	22 (35.5)	
>10 mm	913 (66.7)	40 (64.5)	
Missing	161 (–)	3 (–)	
Grade[b]			0.788
Low	284 (20.3)	11 (18.0)	
Intermediate	686 (49.1)	29 (47.5)	
High	426 (30.5)	21 (34.4)	
Missing	133 (–)	4 (–)	
Nodal status[b]			0.770
Negative	999 (75.0)	45 (73.3)	
Positive	364 (25.0)	15 (26.7)	
Missing	166 (–)	5 (–)	

Flanders, Belgium, 2005–2012.
Data are *n* (%).
DCIS, ductal carcinoma *in situ*; FNDA, false-negative diagnostic assessment at T_0; SDC, screen-detected cancer.
[a]*P*-value based on χ^2-test.
[b]For invasive cancers only.

but the first that also quantifies the effect of FNDA. Women with a history of FPR were at an increased risk of developing BC compared with women who were not recalled, but 41% of the risk increase was due to FNDA. The type of recall was clearly correlated with the risk. After excluding FNDA, an FPR with BI-RADS 4 or 5 had an HR of 2.73 (95% CI: 2.05–3.64), whereas the risk after an FPR with BI-RADS 3 had a lower HR (HR 1.44; 95% CI: 1.28–1.61).

When we included all BC, our risk estimates (HR 1.90; 95% CI: 1.72–2.09) were comparable to the findings in other countries: a Spanish study (Castells *et al.*, 2013) found odds ratios between 1.81 and 2.69, a Danish study (von Euler-Chelpin *et al.*, 2012) found a relative risk (RR) of 1.67, and two studies performed in the USA (Barlow *et al.*, 2006; Henderson *et al.*, 2015) found HR of up to 1.76 and RR of 1.69. When we excluded FNDA, our risk estimates (HR 1.53; 95% CI: 1.38–1.70) were slightly higher than those in the only other study (odds ratio 1.27) that we found that also excluded FNDA cancers (von Euler-Chelpin *et al.*, 2014). When FNDA was not excluded, this particular study also had the lowest risk estimates of the previous studies. Their percentage of FNDA among all cancers after FPR was similar (24.4 vs. our 22.4%).

This additional BC risk should be addressed in three ways: decreasing FNDA, informing patients, and considering increased surveillance of women with FPR. As regards decreasing FNDA, clearly there is room for improvement of diagnostic assessment. Of all cancers that were seen at screening, 4.3% were subsequently missed at DA. Avoiding such FNDA would increase the BC detection rate from 6.3 to 6.6‰. A previous study found that the majority of FNDA are due to erroneously interpreting suspicious lesions as benign, disregarding a radiologist's advice to perform a biopsy, and false-negative biopsy results (Burrell *et al.*, 2001; Duijm *et al.*, 2004; Ciatto *et al.*, 2007). Reasons for not performing biopsy include a surgeon's refusal to perform biopsy even after a radiologist's explicit advice to do so (Duijm *et al.*, 2004). It has also been suggested that DA should be performed in special breast care units (Purushotham *et al.*, 2001; Haward *et al.*, 2003). Periodically evaluating the quality of DA together with regular feedback might be the best way forward. As regards informing patients, increasing emphasis is placed on providing absolute risk indicators such as 5-year cumulative BC incidence. Such absolute risk indicators are necessary for informed decision making and provide much more information than the RR (Mathieu *et al.*, 2007; Akl *et al.*, 2011). In absolute terms, the increase in risk corresponded to a 5-year

cumulative BC incidence of 5.7% after a BI-RADS 4 or 5 FPR, versus 1.4% for women who were not recalled. As regards yearly radiological evaluation, this would require further studies to prove effectiveness.

Our study has several strengths and limitations. First, our conclusions rely on the completeness of BCR and CBSS data. Oncological care programs in hospitals and pathology laboratories are required by law to supply the BCR with their data. For the Flemish region, cancer incidence data were available from 1999 until the end of 2012. The completeness of the BCR as regards BC was evaluated using the independent database method and was estimated to be 99.7% until the end of 2012 (Henau *et al.*, 2015). One of the main sources of data for the CBSS is the National Registry of the Ministry of the Interior, which contains information on all people who were at some point registered with a municipality in Belgium. Cross-linking of the National Registry with the CBSS's other sources is performed to obtain data on people who were not registered in the National Registry. The CBSS is thought to have a degree of completeness in excess of 99%.

Second, we defined SDCs as BCs that are diagnosed within 3 months of DA. This method (hereafter referred as the 3-month algorithm) is more restrictive than the definition of SDC that some authors use (all BC found within 12 months of a positive screening are SDC) (Christiansen *et al.*, 2000; Henderson *et al.*, 2015), but in other manuscripts it is often unclear how SDC are defined. The 3-month algorithm has the advantage of not overestimating the number of SDC, but it is possible that we underestimated the number of SDC. The result of this would be that a woman with an SDC is seen as a woman with FPR followed by BC. To check this, we compared the DA conclusion sent by physicians with the results of the 3-month algorithm. Of the 13 334 women who were recalled (who had given informed consent and did not have BC before screening, see flow chart), we excluded 3112 women (23.4%) for whom we did not have a DA conclusion sent by physicians (BC, no BC). When the DA conclusion sent by physicians was compared with the conclusion of the 3-month algorithm for the remaining 10 198 women, we found that 1550 women had an SDC according to both methods, but the DA conclusion sent by physicians found an additional 16 SDC. Three of these (18.8%) could not be SDC as they were not found in the location of the lesion seen at screening. The remaining 13 (81.2%) had a median time between DA and diagnosis of 6 months. These 13 represent 0.8% of all SDC. Although underestimation exists, we conclude that it is limited.

Third, to classify BC after FPR as either FNDA or new cancer, we used an algorithm that mainly used location data, which are routinely collected in both the BCR and the screening program. Compared with radiological review of all files, such an algorithm has several important advantages, including avoiding radiologist subjectivity, a much lower workload, and more complete data (Blanch *et al.*, 2014). We cannot exclude that some cases were categorized as FNDA, although these BC were in reality unrelated to the lesion seen at screening, meaning we would overestimate the FNDA rate. However, our FNDA rate of 0.64% among women recalled to assessment is situated between the 0.50% estimate found by Ciatto *et al.* (2007), the 0.56% found by Burrell *et al.* (2001), and the 1.5% found by von Euler-Chelpin *et al.* (2014).

Moreover, the proportion of all BC that is seen at screening but missed at DA (4.3%) is very similar to the estimate found by Ciatto *et al.* (2007). We conclude that, although some of the cancers classified as FNDA may have in fact been new cancers, this is likely to be limited.

Fourth, noninformative censoring is an important assumption in time-to-event analysis. However, when a woman dies (due to any cause) she will no longer be at risk for the primary outcome. The events of dying and of developing BC are therefore not independent. As the naïve Kaplan–Meier estimator assumes independence of all events and thus censors the deaths, it can lead to an overestimated risk of disease by failing to account for the competing risk of death. This is mostly a problem when the competing risk of death is high due to, for instance, increased age or comorbidities. Even though a standard Cox proportional hazards regression is not adequate in a competing risk setting it can still be used to assess HRs of FPR (Haesook, 2007; Putter *et al.*, 2007). The resulting HR will be slightly biased if the competing event of death due to any cause is very rare. The advantages of using the Cox model in this setting are that it makes multivariate modeling possible and its HR are relatively easy to interpret. To evaluate whether our results could be overestimated we compared application of the Kaplan–Meier method and the cumulative incidence competing risk method (which calculates cumulative incidence accounting for the presence of competing risks) for each group in the study (Verduijn *et al.*, 2011). The Kaplan–Meier method overestimated the 5-year risk for BC by less than 0.06% in each group. We conclude that this bias is very limited and does not influence our results.

Conclusion

Women with FPR are at an increased risk of developing BC in the years after screening, and the type of recall is clearly correlated with the magnitude of the risk. A part of the risk is explained by FNDA, but the risk remains significant after excluding FNDA.

Acknowledgements

The authors thank Patrick Beyltjens for data cleaning and Anne-Marie Depoorter for her valuable comments on the text.

Conflicts of interest

There are no conflicts of interest.

References

Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, *et al.* (2011). Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst Rev* **3**:CD006776.

Alamo-Junquera D, Murta-Nascimento C, Macia F, Bare M, Galceran J, Ascunce N, *et al.* (2011). Effect of false-positive results on reattendance at breast cancer screening programmes in Spain. *Eur J Public Health* **22**:404–408.

Bangsboll-Andersen S, Vejborg I, von Euler-Chelpin M (2008). Participation behaviour following a false positive test in the Copenhagen mammography screening programme. *Acta Oncol* **47**:550–555.

Barlow W, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, *et al.* (2006). Prospective BC risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst* **98**:1204–1214.

Blanch J, Sala M, Ibanez J, Domingo L, Fernandez B, Otegi A, *et al.* (2014). Impact of risk factors on different interval cancer subtypes in a population-based breast cancer screening programme. *Plos One* **9**:1–10.

Burrell HC, Evans AJ, Wilson AR, Pinder SE (2001). False-negative breast screening assessment: what lessons can we learn? *Clin Radiol* **5**:385–388.

Castells X, Molins E, Macia F (2006). Cumulative false positive recall rate and association with participant related factors in a population based BC screening programme. *J Epidemiol Community Health* **60**:316–321.

Castells X, Roman M, Romero A, Blanch J, Zubizarreta R, Ascunce N, *et al.* (2013). Breast cancer detection risk in screening mammography after a false-positive result. *Cancer Epidemiol* **37**:85–90.

Christiansen CL, Wang F, Barton MB, Kreuter W, Elmore JG, Gelfand AE, Fletcher SW (2000). Predicting the cumulative risk of false-positive mammograms. *J Natl Cancer Inst* **92**:1657–1666.

Ciatto S, Houssami N, Ambrogetti D, Bonardi R, Collini G, Del Turco MR (2007). Minority report – false-negative breast assessment in women recalled for suspicious screening mammography: imaging and pathological features, and associated delay in diagnosis. *Breast Cancer Res Treat* **1**:37–43.

Duijm LE, Groenewoud JH, Jansen FH, Fracheboud J, van Beek M, de Koning HJ (2004). Mammography screening in the Netherlands: delay in the diagnosis of breast cancer after breast cancer screening. *Br J Cancer* **91**:1795–1799.

Goossens M, Van Hal G, Van der Burg M, Kellen E, Van Herck K, De Grève J, *et al.* (2014). Quantifying independent risk factors for failing to rescreen in a breast cancer screening program in Flanders, Belgium. *Prev Med* **69**:280–286.

Haesook T (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clin Cancer Res* **38**:559–565.

Haward R, Amir Z, Borrill C, Dawson J, Scully J, West M, Sainsbury R (2003). Breast cancer teams: the impact of constitution, new cancer workload, and methods of operation on their effectiveness. *Br J Cancer* **89**:15–22.

Henau K, Van Eycken E, Silversmit G, Pukkala E (2015). Regional variation of incidence for smoking and alcohol related cancers in Belgium. *Cancer Epidemiol* **39**:55–65.

Henderson L, Hubbard R, Sprague B, Zhu W, Kerlikowske K (2015). Increased risk of developing breast cancer after a false-positive screening mammogram. *Cancer Epidemiol Biomarkers Prev* **24**:1882–1889.

Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, *et al.* (2012). False-positive results in mammographic screening for BC in Europe: a literature review and survey of service screening programmes. *J Med Screen* **19**:57–66.

Hosmer DW, Lemeshow S (2000). *Applied logistic regression*, 2nd ed. New York, NY: John Wiley and Sons.

Intermutualistisch Agentschap (2010). Programma Borstkankerscreening Verslag nr. [7th Breast cancer screening report]. Brussels.

Marmot M, Altman D, Cameron D, Dewar J, Thompson S, Wilcox M (2013). The benefits and harms of BC screening: an independent review. *Br J Cancer* **108**:2205–2240.

Martens P (2015). *Jaarrapport 2015 Bevolkingsonderzoek Borstkanker [Annual report of the Flemish breast cancer screening programme]*. Brugge: Centrum voor Kankeropsporing.

Mathieu E, Barratt A, Davey H, McGeechan K, Howard K, Houssami N (2007). Informed choice in mammography screening: a randomised trial for a decision aid for 70-year-old women. *Arch Int Med* **167**:2039–2046.

Maxwell A, Beattie C, Lavelle J, Lyburn I, Sinnatamby R, Garnett S, Herbert A (2013). The effect of false positive breast screening examinations on subsequent attendance: retrospective cohort study. *J Med Screen* **20**: 91–98.

Peeters P, Mravunac M, Hendriks J, Verbeek AL, Holland R, Vooijs PG (1988). Breast cancer risk for women with a false positive screening test. *Br J Cancer* **58**:211–212.

Perry N, Broeders M, deWolf C, Tornberg S, Holland R, von Karsa L (2008). European guidelines for quality assurance in BC screening and diagnosis. Fourth edition. *Ann Oncol* **19**:4–22.

Purushotham A, Pain S, Miles D, Harnett A (2001). Variations in treatment and survival in breast cancer. *Lancet Oncol* **2**:719–725.

Putter H, Fiocco M, Geskus R (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statist Med* **26**:2389–2430.

Verduijn M, Grootendorst D, Dekker F, Jager K, le Cessie S (2011). The analysis of competing events like cause-specific mortality – beware of the Kaplan–Meier method. *Nephrol Dial Transplant* **26**:56–61.

von Euler-Chelpin M, Risor L, Thorsted B, Vejborg I (2012). Risk of breast cancer after false-positive test results in screening mammography. *J Natl Cancer Inst* **104**:682–689.

von Euler-Chelpin M, Kuchiki M, Vejborg I (2014). Increased risk of BC in women with false-positive test: the role of misclassification. *Cancer Epidemiol* **38**:619–622.