Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Validation of an Algorithm to Ascertain Late Breast Cancer Recurrence Using Danish Medical Registries

This article was published in the following Dove Press journal:
Clinical Epidemiology

Rikke Nørgaard Pedersen[1]
Buket Öztürk[1]
Lene Mellemkjær[2]
Søren Friis[2]
Trine Tramm[3]
Mette Nørgaard[1]
Deirdre P Cronin-Fenton[1]

[1]Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark; [2]Danish Cancer Society Research Center, Copenhagen, Denmark; [3]Department of Pathology, Aarhus University Hospital, Aarhus, Denmark

Purpose: About 70% of women with breast cancer survive at least 10 years after diagnosis. We constructed an algorithm to ascertain late breast cancer recurrence—which we define as breast cancer that recurs 10 years or more after primary diagnosis (excluding contralateral breast cancers)—using Danish nationwide medical registries. We used clinical information recorded in medical records as a reference standard.

Methods: Using the Danish Breast Cancer Group clinical database, we ascertained data on 21,134 women who survived recurrence-free 10 years or more after incident stage I–III breast cancer diagnosed in 1987–2004. We used a combination of Danish registries to construct the algorithm—the Danish National Patient Registry for information on diagnostic, therapeutic and procedural codes; and cancer diagnoses from the Danish Pathology Registry, the Danish Cancer Registry and the Contralateral Breast Cancer database. To estimate the positive predictive value (PPV), we selected 105 patients who, according to our algorithm, had late recurrence diagnosed at Aarhus University Hospital. To estimate the sensitivity, specificity and negative predictive value (NPV), we selected 114 patients diagnosed with primary breast cancer at Aalborg University Hospital. We abstracted clinical information on late recurrence for patients with medical record-confirmed late recurrence at Aarhus University Hospital.

Results: Our algorithm had a PPV of late recurrence of 85.7% (95% CI: 77.5–91.3%), a sensitivity of 100.0% (95% CI, 39.8–100.0%), a specificity of 97.3 (95% CI, 92.2–99.4) and a NPV of 100% (95% CI, 96.6–100.0%).

Conclusion: Our algorithm for late recurrence showed a moderate to high PPV and high sensitivity, specificity and negative predictive value. The algorithm could be an important tool for future studies of late breast cancer recurrence.

Keywords: algorithm, late breast cancer recurrence, breast cancer neoplasm, PPV, sensitivity

Introduction

In 2018, about 2.1 million women were diagnosed with breast cancer worldwide, accounting for 1 in 4 cancer cases among women.[1] The aging population and the improvements in diagnosis and treatment have increased the number of breast cancer (BC) survivors.[2–5] Today, close to 70% can expect to live for at least ten years after primary diagnosis and treatment.[6] Therefore, it is necessary to extend the focus to identify patients at risk of late breast cancer recurrence, which we define as breast cancer recurrence 10 years or more after the primary breast cancer diagnosis.

Correspondence: Rikke Nørgaard Pedersen
Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark
Tel +4587167212
Email rikp@clin.au.dk

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/CLEP.S269962

Breast cancer, and especially estrogen receptor (ER) positive breast cancer, has the ability to recur many years after primary diagnosis.[7] Recently, a meta-analysis of 88 trials involving >60,000 women with estrogen receptor positive tumors, reported that distant breast cancer recurrences continued to occur at a steady rate for at least 20 years after diagnosis. The risk of distant recurrence was associated with the original tumor size and lymph node status, whereby the risk increased from 13% for TNM (tumor node metastasis) stage T1N0 to 41% for T2N4-9 stage.[8] Another study including 3128 breast cancer patients found a cumulative risk of distant recurrence of 21% after 15 years of follow-up.[9]

Information on the epidemiology of late breast cancer recurrence are important for research purposes but also for surveillance, prediction of prognosis, and monitoring improvements in treatment and care. In Denmark, several registries record breast cancer recurrence, among others, the Danish National Patient Registry (DNPR) and the Danish Breast Cancer Group clinical database (DBCG). However, a specific code for recurrence was first implemented in the DNPR from 2012 and recurrences are only routinely registered in the DBCG database within the first 10 years after diagnosis.[10] Recently two Danish studies combined data from the DNPR and the Danish National Pathology Registry to construct an algorithm to identify patients with colorectal cancer[11] and bladder cancer[12] recurrence, respectively. Two other Danish studies developed and validated algorithms to ascertain breast cancer recurrence using Danish administrative data.[13,14] The latter is from our previous work where we found a positive predictive value (PPV) of 71%.[14] Neither of these algorithms focused on late breast cancer recurrence.

We therefore used a combination of Danish registries to construct an algorithm to ascertain late breast cancer recurrence among Danish women diagnosed with stage I–III operable breast cancer by modifying our previous breast cancer algorithm to identify late breast cancer recurrence. We aimed to: 1) examine the PPV for our late breast cancer recurrence algorithm; 2) estimate the sensitivity, specificity and negative predictive value (NPV) of the algorithm; 3) assess clinical information recorded in the medical charts of late breast cancer recurrence patients regarding diagnostic procedures, anatomical location and initial treatment of the late breast cancer recurrence.

Materials and Methods

Denmark is divided into five regions that are comparable with respect to demographic and socioeconomic characteristics as well as hospital structure and health care usage.[15] The Danish National Health Service provides free universal tax-supported healthcare, guaranteeing free access to medical treatment for all residents.[16]

Data Sources

We used the civil personal registration (CPR) number—a unique personal identification number used in all Danish registries—to enable individual-level data linkage across the registries and databases.[16] These included the DBCG, the Danish Civil Registration System (CRS), the DNPR, the Danish Cancer Registry (DCR), the Danish Pathology Registry and the Contralateral Breast Cancer database.

The DBCG database[17] has registered almost all women with invasive breast cancer in Denmark since 1976. The completeness of registration is approximately 95%.[10] Standardized data abstraction forms are used to register prospectively recorded data on patient, tumor and treatment characteristics. All patients undergoing breast cancer surgery are followed-up twice yearly for the first five years after diagnosis and annually the next five years as long as they are in active therapy.[18] The CRS was established in 1968 and includes registration of data such as vital status and emigration.[16] The DNPR has recorded information on non-psychiatric inpatient admissions since 1977 and outpatient hospital contacts and psychiatric admissions since 1995. Each discharge or outpatient visit is recorded with one primary diagnosis and one or more secondary diagnoses classified according to the International Classification of Diseases (ICD) (eighth revision until 1994 and tenth revision thereafter).[19,20] Surgeries in the DNPR are coded according to the Danish Version of the Nordic Medico-Statistical Committee Classification of Surgical Procedures.[20] Cancer treatments are primarily registered with a Health Care Classification System code (SKS treatment code).[20] The Danish Pathology Registry has routinely recorded information on all pathology examinations in Denmark since 1997 and is complete since 2000. This Registry uses the Systematized Nomenclature of Medicine (SNOMED) classification system enabling identification of specimens of malignant morphology (codes M8 and M9).[21] The DCR has recorded incident cases of cancer in Denmark since 1943 and has been shown to have accurate and virtually complete information on incident cancer cases.[22] However, the registration of contralateral breast cancer (CBC) (following historical coding rules) is insufficient for research purposes.

Consequently, a database with CBCs during 1978–2013 was established.[23]

Source Population

The source population consisted of all women in Denmark diagnosed with stage I–III operable breast cancer in DBCG between January 1, 1987 and December 31, 2004 who were alive, living in Denmark, and without a recurrence or second cancer (CBC or other new primary tumor) 10 years after diagnosis. Information about recurrences or a second cancer within the first 10 years was obtained from DBCG. We also used the CBC database to exclude any patients diagnosed with a CBC the first 10 years after primary breast cancer diagnosis. Furthermore, we used the DNPR to exclude patients with a metastases code within the first ten years after breast cancer diagnosis.

Information about emigration and vital status was obtained from the CRS.

Study Population

To ensure sufficient numbers of late breast cancer recurrence cases, we used two different study populations to address our aims (Figure 1). To compute PPVs, we included all breast cancer patients in the source population who, according to our algorithm, developed a late breast cancer recurrence and were diagnosed at Aarhus University Hospital. Hereafter, we refer to this study population as study population I. We stratified the patients by estrogen receptor (ER) status (ER-positive, ER-negative and ER-missing) obtained from DBCG, due to current cut-off guidelines at the time of diagnosis. Estrogen receptor status was first routinely tested from 1997. We assigned a random number to each patient in the three strata and then ranked the patients within each strata according to the random number. We randomly selected 75 patients in the ER positive strata (60%), 13 patients in the ER negative strata (10%) and 38 patients in the ER missing strata (30%) yielding a total of 126 patients and retrieved their medical records from the hospital archives. We stratified by ER status to ensure sufficient ER positive patients as these have highest risk of late recurrence.[24,25]

To compute sensitivity, specificity and NPV, we included all breast cancer patients from the source population who were diagnosed at Aalborg University Hospital during 1987–2004 regardless of any late breast cancer recurrence status according to our algorithm. Hereafter, we refer to this

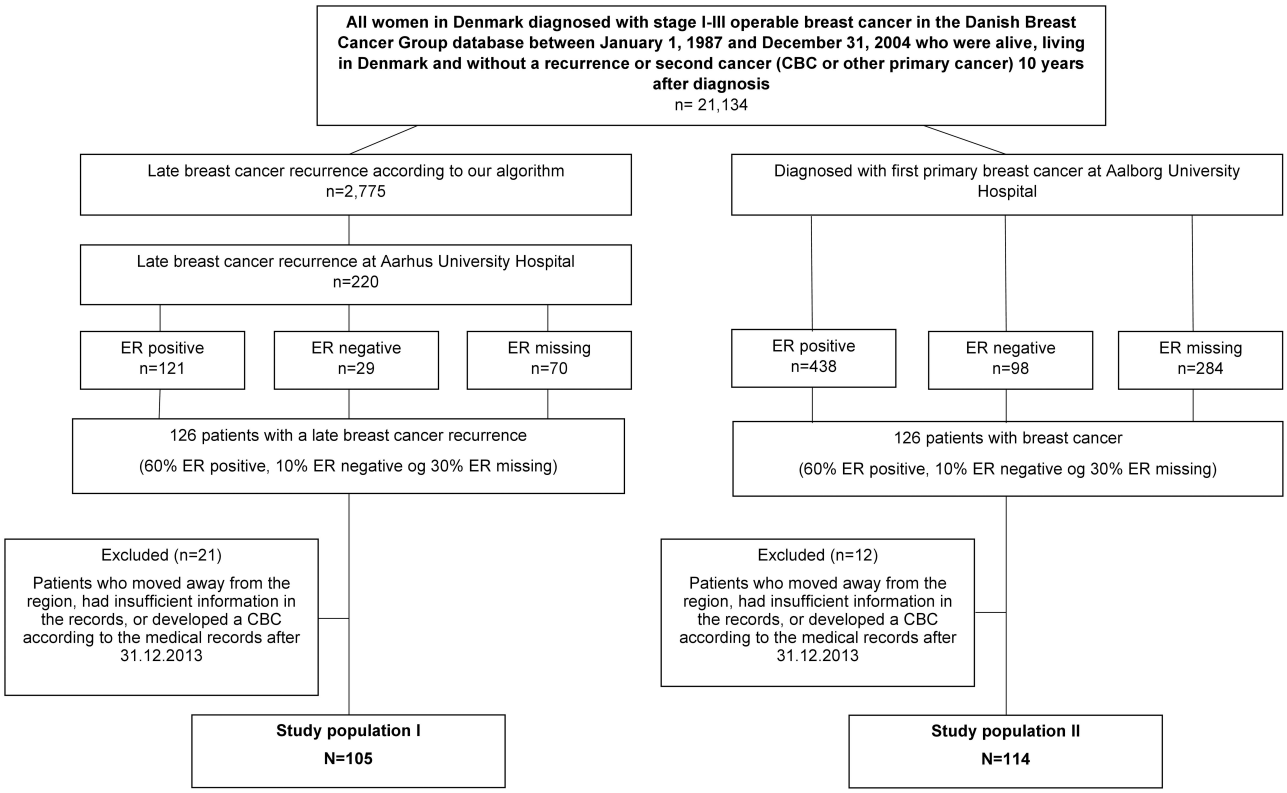

Figure 1 Flowchart of study populations.
Abbreviations: CBC, contralateral breast cancer; ER, estrogen receptor.

study population as study population II. Again, we stratified by ER status, ranked the patients within each strata, and randomly selected 75 patients in the ER positive strata, 13 patients in the ER negative strata and 38 patients in the ER missing strata yielding 126 breast cancer patients in total.

Since the CBC database was updated until 2014, we excluded women who developed a CBC according to the medical records after December 31, 2013.

We retrieved information on late recurrence from medical records.

Algorithm for Late Breast Cancer Recurrence

Late breast cancer recurrence was defined as any local, regional or distant recurrent breast cancer diagnosed ≥10 years after primary breast cancer diagnosis (excluding contralateral breast cancers). We identified recurrences during follow-up if at least one of the following criteria was registered 10 years or more after the primary breast cancer surgery (Figure 2 and Appendix): I) DNPR-registered metastases codes (ICD10 DC76-DC80); II) Pathology-registered SNOMED combinations. Combinations were T code (topography/location) in the breast with morphology codes M8 or M9 with 4, 6, 7, 9 in the fifth position (eg, M8XXX4), 2) any T code (excluding the breast T codes) with morphology codes M8 or M9 with the numbers 6 or 9 in the fifth position. III) DNPR-registered cancer-directed treatment codes (SKS treatment codes) including radiotherapy (BWG), chemotherapy (BWHA) and endocrine therapy (BWHC, BHHH, BOHJ13); IV) DNPR-registered surgical codes (according to the Danish Version of the Nordic Medico-Statistical Committee Classification of Surgical Procedures) including mastectomy (KHAC), breast conserving surgery (KHAB), and resection of the chest wall (KGAE16); V) A code specific for local breast cancer recurrence (ICD10 DC509X) or a code specific for recurrence surgery in the DNPR (KHAF).

To avoid inclusion of a CBC or another new primary cancer in the late breast cancer algorithm, we disregarded the late breast cancer recurrence diagnosis if a second cancer diagnosis (CBC or other new primary tumor) was registered in the DCR, DNPR or the Contralateral Breast Cancer Database up to 90 days after the algorithm criteria date. Another new primary tumor was defined as a new primary cancer that was different from breast cancer (ICD10 C50) and non-melanoma skin cancer (ICD10 C44).

We noted that recommendations to prolong endocrine therapy to ten years after diagnosis[26] were disseminated into clinical practice during our study period. We therefore included a stipulation in the algorithm that endocrine therapy could be an indicator of late recurrence only if the code was present more than two years after latest endocrine therapy code.

The recurrence date estimated by the algorithm was defined as the date of the first registered algorithm criteria. The latest date a recurrence could be registered was December 31, 2017.

Covariates

We obtained information from the DNPR on potential comorbid diseases up to ten years before primary breast

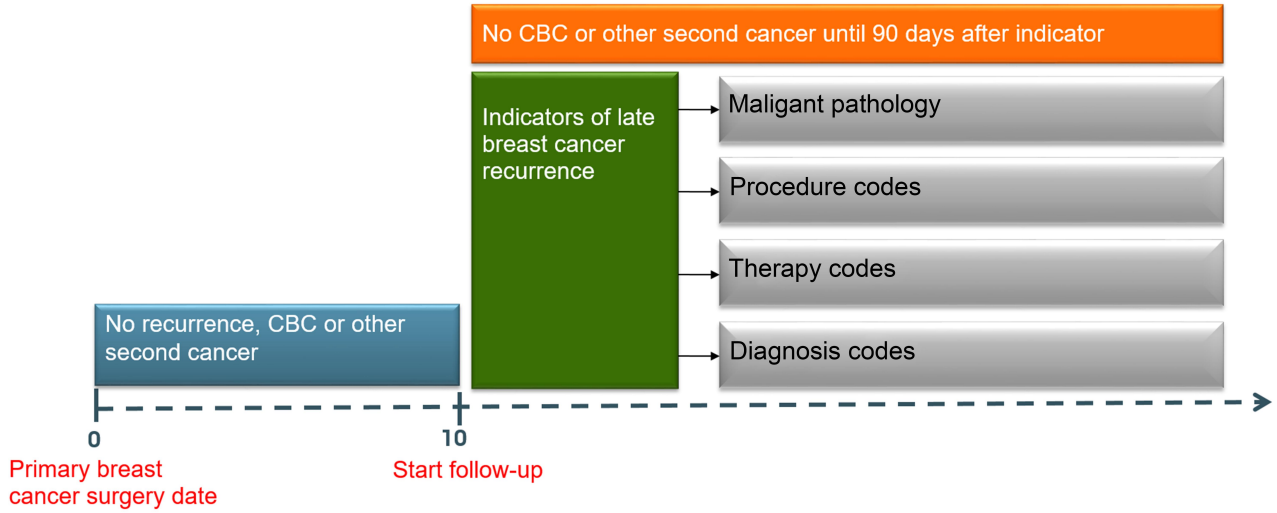

Figure 2 Overview of the algorithm.
Abbreviation: CBC, contralateral breast cancer.

cancer diagnosis and summarized them using the Charlson Comorbidity Index (CCI)[27] modified to exclude breast cancer diagnoses. From the DBCG, we also obtained information about age, menopausal status at diagnosis, type of surgery, WHO histological tumor type and grade, lymph node status, tumor size, ER status, receipt of adjuvant chemotherapy, endocrine therapy (ET) and/or radiation therapy. Stage was calculated using the TNM staging system.

Reference Standard

To validate the late breast cancer recurrence algorithm, we used medical record review as the reference standard. We retrieved medical records from Aarhus University Hospital in the Region of Central Denmark and Aalborg University Hospital in the Region of Northern Denmark. We developed a medical record abstract form and accompanying codebook, which was used to guide the review. Two reviewers retrieved data on late breast cancer recurrence from the medical record. Patients were considered to have a late breast cancer recurrence if their medical record documented a pathological diagnosis or a clinical diagnosis based on mammography, ultrasound, X-ray, CT scan or MRI. Data from the medical records were entered into a secure REDCap data collection platform.

Statistical Analyses

For study populations I and II, we present descriptive characteristics outlining the distribution of patient, tumor- and treatment characteristics of the primary breast cancer.

We computed the PPV and associated 95% confidence intervals (95% CI) of our late breast cancer recurrence algorithm as the proportion of cases identified by our algorithm that were confirmed by the medical record review. PPVs were estimated for the overall algorithm and for the individual algorithm criteria and patient- and tumor characteristics at baseline, respectively. We used Lin´s concordance correlation coefficient (CCC)[28] to find the strength of agreement between the date of recurrence identified by our algorithm and the date identified by the medical chart review. The sensitivity and associated 95% CI was estimated with the numerator being the number of patients listed with a late breast cancer recurrence in both the algorithm and in the medical records, and the denominator being the total number of patients with late breast cancer recurrence documented in the medical records. The specificity was estimated with the numerator being the number of patients listed without a late breast cancer recurrence in both the algorithm and medical records, and the denominator being the total number of patients without a late breast cancer recurrence documented in the medical records. We computed the NPV and associated 95% CI as the proportion of patients without a breast cancer recurrence according to our algorithm and confirmed by the medical record review. From the sensitivity and specificity, we calculated positive and negative likelihood ratios, and associated 95% CIs.

To further describe the patients with late breast cancer recurrence, we reported the following characteristics of record-confirmed late breast cancer recurrence patients from Aarhus University Hospital: diagnostic procedures, date of late breast cancer recurrence, anatomical location of recurrence, treatment type and date of first treatment.

We conducted the following sensitivity analyses: I) Considered CBC and recurrence as one outcome in the reference standard as these events are often pooled together in research on disease-free survival; 2) omitted surgical codes from the algorithm; 3) omitted surgical codes from the algorithm and pooled CBC and recurrence as one outcome in the reference standard.

Statistical analyses were performed using Stata 13 (StataCorp LP, College Station, TX, USA)

This study was approved by DBCG, the Danish Data Protection Agency (Aarhus University, J. nr. 2016–051-000001, record number 552) and the Danish Patient Safety Authority (j. nr. 3–3013-2295/1, 3–3013-3302/1, 3–3013-3153/1, 3–3013-3136/1).

Results

Positive Predictive Values

After exclusion of five patients with a missing record or insufficient information in the records and 16 patients who developed a CBC after December 31, 2013, we included 105 potential cases of late breast cancer recurrence (Figure 1). Descriptive characteristics of study population I are presented in Table 1. Of the 105 potential late breast cancer recurrence cases, 90 were confirmed in the medical record review, PPV = 86% (95% CI; 77–91%). Five of the non-confirmed cases were a new primary breast cancer. Six of the non-confirmed cases were only registered due to the surgical codes. When we classified a CBC as a recurrence too, our PPV increased to 90% (95% CI; 83–95%). The PPVs varied by algorithm criteria and were highest for the specific diagnosis and procedure codes for recurrence (100%) and for the pathology codes (95%). The lowest PPV was for the procedural codes for surgery (including mastectomy, BCS, etc.) where the PPV was 70% (Table 2). Among the patients, who were only identified by the algorithm via a procedural code,

Table 1 Descriptive Characteristics on Patients Diagnosed with Stage I–III Operable Breast Cancer Registered in the DBCG Between January 1, 1987 and December 31, 2004 Who Were Alive, Living in Denmark and Without a Recurrence or Second Cancer 10 Years After Diagnosis, According to Study Population

Characteristics	Study Population I[a]	Study Population II[b]
Total Numbers of Patients	105 (100%)	114 (100%)
Age at primary breast cancer diagnosis (years)		
<40	7 (6.7)	<5
40–49	38 (36.2)	22 (19.3)
50–59	40 (38.1)	39 (34.2)
60–69	15 (14.3)	38 (33.3)
≥70	5 (4.7)	<15
Age, median (years)	52	57
Calendar period of primary breast cancer		
1987–1991	23 (21.9)	15 (13.2)
1992–1996	33 (31.4)	24 (21.1)
1997–2000	30 (28.6)	26 (22.8)
2001–2004	19 (18.1)	49 (43.0)
Menopausal status at primary breast cancer diagnosis		
Premenopausal	48 (45.7)	27 (23.7)
Postmenopausal	46 (43.8)	77 (67.5)
Unknown	11 (10.5)	10 (8.8)
Charlson Comorbidity Index Score		
0	61 (58.1)	45 (39.5)
1–2	38 (36.2)	63 (55.3)
≥3	6 (5.7)	6 (5.2)
Stage[c]		
I	45 (42.8)	49 (43.0)
II	54 (51.4)	60 (52.6)
III	<10	5 (4.4)
Unknown	<5	–
Grade		
Low	29 (27.6)	32 (28.1)
Moderate	45 (42.9)	42 (36.8)
High	11 (10.5)	14 (12.3)
Unknown	20 (19.0)	26 (22.8)
Number of positive lymph nodes		
0	58 (55.2)	74 (64.9)
1–3	42 (40.0)	35 (30.7)
≥4	5 (4.8)	5 (4.4)
Tumor size		
<20mm	72 (68.6)	75 (65.8)
>20mm	33 (31.4)	39 (34.2)
ER status		
ER+	60 (57.1)	70 (61.4)
ER-	9 (8.6)	12 (10.5)
Unknown	36 (34.3)	32 (28.1)
Type of primary surgery		
Mastectomy	39 (37.1)	62 (54.4)
Mastectomy + RT	18 (17.2)	20 (17.5)
BCS + RT	48 (45.7)	32 (28.1)
Adjuvant chemotherapy received		
No	73 (69.5)	93 (81.6)
Yes	32 (30.5)	21 (18.4)
Endocrine therapy received		
No	68 (64.8)	62 (54.4)
Yes	37 (35.2)	52 (45.6)

Notes: Cell sizes less than 5 are reported in aggregate to reduce identifiability of individuals in the data. [a]Women diagnosed with a late recurrence according to our algorithm at Aarhus University Hospital. [b]Women with primary breast cancer diagnosed at Aalborg University Hospital. [c]Stage was calculated using the TNM staging system.

Abbreviations: DBCG, Danish Breast Cancer Group; ER, estrogen receptor status; BCS, breast conserving surgery; RT, radiation therapy.

33% had a recurrence documented in the medical records. When omitting these surgical codes from the algorithm the PPV increased to 91%. If we omitted the surgical codes from the algorithm and did not distinguish a CBC from a late breast cancer recurrence, our PPV was 94% (84.5–95.7%). The PPV´s did not vary by patient- and tumor characteristics at primary diagnosis (Supplementary Table 1).

The recurrence date according to the algorithm was concordant with the recurrence date in the medical records (CCC= 0.996). The median difference between the recurrence date estimated by the algorithm and the recurrence date documented in the medical records was 11 days (IQR: 4–20 days).

Sensitivity, Specificity and Negative Predictive Values

Among the 126 patients sampled from Aalborg University Hospital, we excluded 12 patients who moved away, had missing

Table 2 Positive Predictive Values (PPVs) and Corresponding 95% Confidence Intervals (CIs) of Late Breast Cancer Recurrence Identified by Our Algorithm During Follow-Up

	Confirmed Cases/Potential Cases	PPV (%)	95% CI
Algorithm criteria, overall	90/105	85.7	77.5–91.3
DNPR-registered metastases[a]	63/69	91.3	81.6–96.1
Pathology codes for recurrence[b]	63/66	95.5	86.5–98.6
DNPR-registered treatment codes[c]	81/92	88.0	79.5–93.3
Neo-adjuvant/adjuvant treatment	75/80	93.8	85.6–97.4
Surgery (mastectomy, BCS, etc.)	16/23	69.6	46.6–85.7
Specific codes for recurrence in DNPR[d]	14/14	100.0	–
Local breast cancer recurrence	13/13	100.0	–
Recurrence surgery	<5/<5	100.0	–

Notes: Cell sizes less than 5 are reported in aggregate to reduce identifiability of individuals in the data. [a]ICD10 DC76-DC80. [b]T code in the breast with morphology codes M8 or M9 with 4, 6, 7 or 9 in the fifth position (eg, M8XXX4), or any T code (excluding the breast T codes) with morphology codes M8 or M9 with the numbers 6 or 9 in the fifth position. [c]SKS treatment codes: BWG, BWHA, BWHC, BHHH, BOHJ13; Danish version of the Nordic Medico-Statistical Committee Classification of Surgical Procedures: KHAC, KHAB, KGAE16. [d]ICD10 DC509X; Danish version of the Nordic Medico-Statistical Committee Classification of Surgical Procedures: KHAF.
Abbreviations: DNPR, Danish National Patient Registry; BCS, breast conserving surgery; PPV, positive predictive value.

records or insufficient information in the records, or developed a CBC after December 31, 2013. Therefore, 114 were included in study population II (Figure 1). Their descriptive characteristics are presented in Table 1. Of these patients, seven developed a late breast cancer recurrence according to our algorithm and less than 5 were confirmed by the medical record review. One hundred and seven patients did not develop a late recurrence according to our algorithm; these were confirmed by the medical record review. The sensitivity of the late breast cancer recurrence algorithm was 100% (95% CI: 40–100%), the specificity was 97% (95% CI: 92–99%) and the NPV was 100% (97–100%) (Table 3). The positive likelihood ratio was 37 (95% CI: 12.-112) and the negative likelihood ratio was 0.

Information in Medical Records of Confirmed Late Breast Cancer Recurrence Cases

Among the 90 patients with a medical record-confirmed late breast cancer recurrence diagnosis at Aarhus University Hospital, 75 (83%) had a biopsy-verified diagnosis. Seventeen (19%) developed local recurrence, below 15 (<15%) developed regional recurrence, 61 (68%) developed distant recurrence, and below 5 (<5%) had a unknown extent of disease. The most frequent sites of recurrence were the bones (49%), lungs (30%) and lymph nodes (34%) (Supplementary Table 2). Among the

Table 3 Estimation of Sensitivity, Specificity, Negative Predictive Value and Likelihood Ratios Among Women Diagnosed with Primary Breast Cancer at Aalborg University Hospital, 1987–2004

Algorithm	Medical Record Review		
	Late Recurrence	No Late Recurrence	In Total
Late recurrence	<5	<5	7
No late recurrence	0	107	107
In total	<5	<114	114
Results (95% confidence intervals)			
Sensitivity (TP/(TP+FN))	100.0% (39.8–100.0%)		
Specificity (TN/(TN+FP))	97.3% (92.2–99.4%)		
NPV (TN/(TN+FN))	100.0% (96.6–100.0%)		
PLR (sensitivity/1- specificity)	36.7 (12.0–111.9)		
NLR (1-sensitivity/specificity)	0		

Notes: Cell sizes less than 5 are reported in aggregate to reduce identifiability of individuals in the data.
Abbreviations: TP, true positive; FN, false negative, TN; true negative; NPV, negative predictive value, PLR, positive likelihood ratio; NLR, negative likelihood ratio.

90 confirmed late breast cancer recurrence patients, below 15 (<15%) underwent a mastectomy as initial treatment (within the first year after late breast cancer recurrence), none underwent a BCS and below 5 (<5%) had unspecified surgery. Of the confirmed late recurrence patients, 22 (24%) received radiotherapy, 24 (27%) received chemotherapy, 61 (68%) received endocrine therapy, 7 (8%) received trastuzumab and 35 (39%) received bisphosphonates within the first year after diagnosis (Supplementary Table 2).

Discussion

Main Findings

Our algorithm to identify late breast cancer recurrence using Danish nationwide registered data showed moderate to high PPV and high sensitivity, specificity and NPV. The PPVs varied by algorithm criteria and the lowest PPV was calculated for the procedural codes for surgery. The recurrence date according to the algorithm was concordant with that in the medical records.

Comparison with Other Studies

Our PPV was lower than that reported in a previous breast cancer study by Aagaard Rasmussen et al (94%).[13] This may be due to several reasons. Their reference standard did not distinguish recurrent from second primary breast cancer. Therefore, it is possible that some of their proposed recurrences were new primary breast cancers. We used the contralateral database to exclude second primary breast cancers.[23] The contralateral database used the Cancer Registry to identify contralateral breast cancers and Cancer Registry notification forms or records in the Pathology Registry to ascertain the date of CBC diagnosis.[23] Additional CBCs were retrieved from the DBCG. Most second malignancies in breast cancer survivors occur in the contralateral breast.[23] When we classified a CBC as a recurrence, our PPV increased to 90%. Our lowest PPV was found for the procedural codes. Aagaard Rasmussen et al found a high PPV for procedural codes, but they only included patients with both a procedural code and an ICD code for breast cancer (DC50).[13] Our PPV increased slightly when we added DC50 to the procedural codes. The evidence from our medical record review suggests that surgery many years after a primary breast cancer may represent prophylactic mastectomy, which may explain the observed poorer PPV for procedural codes. Furthermore, Aagaard Rasmussen et al´s reference population had a median follow-up of 7.5 years [interquartile range (ICQ) of 5–9 years] since the primary BC surgery and the algorithm was therefore not validated for use in identifying late breast cancer recurrence.[13]

Our previous algorithm[14] showed lower sensitivity, specificity and PPV than in the present study. However, we did not use the CBC database to exclude second breast cancers. Another Danish study by Lash et al[11] used the same registries and a similar algorithm to identify colorectal cancer recurrence. They found a sensitivity of 95%, a specificity of 97%, a PPV of 86% and a NPV of 99%, similar to our results.

Studies from the US and UK have also identified breast cancer recurrence using administrative data.[29–36] A study from the US by Hasset et al[35] found that identifying recurrence based on chemotherapy alone yielded a PPV of only 11% and was a source for many false positives. However, in our study, we found a PPV of 93% in the algorithm criteria where both chemotherapy, endocrine therapy and radiation therapy were included. A patient was considered as having late breast cancer recurrence if they received just one of the treatments and was not registered with a CBC or other new primary cancer up to 90 days after the late breast cancer recurrence diagnosis. Another study from the US by Chawla et al suggested that metastases codes were not suitable for identifying recurrence.[36] Nonetheless, both our study and that by Aagaard Rasmussen et al showed valid results.[13]

Strength and Limitations

A major strength of this study is that the algorithm was based on high quality Danish national registries with complete data and follow-up.[16,20,21] The risk of selection bias was minimal due to the free universal tax-supported healthcare for all residents provided by the Danish Health Service.[16] We were able to access pathology records incorporating SNOMED codes thereby distinguishing recurrent disease (both local and metastatic) from new primary tumors. The risk of misclassification in our reference standard was minimal, as the medical records also included description of the pathology examination. We calculated PPVs by calendar period and found similar results, indicating that the database quality and completeness was high over time. The occurrence of false positives is a problem in cancer algorithms as codes related to the primary tumor are misinterpreted as a recurrence. Given the time period for late recurrence (more than 10 years after primary diagnosis), therapies for the primary tumor should have been completed. The maximum negative predictive value indicated that all breast

cancer patients classified as recurrence free by our algorithm were truly recurrence free. We observed similar PPVs irrespective of ER status, indicating that the algorithm can be used on all breast cancer patients. The high sensitivity leads to few false negative results and therefore there is low risk of missing patients with late recurrence. Furthermore, our algorithm can be externally validated in other geographical areas.

Several factors should be considered when interpreting our study. We restricted the study to patients diagnosed at two major hospitals. Another concern is that elderly or comorbid patients may be missed by the algorithm. These patients may be less likely to receive pathological diagnostic work-up (biopsy) and treatment for a late recurrence. Furthermore, our algorithm did not include SNOMED codes with "3" in fifth position (eg, M8XXX3), which indicates a malignant tumor. These codes are sometimes used in combination with a text annotation indicating a possible recurrence. In the early phase of our study, we included codes with "3" in fifth position, but we found more recurrences than we expected from previous literature, and therefore decided not to include them. Also, we expect that any local recurrences would be identified by one of the other criteria in our algorithm. This is supported by the observed high sensitivity. Late recurrence is a rare event – we observed few patients who developed a late breast cancer recurrence in our random sample cohort from Aalborg University Hospital (Study population II). As a result, we are unable to stratify the sensitivity estimates by algorithm criteria, patient- and tumor characteristics. A final limitation is that the PPV for late recurrence suggested that 14% did not have a late breast cancer recurrence. Accordingly, the algorithm may overestimate the extent of late recurrence in incidence studies. Furthermore, misclassification of outcomes may bias ratio measures and should be considered in future studies.

Conclusion

We constructed an algorithm to identify late breast cancer recurrence using routinely collected data in Danish administrative registries. The algorithm showed high validity and could be an important tool in future studies of late breast cancer recurrence.

Data Accessibility

The data are not publicly available due to privacy and ethical restrictions.

Abbreviations

BCS, breast conserving surgery; CBC, contralateral breast cancer; CCC, concordance correlation coefficient; CCI, Charlson Comorbidity Index; CPR, Civil Personal Registration; CRS, Danish Civil Registration System; CT scan, computerized tomography scan; DBCG, Danish Breast Cancer Group; DCR, Danish Cancer Registry; DNPR, Danish National Patient Registry; ER, estrogen receptor; FN, false negative; ICD, International Classification of Diseases; MRI, Magnetic Resonance Imaging; NPV, negative predictive value; PPV, positive predictive value; RT, radiation therapy; SNOMED, Systematized Nomenclature of Medicine; TNM, Tumor Node Metastasis; TN, true negative; TP, true positive.

Acknowledgements

This work was supported by grants from the Danish Cancer Society (R147-A10100-16-S45) and Aarhus University. TT is funded by the Danish Cancer Society. The funding agencies had no role in design of the study; the collection, analysis and interpretation of the data; the writing of the article; or the decision to submit the article for publication.

The Department of Clinical Epidemiology is involved in studies that receive funding from various companies as research grants to (and administered by) Aarhus University. None of these studies have any relation to the present work.

Disclosure

Trine Tramm reports personal fees from Roche, personal fees from Pfizer, outside the submitted work. The authors declare no other conflicts of interest for this work.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424. doi:10.3322/caac.21492
2. Hovaldt HB, Suppli NP, Olsen MH, et al. Who are the cancer survivors? A nationwide study in Denmark, 1943-2010. *Br J Cancer*. 2015;112(9):1549–1553. doi:10.1038/bjc.2015.68
3. Peto R, Boreham J, Clarke M, Davies C, Beral V. UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years. *Lancet (London, England)*. 2000;355(9217):1822. doi:10.1016/S0140-6736(00)02277-7
4. Cianfrocca M. Overcoming recurrence risk: extended adjuvant endocrine therapy. *Clin Breast Cancer*. 2008;8(6):493–500. doi:10.3816/CBC.2008.n.059
5. Park J-H, Anderson WF, Gail MH. Improvements in US breast cancer survival and proportion explained by tumor size and estrogen-receptor status. *J Clin Oncol*. 2015;33(26):2870–2876. doi:10.1200/JCO.2014.59.9191

6. Tryggvadóttir L, Gislum M, Bray F, et al. Trends in the survival of patients diagnosed with breast cancer in the Nordic countries 1964-2003 followed up to the end of 2006. *Acta Oncol*. 2010;49(5):624–631. doi:10.3109/02841860903575323
7. Richman J, Dowsett M. Beyond 5 years: enduring risk of recurrence in oestrogen receptor-positive breast cancer. *Nat Rev Clin Oncol*. 2019;16(5):296–311. doi:10.1038/s41571-018-0145-5
8. Pan H, Gray R, Braybrooke J, et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *N Engl J Med*. 2017;377(19):1836–1846. doi:10.1056/NEJMoa1701830
9. Kostev K, Kalder M. 20-year risk of breast cancer recurrence. *Breast Cancer Res Treat*. 2018;168(3):765–766. doi:10.1007/s10549-017-4636-3
10. Møller S, Jensen M-B, Ejlertsen B, et al. The clinical database and the treatment guidelines of the Danish Breast Cancer Cooperative Group (DBCG); its 30-years experience and future promise. *Acta Oncol*. 2008;47(4):506–524. doi:10.1080/02841860802059259
11. Lash TL, Riis AH, Ostenfeld EB, Erichsen R, Vyberg M, Thorlacius-Ussing O. A validated algorithm to ascertain colorectal cancer recurrence using registry resources in Denmark. *Int J Cancer*. 2015;136(9):2210–2215. doi:10.1002/ijc.29267
12. Rasmussen LA, Jensen H, Virgilsen LF, Jensen JB, Vedsted P. A validated algorithm to identify recurrence of bladder cancer: a register-based study in Denmark. *Clin Epidemiol*. 2018;10:1755–1763. doi:10.2147/CLEP.S177305
13. Aagaard Rasmussen L, Jensen H, Flytkjær Virgilsen L, Jellesmark Thorsen LB, Vrou Offersen B, Vedsted P. A validated algorithm for register-based identification of patients with recurrence of breast cancer-based on Danish Breast Cancer Group (DBCG) data. *Cancer Epidemiol*. 2019;59:129–134. doi:10.1016/j.canep.2019.01.016
14. Cronin-Fenton DP, Kjærsgaard A, Nørgaard M, et al. Clinical outcomes of female breast cancer according to BRCA mutation status. *Cancer Epidemiol*. 2017;49:128–137. doi:10.1016/j.canep.2017.05.016
15. Henriksen DP, Rasmussen L, Hansen MR, Hallas J, Pottegård A. Comparison of the five danish regions regarding demographic characteristics, healthcare utilization, and medication use – a descriptive cross-sectional study. *PLoS One*. 2015;10(10):e0140197. doi:10.1371/journal.pone.0140197
16. Schmidt M, Pedersen L, Sørensen HT. The Danish civil registration system as a tool in epidemiology. *Eur J Epidemiol*. 2014;29(8):541–549. doi:10.1007/s10654-014-9930-3
17. Jensen AR, Storm HH, Møller S, Overgaard J. Validity and representativity in the Danish Breast Cancer Cooperative Group – a study on protocol allocation and data validity from one county to a multicentre database. *Acta Oncol*. 2003;42(3):179–185. doi:10.1080/02841860310000737
18. Kaufmann N. Pakkeforløb for brystkræft. 2018.
19. Lynge E, Sandegaard JL, Rebolj M. The Danish national patient register. *Scand J Public Health*. 2011;39(7_suppl):30–33. doi:10.1177/1403494811401482
20. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish national patient registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449–490. doi:10.2147/CLEP.S91125
21. Erichsen R, Lash TL, Hamilton-Dutoit SJ, Bjerregaard B, Vyberg M, Pedersen L. Existing data sources for clinical epidemiology: the Danish national pathology registry and data bank. *Clin Epidemiol*. 2010;2:51–56. doi:10.2147/CLEP.S9908
22. Storm HH, Michelsen EV, Clemmensen IH, Pihl J. The Danish cancer registry – history, content, quality and use. *Dan Med Bull*. 1997;44(5):535–539.
23. Rasmussen CB, Kjær SK, Ejlertsen B, et al. Incidence of metachronous contralateral breast cancer in Denmark 1978–2009. *Int J Epidemiol*. 2014;43(6):1855–1864. doi:10.1093/ije/dyu202
24. Han HH, Lee SH, Kim BG, Lee JH, Kang S, Cho NH. Estrogen receptor status predicts late-onset skeletal recurrence in breast cancer patients. *Medicine (Baltimore)*. 2016;95(8):e2909. doi:10.1097/MD.0000000000002909
25. Saphner T, Tormey DC, Gray R. Annual hazard rates of recurrence for breast cancer after primary therapy. *J Clin Oncol*. 1996;14(10):2738–2746. doi:10.1200/JCO.1996.14.10.2738
26. Howell A, Cuzick J, Baum M, et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or iCombination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet (London, England)*. 2005;365(9453):60–62. doi:10.1016/S0140-6736(04)17666-6
27. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373–383. doi:10.1016/0021-9681(87)90171-8
28. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45(1):255–268. doi:10.2307/2532051
29. Ritzwoller DP, Hassett MJ, Uno H, et al. Development, validation, and dissemination of a breast cancer recurrence detection and timing informatics algorithm. *J Natl Cancer Inst*. 2018;110(3):273–281. doi:10.1093/jnci/djx200
30. Chubak J, Onega T, Zhu W, Buist DSM, Hubbard RA. An electronic health record-based algorithm to ascertain the date of second breast cancer events. *Med Care*. 2017;55(12):e81–e87. doi:10.1097/MLR.0000000000000352
31. Chubak J, Yu O, Pocobelli G, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *J Natl Cancer Inst*. 2012;104(12):931–940. doi:10.1093/jnci/djs233
32. Kroenke CH, Chubak J, Johnson L, Castillo A, Weltzien E, Caan BJ. Enhancing breast cancer recurrence algorithms through selective use of medical record data. *J Natl Cancer Inst*. 2016;108(3). doi:10.1093/jnci/djv336
33. Lamont EB, Herndon IIJE, Weeks JC, et al. Measuring disease-free survival and cancer relapse using medicare claims from CALGB breast cancer trial participants (companion to 9344). *J Natl Cancer Inst*. 2006;98(18):1335–1338. doi:10.1093/jnci/djj363
34. Warren JL, Mariotto A, Melbert D, et al. Sensitivity of medicare claims to identify cancer recurrence in elderly colorectal and breast cancer patients. *Med Care*. 2016;54(8):e47–e54. doi:10.1097/MLR.0000000000000058
35. Hassett MJ, Ritzwoller DP, Taback N, et al. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using two large contemporary cohorts. *Med Care*. 2014;52(10):e65–e73. doi:10.1097/MLR.0b013e318277eb6f
36. Chawla N, Yabroff KR, Mariotto A, McNeel TS, Schrag D, Warren JL. Limited validity of diagnosis codes in medicare claims for identifying cancer metastases and inferring stage. *Ann Epidemiol*. 2014;24(9):666–672, 672.e1-2. doi:10.1016/j.annepidem.2014.06.099