Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Machine Learning Algorithm to Estimate Distant Breast Cancer Recurrence at the Population Level with Administrative Data

Hava Izci[1], Gilles Macq[2], Tim Tambuyzer[2], Harlinde De Schutter[2], Hans Wildiers[1,3], Francois P Duhoux[4], Evandro de Azambuja[5], Donatienne Taylor[6], Gracienne Staelens[7], Guy Orye[8], Zuzana Hlavata[9], Helga Hellemans[10], Carine De Rop[11], Patrick Neven[1,3], Freija Verdoodt[2]

[1]KU Leuven - University of Leuven, Department of Oncology, Leuven, B-3000, Belgium; [2]Belgian Cancer Registry, Research Department, Brussels, Belgium; [3]University Hospitals Leuven, Multidisciplinary Breast Center, Leuven, B-3000, Belgium; [4]Department of Medical Oncology, King Albert II Cancer Institute, Cliniques Universitaires Saint-Luc, Brussels, Belgium; [5]Institut Jules Bordet and l'Université Libre de Bruxelles (U.L.B), Brussels, Belgium; [6]CHU UCL Namur, Site Sainte-Elisabeth, Namur, Belgium; [7]Multidisciplinary Breast Center, General Hospital Groeninge, Kortrijk, Belgium; [8]Department of Obstetrics and Gynecology, Jessa Hospital, Hasselt, Belgium; [9]Department of Medical Oncology, CHR Mons-Hainaut, Mons, Hainaut, Belgium; [10]Department of Obstetrics and Gynaecology, AZ Delta, Roeselaere, Belgium; [11]Department of Obstetrics and Gynaecology, Imelda Hospital, Bonheiden, Belgium

Correspondence: Hava Izci, KU Leuven, Department of oncology, Herestraat 49 Box 7003-06, Leuven, 3000, Belgium, Email hava.izci@kuleuven.be

Purpose: High-quality population-based cancer recurrence data are scarcely available, mainly due to complexity and cost of registration. For the first time in Belgium, we developed a tool to estimate distant recurrence after a breast cancer diagnosis at the population level, based on real-world cancer registration and administrative data.

Methods: Data on distant cancer recurrence (including progression) from patients diagnosed with breast cancer between 2009–2014 were collected from medical files at 9 Belgian centers to train, test and externally validate an algorithm (i.e., gold standard). Distant recurrence was defined as the occurrence of distant metastases between 120 days and within 10 years after the primary diagnosis, with follow-up until December 31, 2018. Data from the gold standard were linked to population-based data from the Belgian Cancer Registry (BCR) and administrative data sources. Potential features to detect recurrences in administrative data were defined based on expert opinion from breast oncologists, and subsequently selected using bootstrap aggregation. Based on the selected features, classification and regression tree (CART) analysis was performed to construct an algorithm for classifying patients as having a distant recurrence or not.

Results: A total of 2507 patients were included of whom 216 had a distant recurrence in the clinical data set. The performance of the algorithm showed sensitivity of 79.5% (95% CI 68.8–87.8%), positive predictive value (PPV) of 79.5% (95% CI 68.8–87.8%), and accuracy of 96.7% (95% CI 95.4–97.7%). The external validation resulted in a sensitivity of 84.1% (95% CI 74.4–91.3%), PPV of 84.1% (95% CI 74.4–91.3%), and an accuracy of 96.8% (95% CI 95.4–97.9%).

Conclusion: Our algorithm detected distant breast cancer recurrences with an overall good accuracy of 96.8% for patients with breast cancer, as observed in the first multi-centric external validation exercise.

Keywords: machine learning, breast cancer, distant metastases, recurrences, algorithm, administrative data

Introduction

Cancer recurrence is considered to be an important cancer outcome metric to measure the burden of the disease and success of (neo)adjuvant therapies. Despite this, high-quality breast cancer recurrence rates currently remain unknown in most countries, including Belgium. To date, cancer recurrence is not systematically registered in most population-based cancer registries, due to the difficulty and labor-intensity of registering follow-up for recurrences.

Recurrence definitions used for registration purposes differ among countries, due to the lack of consensus regarding a standardized clinical definition. Defining recurrence clinically is a challenge, since various methods exist to detect

Received: 3 December 2022
Accepted: 1 April 2023
Published: 5 May 2023

recurrences after (neo)adjuvant treatments of a patient such as physical examination, pathological examination, imaging, or tumor markers. Unlike the guidelines and definitions that currently exist in the clinical trial setting,[1,2] no guidelines are set to correctly and consistently register a recurrence in a patient with stage I–III breast cancer at diagnosis.

Real-world recurrence data could give an estimation of cancer burden and efficacy of cancer treatment modalities outside a conventional clinical trial setting, which could eventually lead to improvements in quality of care.[3,4] Administrative data from health insurance companies on medical treatments and procedures, also known as bill claims, and hospital discharge data could represent an alternative source for the assessment of disease evolution after breast cancer treatment.

Recently, machine learning algorithms based on classification and regression trees (CART) have been developed to detect cancer recurrence at the population level using claims data.[5] However, only in a limited number of countries, research teams were able to successfully construct algorithms to detect breast cancer recurrences, and only for a small number of centers (USA,[6,7] Canada,[8,9] Denmark[10,11] and Sweden)[12] Our aim was to develop, test and validate an algorithm using administrative data features allowing the estimation of breast cancer recurrence rates for all Belgian patients with breast cancer.

Methods

Study Population

To construct and validate an algorithm to detect distant recurrences, female patients with breast cancer diagnosed between January 1, 2009 and December 31, 2014 were included from nine different centers located in all three Belgian regions. We did not include patients with stage IV breast cancer at diagnosis, patients with a history of cancer (any second primary cancer, multiple tumors, and contralateral tumors), or patients who could not be coupled to administrative data sources. All breast cancers, regardless of molecular subtype, were included. Among the nine centers were centers from the Flemish region (University Hospitals Leuven, General Hospital Groeninge, Jessa Hospital, Imelda Hospital, and AZ Delta), Brussels-Capital region (Cliniques universitaires Saint-Luc and Institut Jules Bordet) and Walloon region (CHR Mons-Hainaut and CHU UCL Namur). For all nine centers, 300 patients were included per center, by randomly selecting from the study population 50 patients per incidence year. The study population of six centers was divided by randomization (60–40% split-sample validation) into a training set to develop the algorithm, and an independent test set to perform an internal validation.[13] The algorithm was additionally validated with an external validation set of the three remaining centers, to check reproducibility of the algorithm in a dataset with patients from other centers.

Definition of Distant Recurrence: Manual Chart Review

For the selection of the nine centers, we aimed for a reasonable variety of center characteristics based on teaching vs non-teaching hospital, the spread across the three regions in Belgium, and center size.

For each patient in the study population, recurrence status (yes, no, unknown) and recurrence date (day, month, year) were extracted and collected from electronic medical files and reviewed by trained data managers from each of the nine hospitals. Recurrence was defined as the occurrence of a distant recurrence or metastasis between 120 days after the primary diagnosis and within 10 years of follow-up after diagnosis or end of study (December 31, 2018). Data managers were instructed to consider death due to breast cancer in our definition of a recurrence. Loco-regional recurrence, was not considered as an outcome in our study. Both patients with a progression (without a disease-free interval) and patients with a recurrence (with a disease-free interval) were considered as outcome in our definition of recurrence. Patients with an unknown recurrence status, due to the lack of follow-up for example, were excluded from the analysis. Patients with a recurrence within 120 days were considered de novo stage IV and therefore excluded because interference of first-line treatment complicates recurrence detection. Starting from diagnosis to detect recurrent disease might cause more false positive recurrence cases due to the treatment of the initial breast cancer overlapping with the immediate first-line treatment due to metastatic disease. Recurrence diagnosis date was the time-point (described in day, month, and year), confirmed by pathological examination, imaging (CT, PET-CT, bone scintigraphy or MRI scan), or defined by physicians in the multidisciplinary team meeting (MDT).

Administrative Data Sources and Linkage

In the course of an extensive data linking process with pseudonymization of the patient data, the recurrence data from the hospitals (i.e., gold standard) were linked to several population-based data sources. These included cancer registration data from the Belgian Cancer Registry (BCR), and administrative data sources, including claims or reimbursement data (InterMutualistic Agency, IMA),[14] hospital discharge data (Technische Cel, TCT),[15] information on vital status (Crossroads Bank for Social Security, CBSS)[16] and cause of death ("Agentschap Zorg en Gezondheid", "Observatoire de la Santé et du Social de Bruxelles-Capitale", and "Agence pour une Vie de Qualité" – AVIQ).[17] Information on data sources and data used is presented in Appendix 1.

Pre-Processing and Feature Extraction

To build a robust algorithm to detect distant recurrences, pre-processing and extraction of features were performed. Expert-driven features to potentially detect recurrences in administrative data were created based on recommendations from breast oncologists (P.N. and H.W.). First, a comprehensive list of reimbursement codes for diagnostic and therapeutic procedures and medications was selected, and code groups were created based on their relevance for the diagnosis and/or treatment of distant metastasis in breast cancer patients (See Appendix 2).

Potential features were further refined based on the exploration of data from patients with a recurrence, including time-frames starting from time points after diagnosis (0 days, 90 days, 160 days, 270 days, and 365 days after diagnosis). We assessed different time-frames to obtain the most accurate feature to detect recurrences, and because starting from the date of diagnosis might result in noise from the treatment of the initial breast cancer. We additionally created features based on count of codes, by assessing the maximum number of codes per year or per pre-defined time-frame (starting from 0, 90, 160, 270, and 365 days after diagnosis) (Table 1). The best performing time-frame was selected for each feature by maximizing the Youden's J index:[18]

$$J = Sensitivity + Specificity - 1$$

Feature Selection and Model Development

After a feature list was obtained (as described in previous section), this list was narrowed down based on the ensemble method of bootstrapping.[19] In total 1000 bootstrap samples were used to generate 1000 classification and regression trees (CART) using the same training set, and to select best-performing features based on the frequency of the features.[19,20]

Cost-complexity pruning was applied for each bootstrap sample, to obtain the best performing model and avoid over-fitting of the model to the dataset.[20] CART inherently uses entropy for the selection of nodes or features. The higher the entropy, the more informative and useful the feature is.[20] A 10-fold cross-validation was also performed to ensure robustness of the model in different training sets. Collinearity of the selected features was accounted for by the one standard error (1-SE) rule, to eliminate redundant features. The 1-SE rule selects the least complex tree that is within 1 standard error from the best performing tree.[21]

Based on the selected features from the bootstrapping, a principal CART model was built to classify patients as having a recurrence or not by using the complete training set.

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and classification accuracy was calculated for evaluating and comparing the performance of the principal CART model. All models were created and trained in SAS 9.4 (SAS Institute, Cary, NC, USA) within the SAS Enterprise Guide software (version 7.15 of the SAS System for Windows).

Results

Data for a total of 2507 patients could be retrieved from nine Belgian centers and were included in the final dataset to train, test and externally validate the algorithm (Figure 1 and Table 2). The mean follow-up period was 7.4 years. For the split sample validation, the patients from six centers were split into the training set (N = 975 of which 78 distant recurrences, 8.0%) and internal validation set (N = 713 of which 56 distant recurrences, 7.9%).

Table 1 List of Potential Markers for Recurrence (Available Within Administrative Data) Based on Recommendations from Breast Oncologists

Category	Feature	Rule	Considered Time-Frame
Diagnosis	First Multidisciplinary team meeting (MDT)	Presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	Follow-up Multidisciplinary team meeting (MDT)	Presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	Any Multidisciplinary team meeting (MDT)	Presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	Anatomical pathological report	Presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	Imaging (CT, MRI, X-ray, ultrasound scan)	Count per year or presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	CT scan	Presence of 2, 3, 4, 5 or 6 times a year	
		Max number of codes present X times a year	
	MRI scan	Presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	CT or MRI scan	Presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	Test for Tumor marker CA15-3	Count of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
		Max number of codes present X times a year	
	Secondary malignant neoplasm	Presence	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
Treatment	Chemotherapy	Count or presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
		Time-frame of X days between codes	
		Metastasis-specific codes	
	Metastasis-specific agents	Presence of codes (grouped)	
		Presence of codes (single)	
	Targeted therapy	Count or presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
		Time-frame of X days between codes	
	Hormone therapy	Count or presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
		Time-frame of X days between codes	
	Radiotherapy	Count or presence of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
		Time-frame of X days between codes	
	Surgery	Count of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date
	Systemic treatment (Chemotherapy, Hormone therapy, or Targeted therapy)	Count of codes	Starting from 0, 90 days, 160 days, 270 days, or 365 days after diagnosis date

(*Continued*)

Table 1 (Continued).

Category	Feature	Rule	Considered Time-Frame
Patient	Death due to breast cancer	Presence	
	Death	Presence	From 5 years after diagnosis date
	Morphology		
	Age		
	Clinical, pathological or combined stage		

Abbreviations: CT, computer tomography; MRI, magnetic resonance imaging; CA15-3, cancer antigen 15–3.

The external validation set consisted of three independent centers with 819 patients, of which 82 had distant recurrences (10.0%). The training, internal validation, and external validation sets did not have differences in distribution of baseline tumor and patient characteristics (Table 2).

Based on bootstrap aggregation, 1000 CART models were built using the following features: (1) "Presence of a follow-up MDT meeting, starting from 270 days after diagnosis" (feature present in 975 out of 1000 CART models), (2) "Maximum number of CT codes present (with a moving average over time) of 5 or more times a year" (851 CART models), and (3) "Death due to breast cancer" (412 CART models) (see Supplementary Figure 1). Afterwards, the final CART model was constructed with these three features and calculated by using all data of the training set (Figure 2).

The sensitivity of the principal CART model to detect recurrences for the training set was 79.5% (95% confidence interval [CI] 68.8–87.8%), specificity was 98.2% (95% CI 97.1–99.0%), with an overall accuracy of 96.7% (95% CI 95.4–97.7%) (Table 3), and an AUC (area under the curve) of 94.2%. After 10-fold cross-validation within the training set, we found a sensitivity of 71.8% (95% CI 66.4–86.7%), specificity of 98.2% (95% CI 96.3–98.5%) and overall accuracy of 96.1% (95% CI 94.7–97.2%). The internal validation (i.e. based on test set) resulted in a sensitivity of 83.9% (95% CI 71.7–92.4%), a specificity of 96.7% (95% CI 95.0–98.9%),

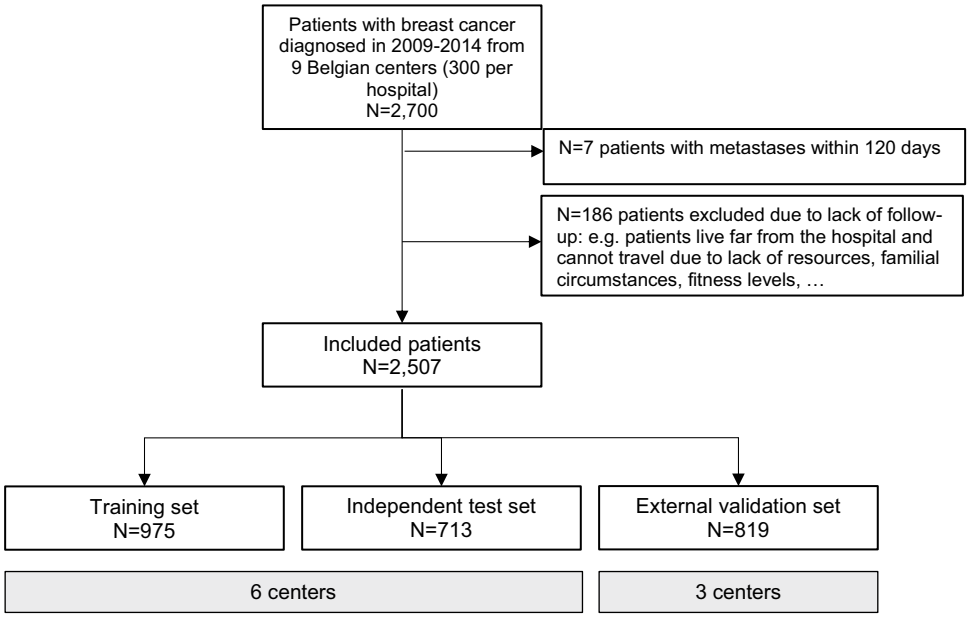

Figure 1 Patient inclusion flow diagram.

Table 2 Baseline Patient and Tumor Characteristics

			6 Belgian Centers		3 Belgian Centers
Characteristic	**Category**	**Statistic**	**Training Set (N=975)**	**Independent Test Set (N=713)**	**External Validation Set (N=819)**
Follow-up (months)		**Mean** **Minimum** **Maximum**	**89.9** **1.5** **131.8**	**87.9** **0.9** **131.9**	**88.7** **1.8** **131.9**
Distant recurrence **Age (years)**	Yes	n (%) Mean Minimum Maximum	78 (8.0%) 59.8 21 93	56 (7.9%) 58.8 25 95	82 (10.0%) 59.9 24 91
Combined stage at diagnosis	I II III Unknown	n (%)	452 (46.4%) 363 (37.2%) 116 (11.9%) 44 (4.5%)	340 (47.7%) 259 (36.3%) 85 (11.9%) 29 (4.1%)	371 (45.3%) 289 (35.3%) 107 (13.1%) 52 (6.3%)
Grade	1 2 3 Unknown	n (%)	124 (12.7%) 461 (47.3%) 340 (34.9%) 50 (5.1%)	94 (13.2%) 348 (48.8%) 242 (33.9%) 29 (4.1%)	135 (16.5%) 353 (43.1%) 244 (29.8%) 87 (10.6%)
Cause of death	Alive Deceased Breast cancer Other	n (%)	846 (86.8%) 43 (4.4%) 86 (8.8%)	608 (85.3%) 44 (6.2%) 61 (8.6%)	703 (85.8%) 49 (6.0%) 67 (8.2%)

and accuracy of 95.7% (95% CI 93.9–97.0%). After external validation was performed on three additional centers, the sensitivity was 84.1% (95% CI 74.4–91.3%), with a specificity of 98.2% (95% CI 97.0–99.1%) and accuracy of 96.8% (95% CI 95.4–97.9%).

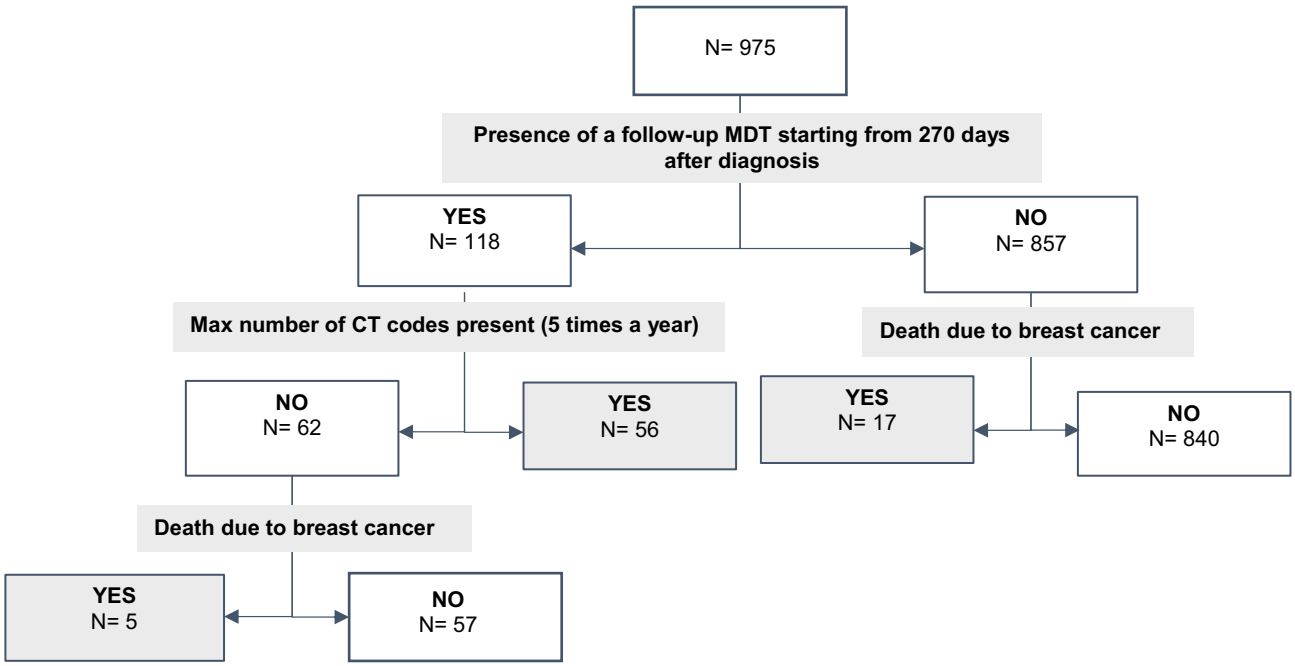

Figure 2 Final CART model to detect recurrences based on the three selected features after bootstrapping. Nodes represent selected features by the algorithm to classify patients.
Abbreviations: MDT, multidisciplinary team meeting; CT, computed tomography scan.

Table 3 Performance of Training Set, Cross Validation, Internal Validation Set and External Validation Set

	TP	TN	FN	FP	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Accuracy, % (95% CI)
Training set (N=975)	62	881	16	16	79.5% (68.8–87.8)	98.2% (97.1–99.0)	79.5% (68.8–87.8)	98.2% (97.1–99.0)	96.7% (95.4–97.7)
Cross validation Training set (N=975)	56	881	22	16	71.8% (66.4–86.7)	98.2% (96.3–98.5)	77.8% (60.5–81.4)	97.6% (97.1–99.0)	96.1% (94.7–97.2)
Internal validation set (N=713)	47	635	9	22	83.9% (71.7–92.4)	96.7% (95.0–98.9)	68.1% (55.8–78.8)	98.6% (97.4–99.4)	95.7% (93.9–97.0)
External validation set (N=819)	69	724	13	13	84.1% (74.4–91.3)	98.2% (97.0–99.1)	84.1% (74.4–91.3)	98.2% (97.0–99.1)	96.8% (95.4–97.9)

Abbreviations: TP, true positives; TN, true negatives; FN, false negatives; FP, false positives; PPV, positive predictive value; NPV, negative predictive value.

Discussion

Main Findings

In this study, we were able to successfully develop a machine learning algorithm to detect distant recurrence in patients with breast cancer, achieving accuracy of 96.8% after external validation in multiple centers across Belgium. The final list of detected parameters were presence of a follow-up MDT meeting, CT scan (max 5 times a year), and death due to breast cancer. Recurrence data are lacking in many population-based cancer registries due to the cost and labor-intensity of registration.[3] True incidence of cancer recurrence should be known across age groups and regions in Belgium, to measure burden of illness and eventually improve quality of care. Current recurrence numbers are often extrapolated from clinical trials, which typically exclude older and frail patients. Older patients are more susceptible to receive under-treatment and to recurrences[22,23] and recurrence numbers could therefore be underestimated.

The administrative data sources used in our algorithm virtually cover all residents of Belgium,[14] which was useful to achieve population-based recurrence data. We were also able to accomplish a multi-centric study by developing the training model and performing an external validation based on data of multiple centers. Likewise, it is highly important to have a relatively large population and reliable gold standard to develop and train a machine learning model in these studies, to avoid prolonging and complicating the feature selection process due to conflicting recurrence and treatment data occurrence.

The definition of a distant recurrence in medical files was the occurrence of a distant recurrence or metastases after a period of 120 days. This time-frame until detection of recurrence varied among previous studies.[24–27] Most common exclusions were done either from 120 days (Chubak et al 2012) or 180 days after diagnosis (A'mar et al 2020). Disease progression can be difficult to measure accurately and can be overestimated because of timing of therapeutic procedures that might be delayed. The limitation of our study was that we could not make a distinction between disease progression and disease recurrence. Defining medical recurrence in the clinic is a challenge, which makes it more difficult to define recurrence with a proxy based on administrative data.[28] Therefore, setting a clear definition of window of treatment and the time-frame for detection of recurrence is considered important for future studies.

We chose to restrict our definition to distant recurrences to achieve a straightforward feature selection. We included death due to breast cancer as an outcome in our definition of recurrences. Cause-specific death and accurate source of cause of death is of utmost importance when studying recurrences, since recurrence and death are closely related to each other.[29]

The machine learning algorithm used in this study was a decision tree, i.e. the Classification And Regression Tree (CART) with the ensemble method. Ensemble learning combines multiple decision trees sequentially (boosting) or in parallel (bootstrap aggregation). The key advantages of using bootstrap aggregation are: better predictive accuracy, less variance, and less bias than a single decision tree. Similarly, latest studies more often make use of ensemble methods.[7,9,12]

Within the recurrence detection features that were selected from the bootstrapping method for the cohort of six different Belgian centers, no treatment features were selected, which could indicate that there are more inter-center similarities for diagnostic regimens and more differences in terms of treatment regimens. During pre-processing of the features, we did additional checks of features to improve accuracy of the model. For instance, we generated a treatment

feature that only included metastases-specific chemotherapy agent codes. However, this feature was not included in the final model. Next, we tried out a model without diagnostic features, but this did not improve accuracy. Previous studies mostly make use of metastatic diagnosis codes (secondary malignant neoplasm or SMN code from ICD-9 or ICD-10) in their algorithm, which would be useful if highly reliable. We also checked subgroups by testing out different models for patients younger or older than 70 years, and different incidence years. We applied the algorithm on subgroups based on age or incidence years, to check if the algorithm accuracy performed better in specific subgroups. As expected, we found higher performance in younger patients (Supplementary Table 1).

Our algorithm performance was comparable to previous studies using decision trees.[9,12,24,30–32] We found greater accuracy compared with the pooled accuracy of previous algorithms.[5]

Although algorithms with highest overall accuracy are often sought-after in earlier studies, some studies also provide multiple algorithms to choose from based on their preference, e.g. high-sensitivity or high-specificity algorithms.[6,10,24,26,30] Finally, we also investigated the false negative cases from University Hospitals in Leuven to explain why these cases were misclassified. We found that in most false negative cases, the patients were missed due to the lack of attestation of the claims or management of the patients' procedures. These cases were most likely patients for which there was a decision to withhold treatment because of comorbid disease, older age, the prognosis of the recurrence, or patients' treatments were reimbursed by the sponsor of a clinical trial.

Previously, algorithms based on administrative claims data to detect breast cancer recurrences at the population level have been established.[5,7–10,12] For example research groups from the USA, Canada, and Sweden have built algorithms to detect recurrences in a delimited region within a population. Recent results from these groups have proven that machine learning algorithms based on administrative data can be used to detect recurrences, in the absence of systematic registration. These studies, however, only encompassed a few centers and were thus not validated in a larger cohort of a population. Moreover, most of these algorithms included complete metastasis-specific International Classification of Diseases (ICD)-codes to detect recurrences. Since metastasis-specific codes are not complete in our database, we were not able to use this code in our algorithm. Particularly, the Danish registry has actively collected recurrence information in the Danish Breast Cancer Group (DBCG) clinical database, which they were able to use to construct and validate population-based recurrence-algorithms to complete their recurrence database.[10,11] Additionally, they were able to look into long-term recurrences beyond 10 years after incidence date.[4,33]

The objective of this study was to develop an algorithm that could be used on a nation-wide level to estimate population-wide distant recurrences. Compared with other studies, we used a large sample size and reported both internal and external validation, which was hardly reported in earlier studies.[5] Another strength of our study was that unlike many other studies from the USA using Medicare claims,[34–38] we were able to include all eligible patients with a breast cancer diagnosis, and not just patients older than 65 years.

Although we used different diagnosis and treatment code sources, it should be noted that treatment regimens often change over time and adaptation of the features should be performed for later use. Adapting the algorithm based on changes in diagnosis or treatment regimens might be necessary to obtain accurate recurrence rates of more incidence years in the future. Ideally, we would also prefer to have long-term follow-up and claims data for patients to detect long-term recurrences. However, due to regulations and the large bulk of data that is generated, a longer follow-up of the codes was not possible within the current study. Longer follow-up of recurrences and administrative data would likely improve the accuracy and lead to a more robust algorithm.

In conclusion, our machine learning algorithm to detect metastatic breast cancer recurrences performed with high accuracy after external validation. Claims data are available for medical procedures and medications, hospital discharge data, vital status and cause of death data on the whole population level, which allows the development of models for Belgium. This substantiates the feasibility to develop and validate recurrence algorithms at the population level and might encourage other population-based registries to develop recurrence models or actively register recurrences in the future as these become progressively important. These rates are valuable to gain more insights about recurrences outside the clinical trial setting and might unveil the importance of active registration of recurrences.

Abbreviations

AUC, Area under the curve; ATC, Anatomical Therapeutic Chemical classification; AVIQ, "Agence pour une Vie de Qualité"; BCR, Belgian Cancer Registry; CA15-3, Cancer antigen 15-3; CART, Classification and regression tree; CBSS,

Crossroads Bank for Social Security; CT, Computed tomography; FN, False negatives; FP, False positives; ICD, International Classification of Diseases and Related Health Problems; IMA, InterMutualistic Agency; MDT, Multidisciplinary team meeting; MRI, Magnetic Resonance Imaging; MZG, "Minimale Ziekenhuis Gegevens"; NPV, Negative predictive value; PPV, Positive predictive value; PET-CT, Positron emission tomography – computed tomography; SE, Standard error; SMN, Secondary malignant neoplasm; TN, True negatives; TP, True positives.

Data Sharing Statement

The data that support the findings of this study are available upon reasonable request. The data can be given within the secured environment of the Belgian Cancer Registry, according to its regulations, and only upon approval by the Information Security Committee.

Ethics Approval and Consent to Participate

This retrospective chart review study involving human participants was in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This study was approved by the Ethics Committee of University Hospitals Leuven (S60928). Informed consent for use of data of all participants was obtained.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by VZW THINK-PINK (Belgium).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Gourgou-Bourgade S, Cameron D, Poortmans P, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). *Ann Oncol*. 2015;26(5):873–879. doi:10.1093/annonc/mdv106
2. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–247. doi:10.1016/j.ejca.2008.10.026
3. Warren JL, Yabroff KR. Challenges and opportunities in measuring cancer recurrence in the United States. *J Natl Cancer Inst*. 2015;107:djv134–djv134. doi:10.1093/jnci/djv134
4. Negoita S, Ramirez-Pena E. Prevention of late recurrence: an increasingly important target for breast cancer research and control. *J Natl Cancer Inst*. 2021. doi:10.1093/JNCI/DJAB203
5. Izci H, Tambuyzer T, Tuand K, et al. A systematic review of estimating breast cancer recurrence at the population level with administrative data. *J Natl Cancer Inst*. 2020;112:979–988. doi:10.1093/jnci/djaa050
6. 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;110:273–281. doi:10.1093/jnci/djx200
7. A'mar T, Beatty JD, Fedorenko C, et al. Incorporating breast cancer recurrence events into population-based cancer registries using medical claims: cohort study. *JMIR Cancer*. 2020;6(2):1–10.
8. Cairncrossh ZF, Nelson G, Shack L, Metcalfe A. Validation in Alberta of an administrative data algorithm to identify cancer recurrence. *Curr Oncol*. 2020;27(3):e343–e346. doi:10.3747/co.27.5861
9. Lambert P, Pitz M, Singh H, Decker K. Evaluation of algorithms using administrative health and structured electronic medical record data to determine breast and colorectal cancer recurrence in a Canadian province: using algorithms to determine breast and colorectal cancer recurrence. *BMC Cancer*. 2021;21(1):1–10. doi:10.1186/s12885-021-08526-9
10. Pedersen RN, Öztürk B, Mellemkjær L, et al. Validation of an algorithm to ascertain late breast cancer recurrence using Danish medical registries. *Clin Epidemiol*. 2020;12:1083–1093. doi:10.2147/CLEP.S269962
11. Rasmussen LA, Jensen H, Virgilsen LF, et al. 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

12. Valachis A, Carlqvist P, Szilcz M, et al. Use of classifiers to optimise the identification and characterisation of metastatic breast cancer in a nationwide administrative registry. *Acta Oncol*. 2021;60(12):1604–1610. doi:10.1080/0284186X.2021.1979645
13. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35:1925–1931. doi:10.1093/eurheartj/ehu207
14. Het Intermutualistisch Agentschap [The Intermutualistic Agency] (IMA) L'Agence InterMutualiste (AIM). https://ima-aim.be/.
15. Technische Cel voor het beheer van de MZG-MFG data [Technical cel for management of MZG-MFG data]- La Cellule Technique pour la gestion des données RHM-RFM. https://tct.fgov.be/.
16. CBSS - Crossroads Bank for Social Security. Available from: https://www.ksz-bcss.fgov.be/nl/documents-list. Accessed April 28, 2023.
17. Agence pour une Vie de Qualité [Walloon Agency for quality of life] (AViQ). https://www.aviq.be/.
18. Smits N. A note on Youden's J and its cost ratio. *BMC Med Res Methodol*. 2010;10(1):1–4. doi:10.1186/1471-2288-10-89
19. Sutton CD. Classification and regression trees, bagging, and boosting. *Handb Stat*. 2005;24:303–329.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. *Classif Regres Trees*. 1984;20:1–358.
21. Chen Y, Yang Y. The one standard error rule for model selection: does it work? *Stats*. 2021;4(4):868–892. doi:10.3390/stats4040051
22. Enger SM, Soe ST, Buist DSM, et al. Breast cancer treatment of older women in integrated health care settings. *J Clin Oncol*. 2006;24 (27):4377–4383. doi:10.1200/JCO.2006.06.3065
23. Han Y, Sui Z, Jia Y, et al. Metastasis patterns and prognosis in breast cancer patients aged ≥ 80 years: a SEER database analysis. *J Cancer*. 2021;12 (21):6445. doi:10.7150/jca.63813
24. Xu Y, Kong S, Cheung WY, et al. Development and validation of case-finding algorithms for recurrence of breast cancer using routinely collected administrative data. *BMC Cancer*. 2019;19(1):1–10. doi:10.1186/s12885-019-5432-8
25. Chubak J, Onega T, Zhu W, et al. An electronic health record-based algorithm to ascertain the date of second breast cancer events. *Med Care*. 2017;55:e81–e87. doi:10.1097/MLR.0000000000000352
26. Kroenke CH, Chubak J, Johnson L, et al. Enhancing breast cancer recurrence algorithms through selective use of medical record data. *J Natl Cancer Inst*. 2016;108. doi:10.1093/jnci/djv336
27. Cronin-Fenton D, Kjærsgaard A, Nørgaard M, et al. Breast cancer recurrence, bone metastases, and visceral metastases in women with stage II and III breast cancer in Denmark. *Breast Cancer Res Treat*. 2018;167(2):517–528. doi:10.1007/s10549-017-4510-3
28. In H, Bilimoria KY, Stewart AK, et al. Cancer recurrence: an important but missing variable in national cancer registries. *Ann Surg Oncol*. 2014;21 (5):1520–1529. doi:10.1245/s10434-014-3516-x
29. Nout RA, Fiets WE, Struikmans H, et al. The in- or exclusion of non-breast cancer related death and contralateral breast cancer significantly affects estimated outcome probability in early breast cancer. *Breast Cancer Res Treat*. 2008;109(3):567–572. doi:10.1007/s10549-007-9681-x
30. 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
31. Nordstrom B, Whyte J, Stolar M, Catherine Mercaldi JK, Kallich JD. Identification of metastatic cancer in claims data. *Pharmacoepidemiology*. 2012;21(2):21–28. doi:10.1002/pds.3247
32. Nordstrom BL, Simeone JC, Malley KG, et al. Validation of claims algorithms for progression to metastatic cancer in patients with breast, non-small cell lung, and colorectal cancer. *Pharmacoepidemiol Drug Saf*. 2015;24(1, SI):511.
33. Pedersen RN, Oztü Rk Esen B€, Mellemkjaer L, et al. The incidence of breast cancer recurrence 10–32 years after primary diagnosis. *J Natl Cancer Inst*. 2021. doi:10.1093/JNCI/DJAB202
34. Lamont EB, Ii JEH, 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
35. Chawla N, Yabroff KR, Mariotto A, et al. Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. *Ann Epidemiol*. 2014;24(9):666–672.e2. doi:10.1016/j.annepidem.2014.06.099
36. Hassett MJ, Ritzwoller DP, Taback N, et al. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using 2 large contemporary cohorts. *Med Care*. 2014;52(10):e65–e73. doi:10.1097/MLR.0b013e318277eb6f
37. Sathiakumar N, Delzell E, Yun H, et al. Accuracy of medicare claim-based algorithm to detect breast, prostate, or lung cancer bone metastases. *Med Care*. 2017;55:e144–e149. doi:10.1097/MLR.0000000000000539
38. McClish D, Penberthy L, Pugh A. Using Medicare claims to identify second primary cancers and recurrences in order to supplement a cancer registry. *J Clin Epidemiol*. 2003;56(8):760–767. doi:10.1016/S0895-4356(03)00091-X