Case-finding Algorithms for Recurrence of Breast Cancer Using Machine Learning

Yuan Xu[1, 2, 3, 5]; Shiying Kong[1, 2, 3]; Winson Y. Cheung[2, 3]; Antoine Bouchard-Fortier[1, 2, 3]; Joseph Dort[1, 2, 3, 4]; Hude Quan [2, 5]; May Lynn Quan[1, 2, 3]

UNIVERSITY OF CALGARY

Introduction

In the era of precision medicine, overall survival is not adequate for assessing healthcare quality, comparing treatment efficacy, or informing decision making for patients with cancer, especially for cancers with long survival times such as breast cancer. Recurrence free survival (RFS) is more frequently investigated given that it provides more relevant information for cancer outcomes. However, cancer recurrence is not explicitly documented in administrative data such as cancer registry data, a widely utilized source for high volume, population based, multi-institutional research.

Currently, chart review is the only reliable way to obtain recurrence status but this is time-consuming and inefficient. This study aims to develop algorithms to detect breast cancer recurrence using the routinely collected administrative data from a Canadian prospective. These algorithms have the potential to be incorporated in the data repository for disease surveillance, monitoring and assessment of quality of care.

Methods

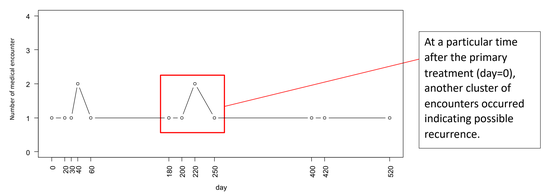
Recurrence of cancer is an event which usually requires intensive health care resources after the initial curative intent treatment such as re-operation, additional chemotherapy or radiation. This may be reflected by an increase of medical encounters. Therefore, physician claims data and other routinely collected administrative data provide a potential source for identifying recurrence.

The study cohort was derived from two population-based cohorts of breast cancer patients in Alberta, Canada with known high recurrence rates. It includes patients who were ≤40 years old and diagnosed between 2007 and 2010, along with patients who were diagnosed between 2012 and 2014 and received a neoadjuvant chemotherapy. Patients who had more than one type of tumor, or had stage IV breast cancer were excluded. The recurrence status was ascertained by primary chart review.

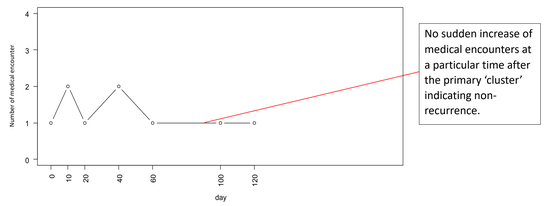
Algorithms were developed using CART (classification and regression tree) model. Study cohort was randomly divided into a training (60%) and a testing (40%) set. By setting different costs for misclassification, we developed different algorithms prioritizing sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Key variables for CART model included:

- Occurrence of second treatments (surgery, radiation therapy, or chemotherapy), or specific procedures (breast imaging, breast mammography or breast biopsy) after primary treatment.
- New cluster of visits to oncologist.
- Death from breast cancers after primary treatment/diagnosis.
- Characteristics of primary cancer including patient's age at diagnosis, tumor stage and type of surgery.



demonstration of the pattern of medical encounters (such as chemotherapy, radiation, surgery, and specialty care visit) for recurrence



demonstration of the pattern of medical encounters for non-recurrence

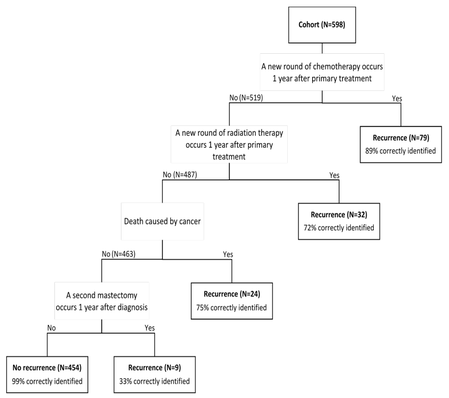
Parameter	All patients	Patients had no recurrence	Patients had recurrence
Mean Age (STD)	44.9 (12.3)	45.2 (12.5)	43.8 (11.3)
Mean length of follow up (STD), y	4 (1.9)	4.1 (1.8)	3.8 (2)
Stage			
Early-stage cancer (0-II)	406 (67.9%)	356 (74.6%)	50 (41.3%)
Advanced stage cancer (III)	192 (32.1%)	121 (25.4%)	71 (58.7%)
Tumor grade			
1	32 (5.3%)	32 (6.7%)	0 (0.0%)
2	211 (35.3%)	167 (35.0%)	44 (36.4%)
3	342 (57.2%)	270 (56.6%)	72 (59.5%)
Unknown	13 (2.2%)	8 (1.7%)	5 (4.1%)
ER status			
Positive	439 (73.4%)	355 (74.4%)	84 (69.4%)
Negative	159 (26.6%)	123 (25.6%)	37 (30.6%)
PR status			
Positive	375 (62.7%)	305 (63.9%)	70 (57.8%)
Negative	223 (37.3%)	172 (36.1%)	51 (42.2%)
HER2 status			
Positive	161 (26.9%)	138 (28.9%)	23 (19.0%)
Negative	437 (73.1%)	339 (71.1%)	98 (81.0%)

Parameter	All patients	Patients had no recurrence	Patients had recurrence
Surgery			
No	5 (0.8%)	4 (0.8%)	1 (0.8%)
BCS	159 (26.6%)	141 (29.6%)	18 (14.9%)
Mastectomy	434 (72.6%)	332 (69.6%)	102 (84.3%)
Chemotherapy			
Yes	547 (91.5%)	436 (91.4%)	111 (91.8%)
No	40 (6.7%)	31 (6.5%)	9 (7.4%)
Unknown	11 (1.8%)	10 (2.1%)	1 (0.8%)
Radiation therapy			
Yes	194 (32.4%)	150 (31.5%)	44 (36.4%)
No	209 (35.0%)	169 (35.4%)	40 (33.0%)
Unknown	195 (32.6%)	158 (33.1%)	37 (30.6%)
Hormone therapy			
Yes	214 (35.8%)	177 (37.1%)	37 (30.6%)
No	192 (32.1%)	149 (31.2%)	43 (35.5%)
Unknown	192 (32.1%)	151 (31.7%)	41 (33.9%)
Cancer caused death			
Yes	76 (12.7%)	10 (2.1%)	66 (54.5%)
No	522 (87.3%)	467 (97.9%)	55 (45.5%)

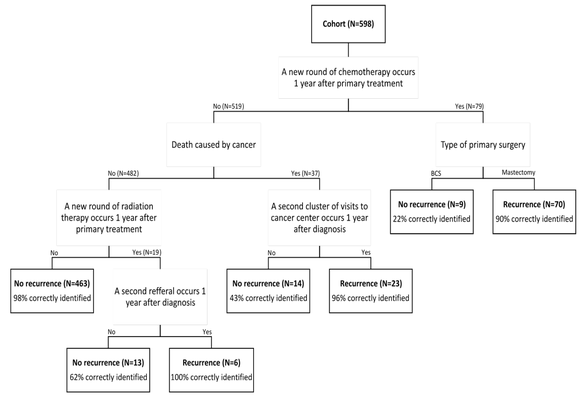
Result

In total, we included 598 patients with stage 0-III breast cancer. Among the 598 patients, we observed a 20.2% (121) recurrence rate along with a median follow-up of 4 years.

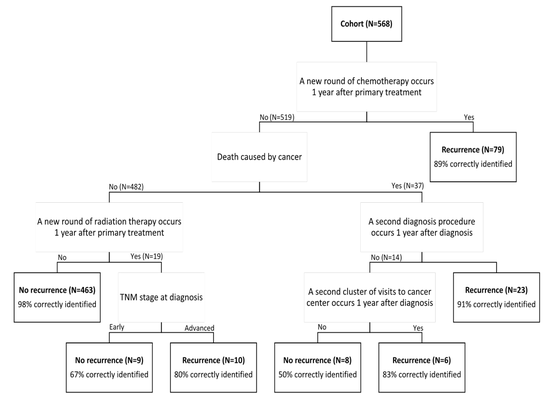
Performance of algorithms were evaluated by comparing the predicted outcome with the gold standard chart review. Validity metrics were calculated including sensitivity, specificity, PPV, NPV, accuracy and their corresponding 95% confidence interval based on an exact binomial distribution.



High sensitivity algorithm with sensitivity 94.2%, specificity 93.7% PPV 79.2%, NPV 98.5%, and accuracy 93.8%.



High PPV algorithm with sensitivity 75.2%, specificity 98.3%, PPV 91.9%, NPV 94.0%, and accuracy 93.6%.



High accuracy algorithm with sensitivity 86.0%, specificity 97.1%, PPV 88.1%, NPV 96.5%, and accuracy 94.8%

Algorithm	Sensitivity (%, 95% CI)	Specificity (%, 95% CI)	PPV (%, 95% CI)	NPV (%, 95% CI)	Accuracy (%, 95% CI)
High sensitivity	94.2 (90.1-98.4)	93.7 (91.5-95.9)	79.2 (72.5-85.8)	98.5 (97.3-99.6)	93.8 (91.9-95.7)
High PPV	75.2 (67.5-82.9)	98.3 (97.2-99.5)	91.9 (86.6-97.3)	94.0 (91.9-96.1)	93.6 (91.7-95.6)
High accuracy	86.0 (79.8-92.1)	97.1 (88.1-98.6)	88.1 (82.3-94.0)	96.5 (94.8-98.1)	94.8 (93.0-96.6)
Combining high sensitivity and high PPV algorithms plus chart review (7.5%)	94.2 (90.1-98.4)	98.3 (97.2-99.5)	93.4 (89.1-97.8)	98.5 (97.4-99.6)	97.5 (96.2-98.7)

Discussion and Conclusion

To the best of our knowledge, this is the first study to develop algorithms for identifying breast cancer recurrence using routinely collected administrative data in a publicly funded health system. One of the most important contributions of this study is the novel methodology used to explore the utility of underlying patterns of medical encounters in identifying recurrences. Moreover, this study also will provide framework for constructing algorithms to identify recurrence of other cancers using administrative data from a health system with universal health insurance coverage. Worth noting, is that we developed various algorithms that can be used for different research purposes.

This study has several limitations. First, the application of the proposed algorithms will need to be validated since they were developed by two population-based breast cancer cohorts with high risk of recurrence and only patients in Alberta were included. Second, PPV would change slightly when the algorithms are applied to populations with different rate of recurrence. Third, our algorithms were not designed to distinguish second primary breast cancers from recurrence.

The proposed algorithms achieved favourably high validity for identifying recurrence using widely available administrative data in a universal health system in Canada. Further study may be needed for external validation of the algorithms for widespread use.

UNIVERSITY OF CALGARY
CUMMING SCHOOL OF MEDICINE
Centre for Health Informatics

LIBIN CARDIOVASCULAR INSTITUTE of Alberta

the methods hub tools for better health.

ACORN ALBERTA CANCER OUTCOMES RESEARCH NETWORK

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

Contact

Yuan Xu
University of Calgary
Email: yuxu@ucalgary.ca

Author affiliation

[1] Department of Surgery, Foothills Medical Centre, University of Calgary, Calgary, Alberta, Canada
[2] Department of Community Health Science, Foothills Medical Centre, University of Calgary, Calgary, Alberta, Canada
[3] Department of Oncology, University of Calgary, Tom Baker Cancer Centre, Calgary, Alberta, Canada.
[4] The Ohlson Research Initiative, Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, Alberta, Canada
[5] Center for Health Informatics , University of Calgary, Calgary, Alberta, Canada