Ann Surg Oncol (2023) 30:2095–2103
https://doi.org/10.1245/s10434-022-12955-6

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

ORIGINAL ARTICLE – GLOBAL HEALTH SERVICES RESEARCH

A New Method of Identifying Pathologic Complete Response After Neoadjuvant Chemotherapy for Breast Cancer Patients Using a Population-Based Electronic Medical Record System

Guosong Wu, PhD[1,2], Cheligeer Cheligeer, PhD[2,3], Anne-Marie Brisson, MD, MSc[4], May Lynn Quan, MD, MSc[1,4,6], Winson Y. Cheung, MD, MPH[5,6], Darren Brenner, PhD[1,4], Sasha Lupichuk, MD, MSc[4], Carolin Teman, MD[7], Robert Barkev Basmadjian, MSc[1], Brittany Popwich, BSc[1,4,6], and Yuan Xu, MD, PhD[1,4,6]

[1]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; [2]The Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; [3]Alberta Health Services, Calgary, AB, Canada; [4]Departments of Oncology, Community Health Sciences, and Surgery, and The Center for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB , Canada; [5]Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; [6]Department of Surgery, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; [7]Department of Pathology and Laboratory Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

ABSTRACT

Background. Accurate identification of pathologic complete response (pCR) from population-based electronic narrative data in a timely and cost-efficient manner is critical. This study aimed to derive and validate a set of natural language processing (NLP)-based machine-learning algorithms to capture pCR from surgical pathology reports of breast cancer patients who underwent neoadjuvant chemotherapy (NAC).

Methods. This retrospective cohort study included all invasive breast cancer patients who underwent NAC and subsequent curative-intent surgery during their admission at all four tertiary acute care hospitals in Calgary, Alberta, Canada, between 1 January 2010 and 31 December 2017. Surgical pathology reports were extracted and processed with NLP. Decision tree classifiers were constructed and validated against chart review results. Machine-learning algorithms were evaluated with a performance matrix including sensitivity, specificity, positive predictive value (PPV), negative predictive value [NPV], accuracy, area under the receiver operating characteristic curve [AUC], and F1 score.

Results. The study included 351 female patients. Of these patients, 102 (29%) achieved pCR after NAC. The high-sensitivity model achieved a sensitivity of 90.5% (95% confidence interval [CI], 69.6–98.9%), a PPV of 76% (95% CI, 59.6–87.2), an accuracy of 88.6% (95% CI, 78.7–94.9%), an AUC of 0.891 (95% CI, 0.795–0.987), and an F1 score of 82.61. The high-PPV algorithm reached a sensitivity of 85.7% (95% CI, 63.7–97%), a PPV of 81.8% (95% CI, 63.4–92.1%), an accuracy of 90% (95% CI, 80.5–95.9%), an AUC of 0.888 (95% CI, 0.790–0.985), and an F1 score of 83.72. The high-F1 score algorithm obtained a performance equivalent to that of the high-PPV algorithm.

Conclusion. The developed algorithms demonstrated excellent accuracy in identifying pCR from surgical pathology reports of breast cancer patients who received NAC treatment.

First Received: 2 August 2022
Accepted: 1 December 2022
Published Online: 21 December 2022

Y. Xu, MD, PhD
e-mail: yuxu@ucalgary.ca

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death among females globally.[1] Curative-intent treatment of early-stage disease includes surgical resection and often other methods such as radiation, systemic therapies, or both.[2,3] Neoadjuvant

(preoperative) chemotherapy (NAC) is recommended in locally advanced cases to downstage the tumor and facilitate surgical resection. More recently, NAC has been offered for operable but high-risk phenotypes such as human epidermal growth factor receptor-2 (HER2)-positive and triple-negative cases (estrogen, progesterone and HER2-negative) when the primary tumor is at least 2 cm and/or if there is any axillary nodal involvement.

In this study, pathologic complete response (pCR) was defined as no remaining invasive cancer in the breast and/or axillary lymph nodes.[3–5] Systematic reviews and meta-analyses suggest that patients who attain pCR after NAC achieve significantly improved overall survival.[6,7] Recent reviews of randomized trials have demonstrated a survival benefit from additional adjuvant systemic therapies for high-risk phenotype tumors depending on whether pCR has been achieved or not.[8] Therefore, the presence of pCR is considered a surrogate end point for favorable long-term outcomes among breast cancer patients and plays a critical role in adjuvant systemic decision-making.[9,10]

Identifying pCR from large population-based electronic datasets that include a free text of patient pathology reports would allow the generation of real-world evidence indicating the impact of adopting new NAC regimens and the downstream effects from the use of adjuvant treatment options. Patients achieving pCR with NAC also could be systematically identified for prospective studies. Additionally, further investigation of factors associated with pCR could be accomplished with the goal of optimizing patient selection for NAC.

Electronic medical records (EMR) represent a promising data source containing patient-level clinical information, including pCR-related unstructured narrative descriptions in pathology reports. Machine-learning techniques such as natural language processing (NLP), which enables the processing and analysis of free text, have been widely applied in the field of medicine for the purpose of disease classification or prediction.[11,12] The development of an accurate NLP algorithm to help researchers and clinicians reliably identify pCR from pathology reports of breast cancer patients is warranted. However, research on this method is scarce.

This study aimed to develop and validate a set of NLP-based algorithms that can used to automate the detection of pCR from diagnostic biopsy pathology reports and final surgical pathology reports of breast cancer patients after NAC treatment embedded within an EMR.

METHODS

Study Population and Setting

This retrospective cohort study included all non-metastatic invasive breast cancer patients admitted to any of the acute care hospitals in Calgary, Alberta, Canada, between 1 January 2010 and 31 December 2017. The study excluded patients with a diagnosis of multiple breast primary tumors. Male patients were excluded from this study because the literature indicated lesser benefit of neoadjuvant therapy for men, and because they often opt for mastectomy.[13] The study followed the Standards for Reporting of Diagnostic Accuracy Study (STARD)[14] and was approved by Health Research Ethics Board of Alberta–Cancer Committee. A waiver of consent was granted.

Data Sources

The data were retrieved from two Alberta provincial databases: the Alberta Cancer Registry (ACR) database and the Sunrise Clinical Manager EMR. The ACR holds the legal mandate to record and maintain data on all new cancer diagnoses and cancer deaths in the province since 1942. During the study period, ACR was used to identify all patients with breast cancer. Structured patient data (e.g., age, year of diagnosis, lateral, biomarkers for breast cancer prognosis) were extracted.

The Sunrise Clinical Manager (SCM) EMR is an inpatient electronic medical record system universally applied in all four acute care hospitals in Calgary. The SCM EMR was used to extract unstructured patient data.[15] Raw free-text data documented in both diagnostic core biopsy reports and final surgical pathology reports, which include all biomarkers and histopathology evaluations, were extracted to develop NLP algorithms. Patient text notes without a lymph node evaluation report were excluded because true pCR achievement cannot be verified without lymph node evaluation. All databases then were linked by patient personal health care number (PHN) and unique lifetime identifiers (ULIs). Patient records without a valid PHN or ULI were excluded.

Chart Review to Ascertain the pCR as a Reference for Validation

To validate the developed algorithms, each pathology report was manually reviewed by a breast radiology fellow. Reports were categorized as "yes" if there was complete pathologic response or "no" if there was evidence of residual disease. The Food and Drug Administration (FDA) defines a pathologic complete response as the absence of

residual invasive malignant cells either with or without the presence of *in situ* disease.[16]

For the purpose of our study, cases with *in situ* disease (e.g., residual ductal carcinoma *in situ*) were not considered as cases with pathologic complete responses. All uncertainties raised from chart review were discussed and resolved to ensure that they satisfied the case definition. The data extraction agreement between the surgical fellow and a senior pathologist were tested to confirm the presence of pCR for the first 10 charts, and the result was excellent (kappa = 1). The result from this review served as the gold standard for validation of the developed pCR machine-learning algorithms.

Data Preprocessing

The free-text data of pathology reports documented in the EMR database were extracted and preprocessed through a series of predefined steps. First, sentence segmentation was used to break the text apart into separate sentences. We used the default sentence separator from SpaCy[17] with some custom Regular Expression (RE)-based rules (Supplementary).

The second step was concept extraction. We extracted biomedically related concepts from each sentence using a publicly available Python package named SciSpacy,[18] which integrated models that trained on a biomedical dataset to detect biomedical concepts from plain text.[17,19]

The third step was negation detection, in which we applied a Python package named NegSpacy. Concepts after a forward negation word (e.g., "not," "free of," "didn't") or before a backward negation word (e.g., "free," "absent") without the appearance of the negation's termination (e.g., "but," "however") were considered as negated concepts and excluded from our dataset.

Finally, we converted extracted concepts into Bag of Words vectors for the downstream pCR classification task.[20] The concepts with negation were differentiated from the detected concepts by assigning different positions in the vector. The rationale was to convert each EMR plain text into a list of occurrences of its contained medical concepts.

Model Development

We implemented decision tree algorithms that derived from the Scikit-learn Python package to summarize general patterns of pCR in free texts for its simplicity and interpretability.[21] The decision tree model is a tree-like classification model that contains decision nodes for splitting a tree by certain medical concepts and leaf nodes for indicating the classification results. The objective of the modeling procedure was to learn the decision tree structure from the data observation, which is an iterative procedure. The procedure selects an optimum decision node (name of medical concepts) for splitting the tree that leads to the precious classification.

Gini-impurity and entropy are two metrics applied mainly to quantify the goodness of the decision node. The decision tree has several adjustable hyperparameters that need to be set manually. By hyperparameter tuning, we explored the optimal combination of decision tree settings for desired outcomes, including high sensitivity, high specificity, and high F1 score. First, this included class weight adjustment. Imbalanced data are among the major challenges of applying machine-learning in the field of medicine.[22] Class weight can be tuned to change the penalty for wrong classification of different categories during the model training, and a balanced weight would optimize the model performance for minor classes.

Second, Gini-impurity and entropy were the criteria for calculating information gain. Decision tree algorithms used information gain to split a node.

Third, a pruning technique was applied to remove nodes that provided less additional information and to avoid overfitting. The complexity parameter of minimal cost-complexity pruning was tuned at the cutoffs of 0.01, 0.015, and 0.02.[23]

Finally, we used the height of the tree to determine the maximum number of decision nodes that one sample will go through to reach the classification result.

The stratified train-test split method was applied to divide data into two sets with a ratio of 8:2 for the purpose of model training and validation while keeping the same pCR rate within each set. We validated and chose trees that prioritized sensitivity, positive predictive value (PPV), and F1 score. Confidence intervals for sensitivity, specificity, and accuracy were calculated using "exact" Clopper-Pearson confidence intervals, PPV, and negative predictive value (NPV) with standard logit confidence intervals,[24] respectively.

Tree Analysis

We investigated all node concepts of the decision tree to ensure a logical clinical relevance associated with pCR and to avoid system errors (e.g., a pathologist's personal writing style). For example, we found that the word "score" was a strong indicator associated with not developing pCR during the initial attempt of tree model development. A review of original reports suggested that residual tumors after NAC often were measured and described via several commonly used measures to assess the tumor size, grade, and receptor status (e.g., Nottingham score, mitotic score, modified Bloom-Richardson [MBR] score, HER2 score). We combined these measures into one concept named

"score" and used it as a root node for retraining decision tree models (see Supplementary for details of concept combinations).

Statistical Analyses

Patient demographic and clinical characteristics were summarized as numbers and percentages or as medians and interquartile ranges (IQRs) as appropriate. The comparison of categorical variables was analyzed by chi-square or Fisher's exact test when applicable. The study evaluated NLP-based pCR algorithms with a performance matrix comprising sensitivity, specificity, PPV, NPV, accuracy, and F1 score. All statistical analyses were performed using Stata 16 software (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC) and Python 3.0.[25]

RESULTS

Patient Characteristics

Of the non-metastatic invasive breast cancer patients in this study, 874 were from Alberta and 425 were from Calgary. The study excluded 74 of these patients (Fig. 1). The final cohort consisted of 351 unique female patients (Table 1). Most of the patients were urban residents (87.5%) with a median age of 49 years (IQR, 42–56 years). The chart review results indicated that 102 (29%) of the patients achieved pCR after NAC. The patients who experienced pCR were more likely to be estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-positive than the patients without pCR. The patients who attained pCR were more likely to survive (96.1% vs 85.1%) by the end of study follow-up period (median, 3 years; IQR, 2–5 years).

Performance of NLP Algorithms

The uncertainty of the developed models was measured by entropy, and a pruning strength of 0.02 was used to generate the optimal trees. Balanced class weight was applied to derive the high-sensitivity model and the high-F1 score model, but not to the high-specificity model.

Using the chart review as our reference, we evaluated the performance of the developed NLP algorithms in a testing dataset. We chose the probability that could maximize the F1 score as the threshold of classification. In the testing dataset, the high-sensitivity algorithm achieved a sensitivity of 90.5% (95% CI, 69.6–98.9%), a specificity of 87.8% (95% CI, 75.2–95.4%), a PPV of 76% (95% CI, 59.6–87.2%), a NPV of 95.6% (95% CI, 85.1–98.8%), an

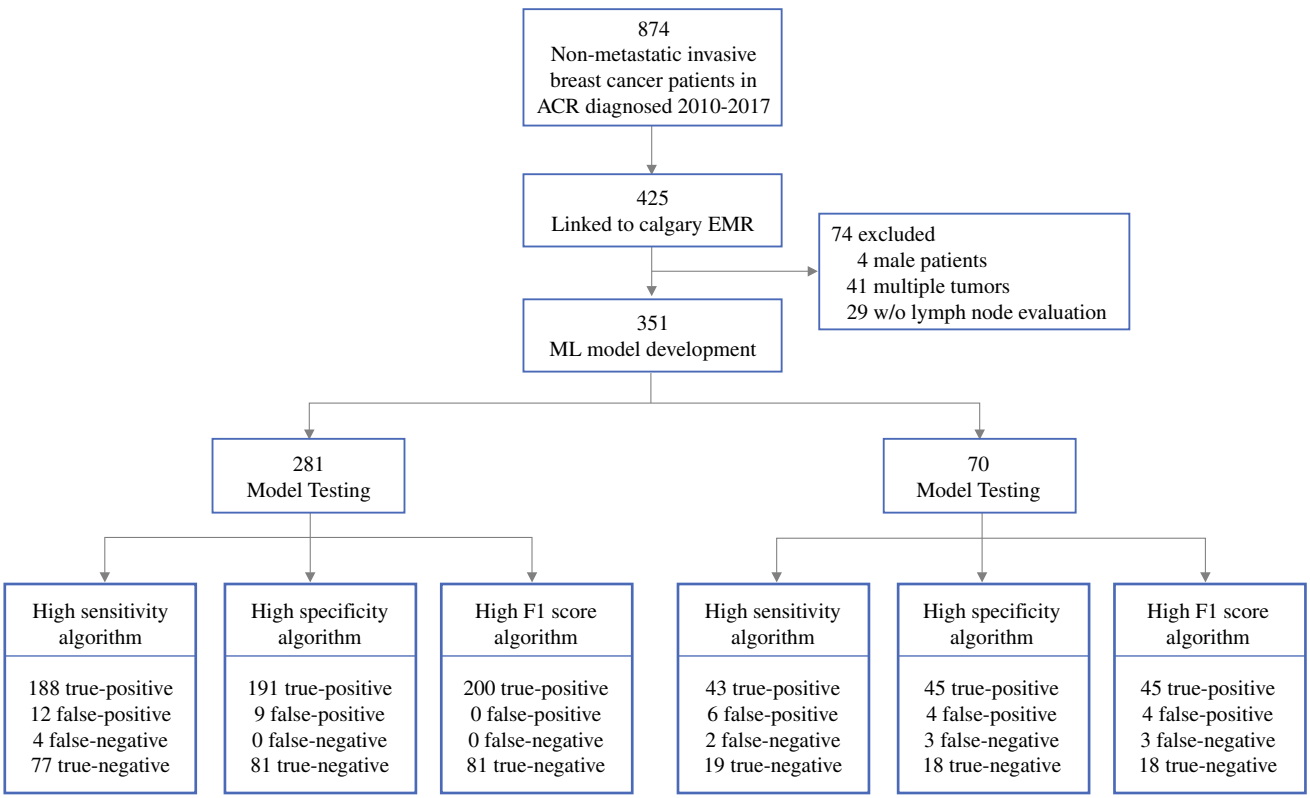

FIG. 1 Flow of participants. *ACR* Alberta cancer registry; *EMR* Electronic medical records; *ML* Machine-learning

TABLE 1 Demographic characteristics of patients with non-metastatic invasive breast cancer in Calgary diagnosed 2010–2017

Characteristic	Total (n = 351) n (%)	Patients with pCR (n = 102) n (%)	Patients without pCR (n = 249) n (%)	p Value
Median age: years (IQR)	49 (42–56)	48 (42–57)	49 (42–56)	0.918
Residency				0.323
Rural	44 (12.5)	10 (9.8)	34 (13.8)	
Urban	307 (87.5)	92 (90.2)	215 (86.2)	
Year of diagnosis				0.144
2010–2012	74 (21.1)	28 (27.4)	46 (18.5)	
2013–2015	162 (46.2)	41 (40.2)	121 (48.6)	
2016–2017	115 (32.7)	33 (32.4)	82 (32.9)	
Median study follow-up: years (IQR)	3 (2–5)	3 (2–5)	3 (2–5)	0.109
Lateral				
Left	172 (49.0)	47 (46.1)	125 (50.2)	0.483
Right	179 (51.0)	55 (53.9)	124 (49.8)	
TNM stage				0.000
1	8 (2.3)	6 (5.9)	2 (0.8)	
2	174 (49.6)	66 (64.7)	108 (43.4)	
3	169 (48.1)	39 (29.4)	139 (55.8)	
ER status				0.000
Positive	251 (71.5)	55 (53.9)	196 (78.7)	
Negative	100 (28.5)	47 (46.1)	53 (21.3)	
PR status				0.000
Positive	209 (59.5)	37 (36.3)	172 (69.1)	
Negative	142 (40.5)	65 (63.7)	77 (30.9)	
HER2 status				0.000
Positive	102 (29.1)	44 (43.1)	58 (23.3)	
Negative	241 (68.6)	58 (56.9)	183 (73.5)	
Unknown	8 (2.3)	0	8 (3.2)	
Status after last visit				0.015
Alive	310 (88.3)	98 (96.1)	212 (85.1)	
Death due to breast cancer	32 (9.1)	3 (2.9)	29 (11.7)	
Death due to other causes	9 (2.6)	1 (1)	8 (3.2)	

pCR Pathologic complete response; *IQR* Interquartile range; *TNM* Tumor-node-metastasis; *ER* Estrogen receptor; *PR* Progesterone receptor; *HER2* HUMAN epidermal growth factor receptor-2

accuracy of 88.6% (95% CI, 78.7–94.9%), an area under the receiver operating characteristic curve [AUC] of 0.891 (95% CI, 0.795–0.987), and an F1 score of 82.61. By optimizing the PPV, the high-PPV algorithm reached a higher PPV of 81.8% (95% CI, 63.4–92.1%) and a higher specificity of 91.8% (95% CI, 80.4–97.7%), but the sensitivity was decreased to 85.7% (63.7–97%) compared with the high-sensitivity algorithm. Overall, the high-PPV algorithm reached a higher F1 score of 83.72 and an accuracy of 90% (95% CI, 80.5–95.9%). The high-F1 score algorithm obtained performance equivalent to that of the high-PPV algorithm (Table 2).

The high-sensitivity algorithm is demonstrated with a classification tree in Fig. 2. The root node contained 351 breast cancer patients, and 102 (29.1%) of these patients were classified as pCR after NAC. This node split was based on the feature "score" to minimize node impurity. The vast majority (97.3%) of the 221 pathology reports that included the feature "score" did not show achievement of pCR. For 130 reports that did not include the "score" feature, the splitting process repeats on the next rule featured "tubule" and denied "residual" and stopped when additional splits yielded no further reductions in node

TABLE 2 Performance of NLP-based algorithms on test dataset

Algorithms	Sensitivity % (95 % CI)	Specificity % (95 % CI)	PPV % (95 % CI)	NPV % (95 % CI)	Accuracy % (95 % CI)	AUC % (95 % CI)	F1 score
High-sensitivity algorithm	90.5 (69.6–98.9)	87.8 (75.2–95.4)	76.0 (59.6–87.2)	95.6 (85.1–98.8)	88.6(78.7–94.9)	0.891 (0.795–0.987)	82.61
High-PPV algorithm	85.7 (63.7–97.0)	91.8 (80.4–97.7)	81.8 (63.4–92.1)	93.75(84.0–97.7)	90.0 (80.5–95.9)	0.888 (0.790–0.985)	83.72
High-F1 score algorithm	85.7 (63.7–97.0)	91.8 (80.4–97.7)	81.8 (63.4–92.1)	93.75(84.0–97.7)	90.0 (80.5–95.9)	0.888 (0.790–0.985)	83.72

CI Confidence interval; *PPV* Positive predictive value; *NPV* Negative predictive value; *AUC* Operating characteristic curve

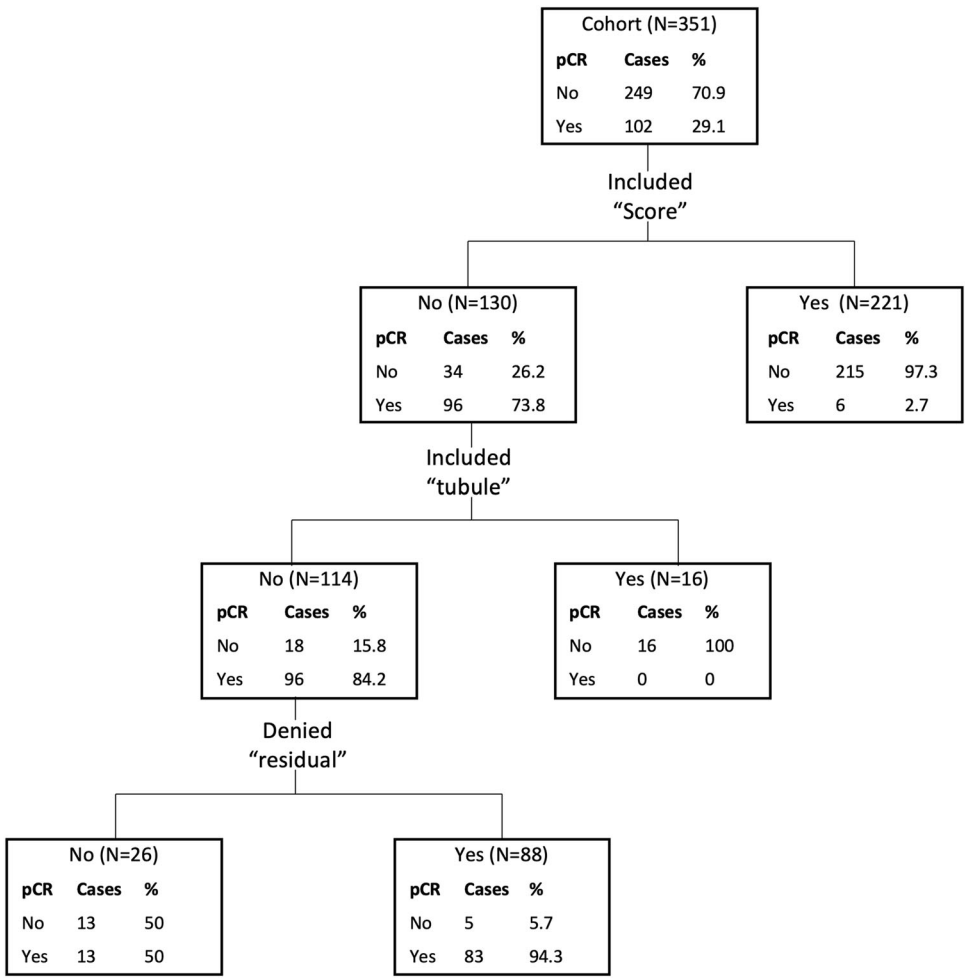

FIG. 2 High-sensitivity algorithm for identifying pathologic complete response after neoadjuvant chemotherapy in breast cancer. Note: "Yes" indicates that the criteria were met, and "No" indicates that the criteria were not met

impurity or reached the maximum height.Figure 3 demonstrates the flow of the high-PPV/F1 score algorithm.

DISCUSSION

This retrospective cohort study derived and validated machine-learning algorithms that can be used to identify pCR from hospital electronic pathology reports after NAC for breast cancer patients. The developed machine-learning algorithms demonstrated high sensitivity, PPV, accuracy, and F1 score. This suggests that the machine-learning algorithms hold immense promise for accelerating research and clinical care of breast cancer patients treated with NAC regimens.

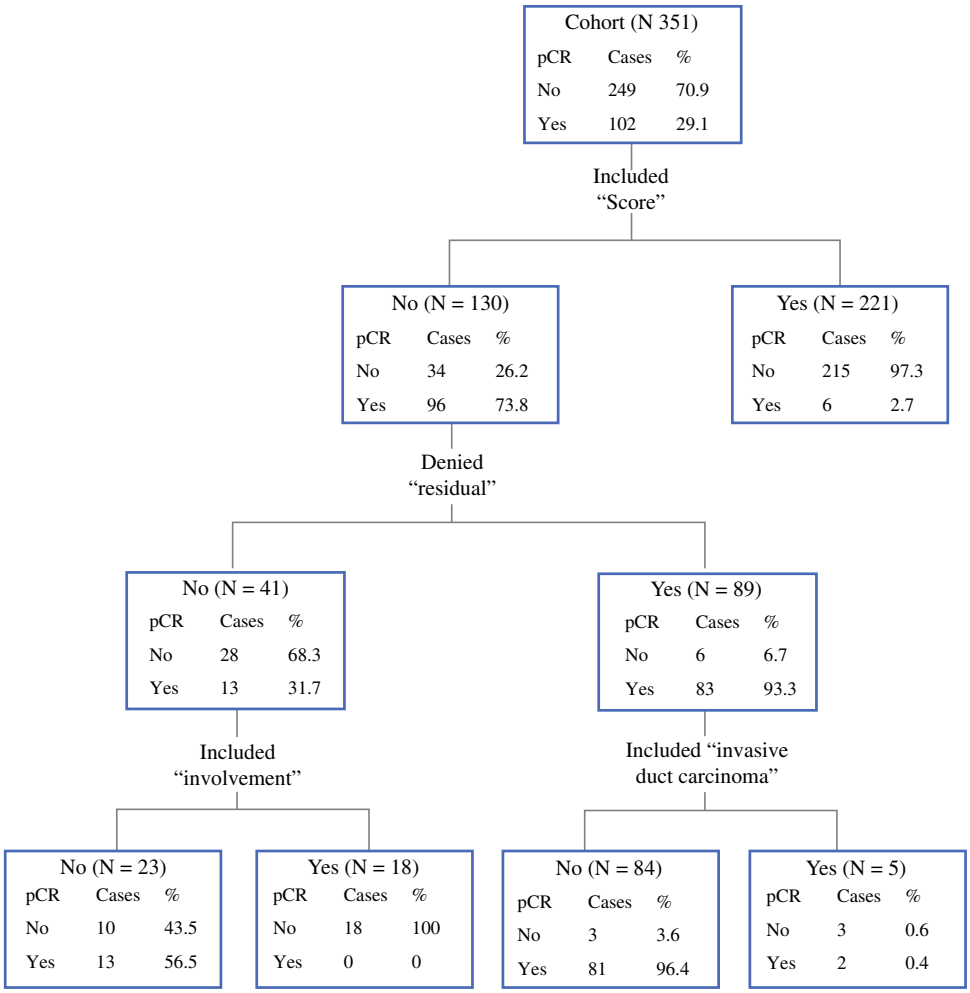

FIG. 3 High PPV/F1 score algorithm for identifying pathologic complete response after neoadjuvant chemotherapy in breast cancer. Note: "Yes" indicates that the criteria were met, and "No" indicates that the criteria were not met. *PPV* Positive predictive value

Manual extraction of patient clinical information (e.g., cancer characteristics) from pathology reports documented in electronic medical records is extremely expensive and time-consuming. Published literature suggests that it is valuable and trustworthy to extract clinical information through the development of machine-learning algorithms.[26] Yala et al.[27] developed a rule-based machine-learning algorithm to parse all carcinoma and atypia categories from patient pathology reports and achieved an accuracy of 90%.

In our population-based cohort study, we extracted pCR information from patient pathology reports using machine-learning algorithms. Prior studies were focused mostly on prediction of pCR with MRI images before surgeries.[28–31] For example, Sutton et al.[31] developed and validated a radiomics classifier that identified breast cancer pCR after NAC with MRI images before curative surgery. The best model achieved an AUC of 0.83 and can be used to assist the radiologist to improve diagnostic accuracy.

The literature suggests that the AUC for machine-learning algorithms to predict pCR of breast cancer patients after NAC from MRI images ranges from 0.72 to 0.97.[28–31] Our algorithms, derived from pathology reports of breast cancer patients after surgery, achieved an accuracy comparable with that of other machine-learning models derived from MRI images. The algorithms provide an innovated route for researchers and clinicians to identify pCR from breast cancer patients after NAC treatment.

Integrating negated concepts into the Bag of Vectors instead of neglecting them contributed to the accuracy and interpretability of the model given that pCR is describing the elimination of residual cancer. The denial of the word "residual" indeed plays an important role for machine-learning models to generate the correct classification. We attempted to build trees with less features by setting the maximum of candidate features to the logarithm of the total number of the whole vocabulary with a base of two or the square root of the total number of features. The results from those efforts indicated that models derived from a full vocabulary generate optimum performance. We adjusted the tree depth from three to five, and no performance boost was observed. As a result, a set of three level trees was selected in this study for its simplicity.

Residual cancer burden (RCB) is beneficial in determining pCR achievement (RCB, 0), but it requires detailed parameters for calculation as well as additional effort of documentation and often is not consistently available on the patient synoptic report. The developed machine-learning algorithms were internally validated and demonstrated excellent performance in identifying pCR among breast cancer patients after NAC from patient pathology reports. We achieved an accuracy of 83.7 and an F1 score of 90 for the developed algorithms.

Overall, the algorithms developed from this study can be translated into research and clinical care of breast cancer patients treated with NAC regimens. Specifically, the high-sensitivity algorithm could be applied when researchers need to identify patients who had a pCR after NAC to conduct large prospective or retrospective cohort studies. The high-PPV algorithm could be used to determine a cohort with the most pCR patients for clinical trials to test for interventions or to compare health care outcomes. The high-F1 score algorithm provides a happy medium between the two algorithms and could be used if research purposes are not fully established.

The methodology frameworks of this study also are applicable in other aspects of breast cancer classification or even other cancer types. The pCR algorithms have a great amount of downstream use such as determining the quality or effectiveness of a new alternative NAC or other systemic therapies as well as long-term survival or quality of life of patients with pCR. For example, the developed algorithms could serve as an initial screening tool for identifying pCR patients from a large patient cohort, followed by a panel review to confirm the achievement of pCR. Clinical specialists can take advantage of this tool to dramatically decrease the amount of time and costs associated with chart reviews. However, we suggest that a final classification of pCR should always be made from human intelligence.

Study Limitations

This study had several limitations. First, it was a population-based retrospective cohort study that included all breast cancer patients who underwent curative surgeries after NAC. The data size was moderate for fitting a machine-learning model.[32] We chose a tree-based classifier because it is robust for prioritizing the important features among high dimensionality.[33] Second, the imbalance data from this study might have compromised the accuracy of model classification.[22] Researchers tend to use case control designs to create a balanced dataset and to improve the performance of a machine-learning model.[34,35] We maintained the original pCR rate in algorithm development, believing this would increase the generalizability of the developed algorithms to real-world data. Third, developed algorithms were internally validated with a preserved dataset. External validation using data from other care systems is required. Finally, the applicability of our algorithms in other care systems must be carefully assessed.

CONCLUSION

In summary, we developed and validated a set of NLP algorithms that accurately capture the pCR from surgical pathology reports among breast cancer patients who underwent surgery after NAC treatment. The algorithms could greatly reduce the costs and labor for identifying pCR in large population-based datasets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1245/s10434-022-12955-6.

ACKNOWLEDGMENT This project, entitled "Building Pipeline to Transform Real-World Data to Evidence to Improve Cancer Care," was supported by the Canadian Cancer Society (CCS).

DISCLOSURE The authors declare that they have are no conflict of interest.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394–424.
2. Waks AG, Winer EP. Breast cancer treatment: a review. *JAMA*. 2019;321:288–300.
3. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primers*. 2019;5:1–31.
4. Dialani V, Chadashvili T, Slanetz PJ. Role of imaging in neoadjuvant therapy for breast cancer. *Ann Surg Oncol*. 2015;22:1416–24.
5. Mamounas EP. Impact of neoadjuvant chemotherapy on locoregional surgical treatment of breast cancer. *Ann Surg Oncol*. 2015;22:1425–33.
6. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384:164–72.
7. Spring LM, Fell G, Arfe A, et al. Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: a comprehensive meta-analysis. *Clin Cancer Res*. 2020;26:2838–48.
8. Pondé NF, Zardavas D, Piccart M. Progress in adjuvant systemic therapy for breast cancer. *Nat Rev Clin Oncol*. 2019;16:27–44.
9. Korn E, Sachs M, McShane L. Statistical controversies in clinical research: assessing pathologic complete response as a trial-level surrogate end point for early-stage breast cancer. *Ann Oncol*. 2016;27:10–5.
10. Cortazar P, Geyer CE. Pathological complete response in neoadjuvant treatment of breast cancer. *Ann Surg Oncol*. 2015;22:1441–6.

11. Locke S, Bashall A, Al-Adely S, et al. Natural language processing in medicine: a review. *Trends Anaesth Crit Care*. 2021;38:4–9.
12. Chowdhary K. Natural language processing. *Fund Artif Intell*. 2020. https://doi.org/10.1007/978-81-322-3972-7_19.
13. Duma N, Hoversten KP, Ruddy KJ. Exclusion of male patients in breast cancer clinical trials. *JNCI Cancer Spect*. 2018. https://doi.org/10.1093/jncics/pky018.
14. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6:e012799.
15. Lee S, Xu Y, D'Souza AG, et al. Unlocking the potential of electronic health records for health research. *Int J Population Data Sci*. 2020. https://doi.org/10.23889/ijpds.v5i1.1123.
16. Pathological Complete Response in Neoadjuvant Treatment of High-Risk Early-Stage Breast Cancer: Use as an Endpoint to Support Accelerated Approval Guidance for Industry. Food and Drug Administration, 2020.
17. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To Appear*. 2017;7:411–20.
18. Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace
19. Neumann M, King D, Beltagy I, et al. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv: arXiv:190207669(2019).
20. A comparison of event models for I Bayes text classification. AAAI-98 workshop on learning for text categorization; 1998. Citeseer.
21. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
22. Bekkar M, Djemaa HK, Alitouche TA. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*. 2013. https://doi.org/10.5121/ijdkp.2013.3402.
23. Breiman L, Friedman JH, Olshen RA, et al. Classification and regression trees. Routledge: New York, NY, 2017.
24. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case–control studies. *Stat Med*. 2007;26:2170–83.
25. van Rossum G. Python reference manual. *Department of Computer Science [CS]* 1995(R 9525).
26. Tang R, Ouyang L, Li C, et al. Machine learning to parse breast pathology reports in Chinese. *Breast Cancer Res Treat*. 2018;169:243–50.
27. Yala A, Barzilay R, Salama L, et al. Using machine learning to parse breast pathology reports. *Breast Cancer Res Treat*. 2017;161:203–11.
28. Cain EH, Saha A, Harowicz MR, et al. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast Cancer Res Treat*. 2019;173:455–63.
29. Li F, Yang Y, Wei Y, et al. Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer. *J Translat Med*. 2021;19:1–13.
30. Qu YH, Zhu HT, Cao K, et al. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (DL) method. *Thorac Cancer*. 2020;11:651–8.
31. Sutton EJ, Onishi N, Fehr DA, et al. A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy. *Breast Cancer Res*. 2020;22:1–11.
32. Song Y-Y, Ying L. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27:130.
33. Myles AJ, Feudale RN, Liu Y, et al. An introduction to decision tree modeling. *J Chemomet*. 2004;18:275–85.
34. Ford E, Rooney P, Oliver S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC Med Informat Decision Making*. 2019;19:1–9.
35. Kim H-E, Kim HH, Han B-K, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digital Health*. 2020;2:e138–48.