original reports

# Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer

Imon Banerjee, PhD[1]; Selen Bozkurt, PhD[1]; Jennifer Lee Caswell-Jin, MD[1]; Allison W. Kurian, MD, MSc[1]; and Daniel L. Rubin, MD, MS[1]

abstract

**PURPOSE** Electronic medical records (EMRs) and population-based cancer registries contain information on cancer outcomes and treatment, yet rarely capture information on the timing of metastatic cancer recurrence, which is essential to understand cancer survival outcomes. We developed a natural language processing (NLP) system to identify patient-specific timelines of metastatic breast cancer recurrence.

**PATIENTS AND METHODS** We used the OncoSHARE database, which includes merged data from the California Cancer Registry and EMRs of 8,956 women diagnosed with breast cancer in 2000 to 2018. We curated a comprehensive vocabulary by interviewing expert clinicians and processing radiology and pathology reports and progress notes. We developed and evaluated the following two distinct NLP approaches to analyze free-text notes: a traditional rule-based model, using rules for metastatic detection from the literature and curated by domain experts; and a contemporary neural network model. For each 3-month period (quarter) from 2000 to 2018, we applied both models to infer recurrence status for that quarter. We trained the NLP models using 894 randomly selected patient records that were manually reviewed by clinical experts and evaluated model performance using 179 hold-out patients (20%) as a test set.

**RESULTS** The median follow-up time was 19 quarters (5 years) for the training set and 15 quarters (4 years) for the test set. The neural network model predicted the timing of distant metastatic recurrence with a sensitivity of 0.83 and specificity of 0.73, outperforming the rule-based model, which had a specificity of 0.35 and sensitivity of 0.88 ($P < .001$).

**CONCLUSION** We developed an NLP method that enables identification of the occurrence and timing of metastatic breast cancer recurrence from EMRs. This approach may be adaptable to other cancer sites and could help to unlock the potential of EMRs for research on real-world cancer outcomes.



ASSOCIATED CONTENT

**Data Supplement**

Author affiliations and support information (if applicable) appear at the end of this article.

**Accepted on August 14, 2019 and published at ascopubs.org/journal/cci on October 4, 2019: DOI https://doi.org/10.1200/CCI.19.00034**



## INTRODUCTION

Although breast cancer mortality has declined over time, early-stage disease may recur as incurable distant metastases 15 years or more after initial treatment.[1] Despite rapid advances in the treatment and prognosis of early-stage breast cancer, far less is known about survival outcomes over time after metastatic recurrence, with few studies at the population level.[2] Yet metastatic recurrence is the path by which patients die of breast cancer, and a better understanding of this process is essential to develop and implement novel therapies that can reduce cancer mortality. Clinical trials enroll less than 5% of adult patients with cancer in the United States and thus are not broadly representative of the population.[3] Population-based cancer registries such as the SEER registry are funded to collect data only on the first course of cancer therapy[4] and cannot conduct the continuous follow-up that is necessary to capture the occurrence and timing of metastatic cancer recurrence. There is growing interest in clinic-based data sources, such as claims and medical record data, which may offer more clinically relevant details about the management and outcomes of distant metastatic recurrence.[5] However, creating such sources requires a substantial amount of manual curation to extract the relevant data elements.

Previous studies have developed algorithms to detect distant metastatic recurrence by either analyzing structured electronic medical record (EMR) data (eg, diagnostic and procedural codes)[6,7] or using natural language processing (NLP) approaches applied to free-text notes[8,9] (Data Supplement). The structured EMR data elements are relatively simple to extract and yield reasonable specificity in identifying metastatic recurrence. However, such approaches often yield low sensitivity for the detection of metastatic disease because the accuracy and completeness of implementation of

## CONTEXT

### Key Objective

Despite rapid advances in prognosis of early-stage breast cancer, far less is known about survival outcomes over time after metastatic recurrence because cancer registries only collect data on the first course of cancer therapy. Our objective was to develop an automated tool to identify patient-specific timelines of metastatic recurrence by parsing narrative clinical notes from an electronic medical record repository.

### Knowledge Generated

We designed a robust neural network model to detect the presence and timing of metastatic recurrence using a variety of clinical text notes from an electronic medical record platform. The neural network model predicted the timing of distant metastatic recurrence with a sensitivity of 0.83 and specificity of 0.73, outperforming the traditional rule-based model ($P < .001$).

### Relevance

The model has great potential to enhance understanding of real-world cancer outcomes and offers a potential advantage in terms of analyzing large volumes of longitudinal free-text notes, reducing manual chart review and the need for feature engineering.

diagnostic and procedure codes are limited by their inflexibility. For example, Nordstrom et al[10] included International Classification of Disease, Ninth Edition, codes for secondary neoplasms and drugs typically used for treating metastatic cancer in a classification and regression trees algorithm. They concluded that the sensitivity and predictive value were low and that additional sources of data on metastatic recurrence should be included.

Physicians' text notes, including progress notes and radiology and pathology reports, may offer the greatest nuance and detail about a patient's clinical status, including the presence and timing of metastatic cancer recurrence. Therefore, parsing of clinical narratives may significantly enhance sensitivity for metastatic recurrence detection. However, data extraction from free-text clinical notes is challenging as a result of their substantial variability. Conversion of clinical notes into a computer-manageable representation requires strategies such as NLP. Previous NLP approaches to detect metastatic recurrence have been limited by use of pathology reports only[11] or by reliance on rule-based pipelines (eg, prior knowledge based, regex)[8,12] that reduce generalizability outside of a single institution.

In the current study, we developed a robust NLP algorithm to detect the presence and timing of metastatic breast cancer recurrence using a variety of clinical text notes from a widely used EMR platform. We aimed to reduce manual intervention to make the recurrence algorithm easier to adopt and more broadly generalizable.

## PATIENTS AND METHODS

### OncoSHARE Cohort

With the approval of the Institutional Review Boards of Stanford University, we trained and validated the NLP algorithms on the OncoSHARE breast cancer research database.[13] OncoSHARE was developed through a collaborative effort that included Stanford University Health Care (SHC), an academic medical center in the San Francisco Bay Area; multiple sites of the Palo Alto Medical Foundation (PAMF), a community-based Sutter Health affiliate; and the California Cancer Registry (CCR), a SEER registry. OncoSHARE comprises a three-way data linkage at the individual patient level, with integration of retrospective EMR data from both SHC and PAMF linked to registry data from the CCR. For the current study, data from SHC EMRs only were used to develop the algorithms.

The structured fields in OncoSHARE include diagnosis, procedure, prescription, and laboratory orders. Unstructured data consist of free-text clinician notes, such as medical and social histories, impressions, and visit summaries. Complementary to EMR, the registry data from CCR contain demographics, tumor characteristics at initial breast cancer diagnosis, and up-to-date patient survival data. For this study, we focused on 8,956 patients treated at SHC, on whom 1,212,400 clinical notes were available to us.

### Manual Chart Review

Among the 8,956 patients with breast cancer, we selected 1,519 patients to establish a set of patients for whom the recurrence status (definite recurrence or no recurrence) was known (ground truth) to be used for training and validation of the NLP methods. To determine whether these patients experienced recurrence, we recruited three senior medical students to undertake a chart review of each patient using a Web-based in-house tool[14] (Data Supplement). The pathology report and progress notes that referred to pathology reports were the definitive source of recurrence information. However, a radiologic report read

 

as suspicious for metastasis, which subsequently led to a positive biopsy, could be used to establish the earliest date of the recurrence event. Each of the reviewers annotated approximately 500 patients in additional to 60 overlapping patients between readers for computing agreement (Data Supplement). The reviewers made the following two annotations for each patient: whether the breast cancer recurred in a distant anatomic region within the follow-up period, and if it did recur, the time stamp of the first encounter when there was evidence of recurrence (eg, pathologic confirmation). Subsequently, two senior oncologists removed the uncertain patients (Data Supplement), and finally, 894 patients served as the ground truth data set.

### Train-Test Set Splitting and Quarterly Division

To evaluate the NLP models against ground truth, we performed a patient-level separation of the 894 annotated patients by randomly selecting 179 patients as the test set (20%) and using the remaining 715 patients as a training and validation set (80%). Following the National Comprehensive Cancer Network guidelines for survillence,[15] we defined the time of recurrence of cancer in terms of the quarter of the year during the follow-up period in which breast cancer recurred, starting from the date of diagnosis (which establishes the beginning of quarter 1). The goal of the NLP methods we developed was to analyze all the clinical texts (ie, radiology and pathology reports and progress notes) for a patient during each quarterly text block and use that information to classify the patient as either having recurrent cancer or no recurrence of cancer (Fig 1). The NLP processing block is composed of basic text cleaning steps (eg, segmentation, signature removal, punctuation removal, number-to-string conversion) followed by named entity tagging. In the study, we dropped the quarters from the follow-up period in which the patient did not have any encounter in the SHC radiology, pathology, or oncology departments. Given that each patient had a distinct follow-up trajectory, the number of quarters varied for each patient. On average, patients in the training set had 19 quarters of follow-up (ie, 5 years of follow-up; range, one to 78 quarters) and patients in the test set had 15 quarters of follow-up (ie, 4 years of follow-up; range, one to 67 quarters). If a patient did not have any visits in a particular quarter (mainly no pathology or radiology reports or progress notes), we dropped that time point from our study.

### Knowledge-Based Processing of Quarterly Text Blocks

To capture the vocabulary for the intended task, we compiled the following two complementary dictionaries: the target term list, which was a publicly available terminology program (Clinical Event Recognizer[16]) extended with 430 additional metastatic terms that were primarily captured by analyzing the training set; and the modifier list, which was a list of modifier terms, including clinical terms related to negations (eg, no, rule out), temporality (eg, history, current), family (eg, mother, sister), anatomic locations (eg, brain, liver), risk, and discussion (eg, risk of, may introduce). A detailed description of the dictionary creation is provided in the Data Supplement. Finally, a keyword-based sentence retrieval method was applied on each quarterly text block, which selected only the sentences that contained at least one of the recurrence-related terms (terms from the target term list) as a named entity and generated a text snippet by combining the sentences extracted from the whole targeted quarter. On average, 17.16 sentences (± 37.43 sentences) were extracted from each quarter with 122.63 words (± 387.54 words).

### NLP Model Development and Evaluation

***Neural network NLP model.*** We developed a neural network model that automatically classifies clinical texts from each quarter of the year and computes a probability to reflect whether the patient's cancer has recurred within that quarter (Fig 2A). The neural network model consists of an input layer (read vectorized text block), hidden layers (transforming input using activation and creating embedding of the text block), a dropout layer (a number of hidden layer outputs are randomly dropped out to reduce overfitting), and finally a softmax layer for computing the probabilistic output. To build the vectorized representation of the text, we generated a vocabulary that describes clinical texts by parsing the unique words present in the training data set (quarters) having an optimal occurrence frequency cutoff (see experiments to identify the optimal frequency cutoff in Data Supplement). We vectorized the quarterly text blocks by representing the texts as a sequence of integer values, where each word is represented as a unique integer if the word also exists in the vocabulary. A mathematical description of the model design can be found in the Data Supplement. To optimize the two core hyperparameters of the network, we experimented with different settings of the number of hidden layers (one to six layers) and vocabulary size (50 to 2000 words), which is presented in the Data Supplement, and we found that two hidden layers and a vocabulary size of 1,000 words optimized the performance on the validation set (ie, 10% of the training data). The dimension of the layers (number of neurons) was determined according to two thirds of the size of vocabulary.

***Rule-based NLP method.*** To compare the benefit of our approach versus alternative, commonly used approaches, we created a rule-based method as a sequential NLP pipeline (Fig 2B) to identify the recurrent status from each quarterly text block. The final rule-based system was defined after an iterative process that consisted of several experiments during dictionary expansion and the rule development process (Data Supplement) from baseline rules to final extended rules. As domain knowledge, we supplied the candidate recurrence identification rules,

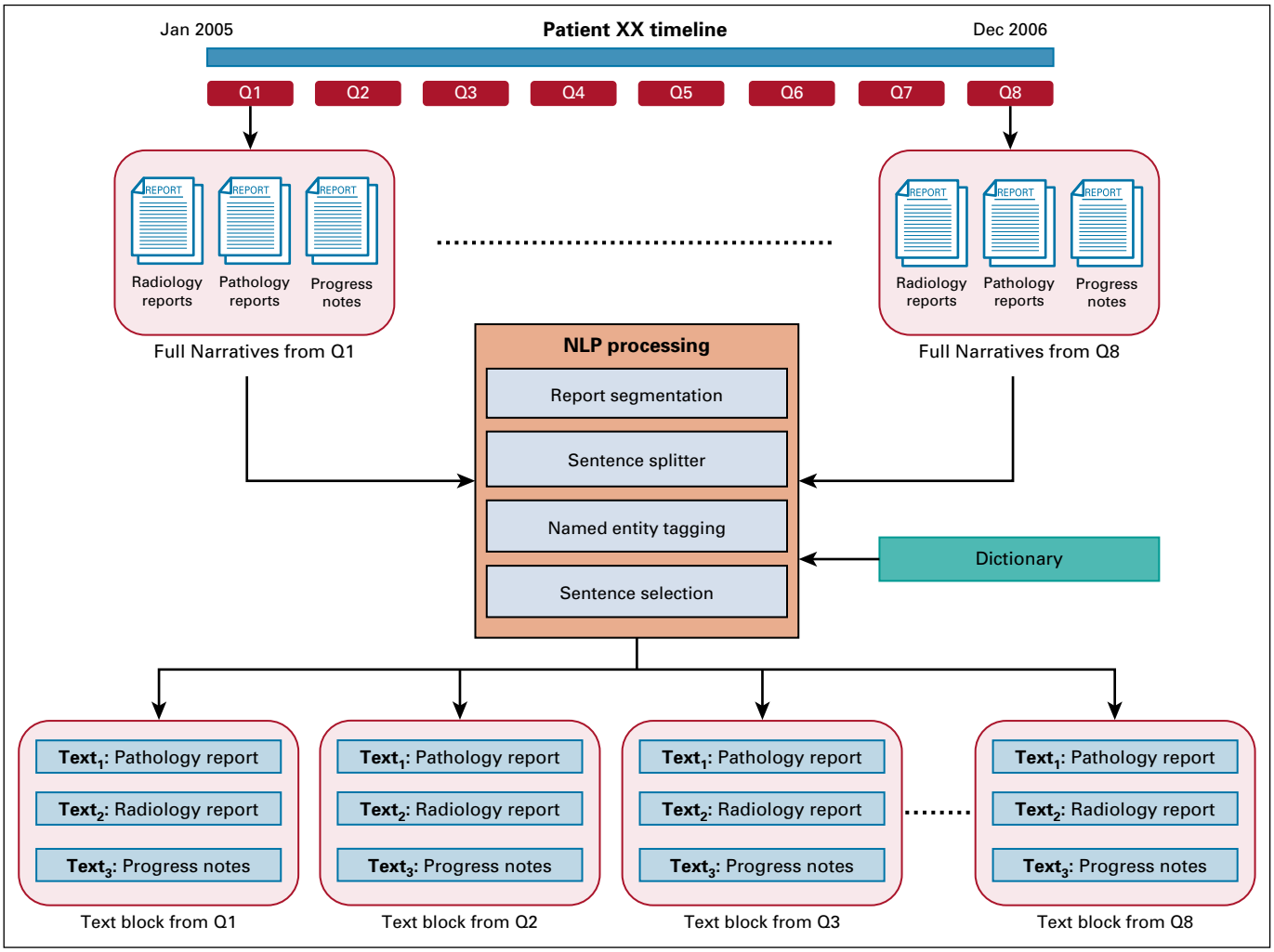


**FIG 1.** Composition of heterogeneous types of clinical narratives in a quarterly (Q) division using basic natural language processing (NLP) steps: a sample scenario of a patient with 2 years of follow-up.

which were generated by consulting the notes in the training set and the prior domain knowledge, including rules defined by previous systems.[8,17] The final rules were also developed in consultation with two oncologists and by performing experiments with different combinations of dictionary terms and candidate rules. The rule-based pipeline includes sequential text-processing tasks (Fig 2B) to read selected sentences from each quarterly text block generated by the knowledge-based processing pipeline and to infer the recurrence status of the patient for the quarter. A detailed description of the model design can be found in the Data Supplement.

## RESULTS

Figure 3 presents a digested workflow of the proposed NLP framework's development and evaluation components, starting from manual chart review and model development and progressing to evaluation and highlighting interactions between the models. Table 1 lists the patient and note characteristics of the data set. Survival status and missing stage information were completed by consulting the CCR.

Table 2 details the performance at the patient and quarter levels of three NLP models for identifying breast cancer recurrence, based on comparing the recurrence timeline generated by the proposed models against manual chart review on the test set (3,434 quarters from 179 patients). The baseline rule-based model presents a model that only includes publicly available Clinical Event Recognizer[16] basic terms and context related to negations and temporality (historical or hypothetical). The extended rule-based model presents our rule-based model with the final ruleset, and neural recurrence presents our neural network model. As seen from Table 2, all of the methods, including the simple rule-based baseline model, performed equally well in identifying the patients with no recurrence. However, the specificity for the definite recurrence class of the

 

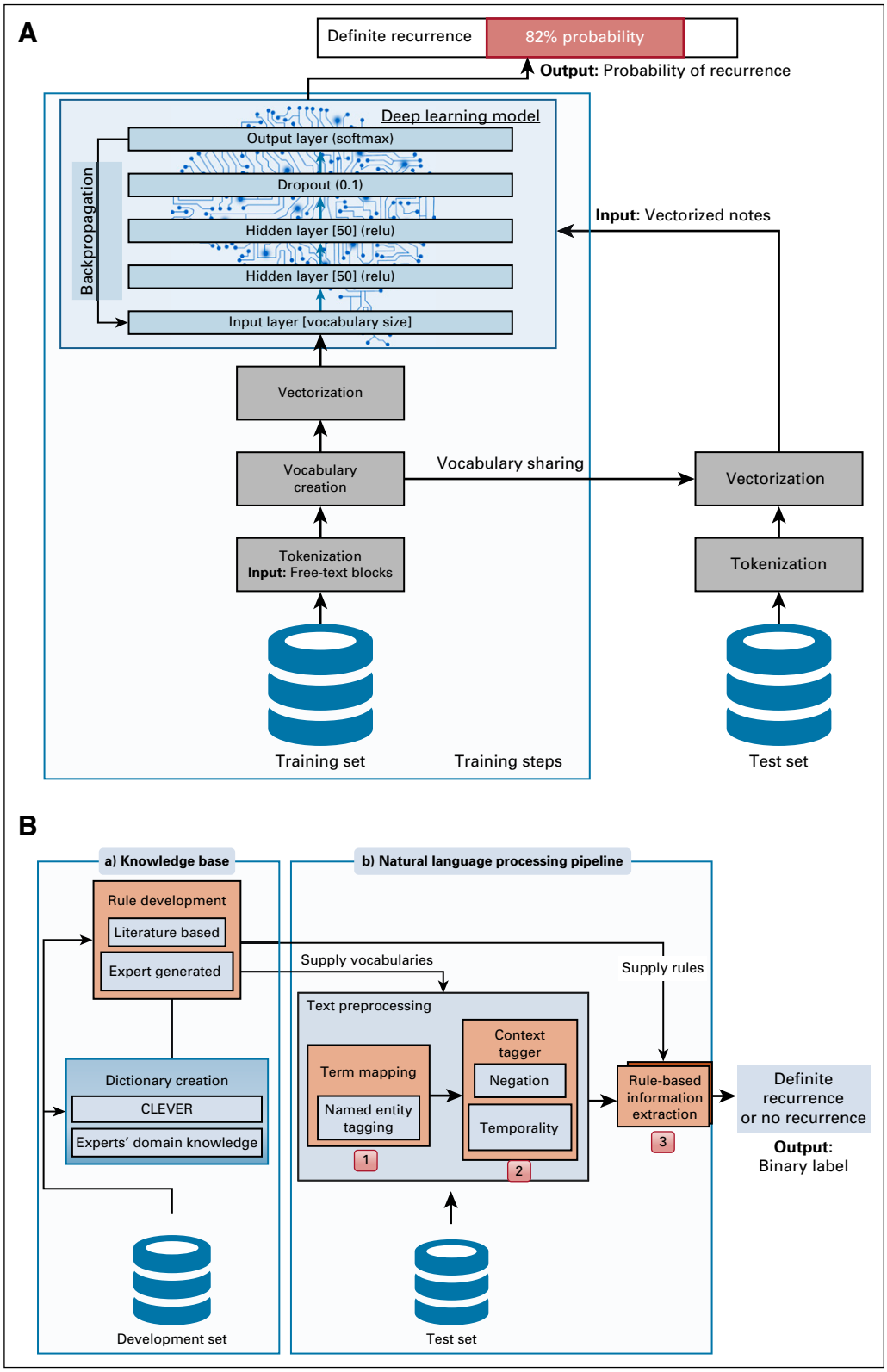


**FIG 2.** Natural language processing (NLP) pipelines for automated detection of breast cancer recurrence. (A) Neural recurrence NLP model where, for simplicity, we presented only the core blocks in the schema with two hidden layers. We experimented with varying the size of the vocabulary and varying the number of input layers, and the results are presented in the Data Supplement. (B) An alternative approach demonstrating a rule-based NLP pipeline where training steps are isolated from testing blocks for better interpretation of the pipeline. CLEVER, Clinical Event Recognizer.

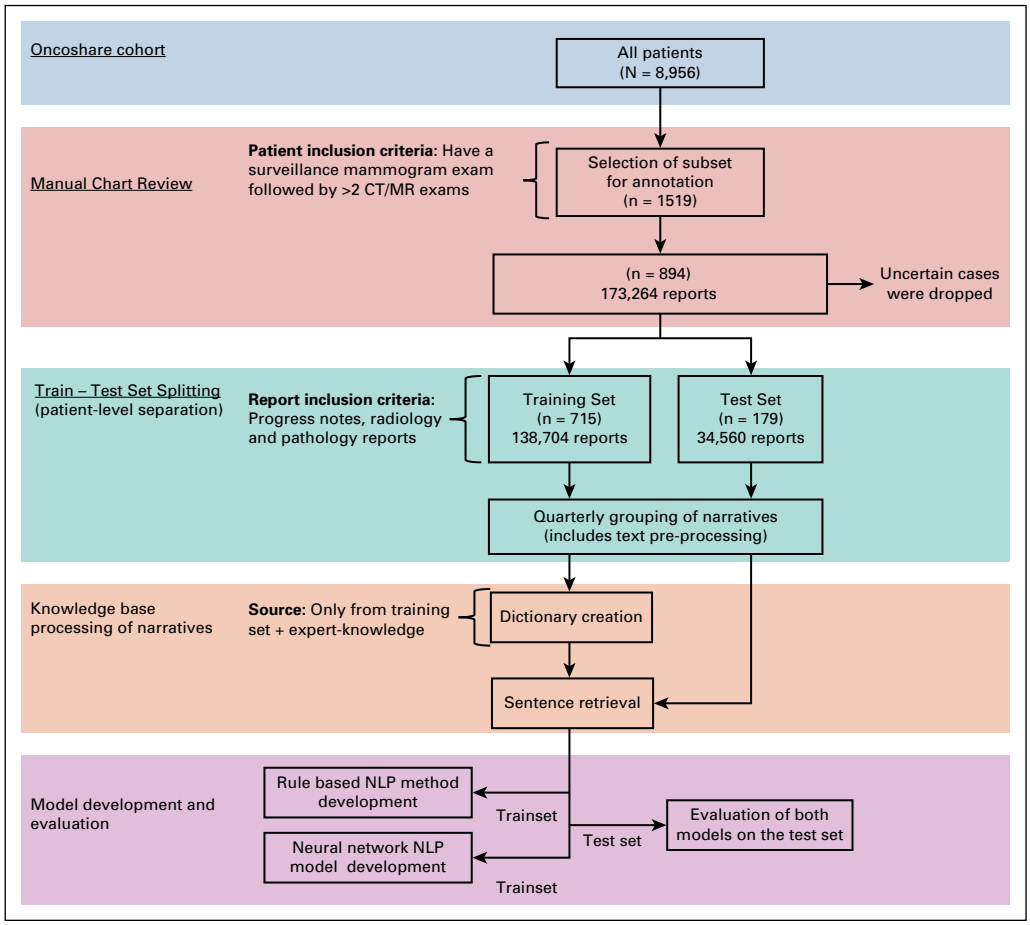


**FIG 3.** Pictorial view of the overall methodology, where each core component is color coded and arrows represent interactions between the components.

rule-based NLP methods is low, and thus, they generate more false-positive results for recurrence, which would require additional manual postprocessing to identify the timeline for definite recurrence. Figure 4A shows an example false-positive result that is incorrectly tagged as definite recurrence by the rule-based system when the text suggests only the suspicion of metastatic disease (concerning for possible metastatic disease).

The neural recurrence model provided a better trade-off between sensitivity and specificity for instances of definite recurrence and outperformed the baseline and extended rule-based models, and it preformed equally well for tagging instances of no recurrence. Moreover, we used a method called sensitivity analysis[18] for computing the relevance of each word in the input text for extracting recurrence. Sensitivity analysis takes the partial derivative of the loss function of the trained neural recurrence model with respect to each input word to derive the importance of the words for the recurrence classification task. The heat maps (Fig 4B) present results of the sensitivity analysis of input for instances of both no recurrence and definite recurrence, where the model places the most emphasis (dark text color) on "finding," "pulmonary," "diagnosis," "brain," and "metastatic," which indicate the possibility of definite recurrence. Figure 4B also shows that the model correctly interpreted uncertainty related to metastatic status in the sentences and categorized it as no recurrence. For instance, "current" and "worrisome" were assigned more importance (dark text color) than "metastases." Table 2 also represents the overall patient-level performance of the NLP methods, showing that the neural recurrence model also outperformed the rule-based systems for identifying patients with metastasis from the EMR system.

## DISCUSSION

Many have commented on the high priority of an efficient and accurate detection strategy for metastatic cancer recurrence in practice-based data.[4,8,19,20] Previous studies have developed cancer recurrence detection algorithms that process EMR data using diagnostic and procedural codes[6,7] and using NLP approaches to evaluate free-text notes.[8,9] These approaches yielded high sensitivity and specificity for identifying patients with metastasis, with somewhat lower accuracy for extraction of recurrence

 

**TABLE 1.** Characteristics of Patients in OncoSHARE: Overall, Training Set, and Test Set

| Characteristic | Whole Data Set (N = 8,956) | Training Set (n = 715) | Test Set (n = 179) |
|---|---|---|---|
| Age at primary diagnosis, years | | | |
| Mean | 54 | 54 | 55 |
| Standard deviation | 13 | 12 | 14 |
| Follow-up duration, years | 6 | 5 | 4 |
| Marital status, No. (%) | | | |
| Single | 1,354 (15) | 107 (15) | 32 (18) |
| Married | 5,869 (65) | 494 (69) | 110 (62) |
| Separated, divorced, or widowed | 1,539 (17) | 100 (14) | 30 (17) |
| Domestic partner | 11 (0.1) | 0 (0) | 0 (0) |
| Unknown | 183 (2) | 14 (0.1) | 2 (1) |
| Ethnicity, No. (%) | | | |
| Hispanic | 842 (9) | 46 (6) | 14 (8) |
| Non-Hispanic | 7,649 (85) | 472 (66) | 117 (65) |
| Unknown | 465 (5) | 197 (28) | 48 (27) |
| Race, No. (%) | | | |
| White | 6,726 (75) | 389 (54) | 100 (56) |
| Asian | 1,353 (15) | 99 (14) | 25 (14) |
| Black | 325 (4) | 19 (3) | 3 (2) |
| Pacific Islander | 49 (1) | 7 (1) | 1 1() |
| Native American | 17 (1) | 2 (0) | 0 (0) |
| Unknown | 486 (5) | 199(28) | 50 (28) |
| Stage, No. (%) | | | |
| 0 | 1,929 (22) | 109 (19) | 21 (12) |
| I | 2,894 (32) | 183 (33) | 39 (22) |
| II | 2,546 (28) | 183 (33) | 58 (32) |
| III | 861 (10) | 61 (11) | 21 (12) |
| IV | 442 (0.05) | 1 (0) | 0 (0) |
| Unknown | 284 (0.03) | 15 (0.03) | 1 (0.01) |
| Receptor status, No. (%) | | | |
| HER2 positive | 1,245 (14) | 74 (10.34) | 21 (11.73) |
| ER/PR positive and HER2 negative | 3,638 (40.62) | 259 (36.22) | 69 (38.55) |
| Triple negative | 944 (10.54) | 59 (8.25) | 14 (7.82) |
| Missing | 3,026 (33.79) | 321 (44.89) | 74 (41.34) |
| Other | 103 (1.15) | 2 (0.27) | 1 (0.56) |
| Grade, No. (%) | | | |
| Grade 1 (well differentiated) | 1,609 (18) | 128 (18) | 44 (25) |
| Grade 2 (moderately differentiated) | 3,363 (38) | 250 (35) | 62 (35) |
| Grade 3-4 (poorly differentiated) | 3,984 (44) | 337 (47) | 73 (41) |
| No. of notes | | | |
| Progress notes | 1,003,210 | 94,208 | 23,552 |
| Radiology reports | 163,098 | 35,004 | 8,761 |
| Pathology reports | 46,092 | 9,492 | 2,247 |

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

**TABLE 2.** Comparison of Performance of NLP Model at the Quarter Level and Patient Level: Classwise Sensitivity, Specificity, F1 Score, and AUC-ROC

| | Quarter-Level Performance of the NLP Methods | | | | | | | Patient-Level Performance of the NLP Methods | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Definite Recurrence | | | No Recurrence | | | | | | |
| Model | Specificity | Sensitivity | F1 Score | Specificity | Sensitivity | F1 Score | AUC-ROC | Specificity | Sensitivity | F1 Score |
| Baseline rule-based model | 0.2 | 0.9* | 0.33 | 0.99* | 0.83 | 0.9 | NA | 0.89 | 0.41 | 0.45 |
| Extended rule-based model | 0.35 | 0.88 | 0.5 | 0.99* | 0.92 | 0.95 | NA | 0.9 | 0.64 | 0.7 |
| Neural recurrence model | 0.82* | 0.73 | 0.77* | 0.99* | 0.99* | 0.99* | 0.9 | 0.95* | 0.93* | 0.94* |

Abbreviations: AUC, area under the curve; NA, not applicable; NLP, natural language processing; ROC, receiver operating characteristic.
*Best performance.

timing, whereas accurate identification of recurrence timing is essential for clinically relevant studies of the epidemiology and outcomes of metastatic cancer. If we can identify when patients experienced metastatic recurrence, we can answer key questions, such as whether new treatments are successfully extending life at the population level and whether the quality of metastatic cancer treatment varies between practice settings.

We developed a fully automated approach to scan free-text EMR progress notes in addition to radiology and pathology reports to generate a patient-level timeline of metastatic breast cancer recurrence for each 3-month (quarter of the year) period (Fig 4C). Our primary contribution in this study is a neural network model that has superior accuracy in dating metastatic recurrence within 3 months, which is also represented in the population-level analysis (Figs 4D and 4E). To our knowledge, our study is the first application of a neural network model to address this problem and show results of high accuracy. Our approach offers an efficient and generalizable strategy to detect and date metastatic recurrence—a clinically important event that is not currently captured in population-based cancer registries. Thus, it has great potential to enhance understanding of real-world cancer outcomes and offers a potential advantage in terms of analyzing large volumes of longitudinal free-text notes, reducing manual chart review and the need for feature engineering.

Despite the carefully harvested rules, rule-based NLP techniques often lack generalizability and require manual effort to tune the methods for a particular data set. Recent advances in machine learning approaches have provided NLP researchers with tools to create automated text sentiment classification models without the requirement for handcrafted feature engineering.[21] Challenges to applying such methods directly for information extraction from clinical text include modeling the ambiguity of the free-text narrative style for clinical reports, lexical variations, use of ungrammatical and telegraphic phases, and frequent appearance of abbreviations and acronyms.[22]

We compared our neural network approach with two different rule-based, NLP-based approaches to detecting metastatic recurrence. The rule-based systems used a manually curated, comprehensive vocabulary of metastatic-related terms and modifiers, which required substantial prior domain knowledge and manual labor to generate rules. We found the rule-based systems' performance to be suboptimal, but open-source vocabularies may be useful to expand seed terms in a different care environment. The neural network model required a sufficiently large training data set but substantially less prior knowledge or manual tuning than the rule-based approach. Our results demonstrate that, with sufficient training data, a neural network model can manage the linguistic ambiguity present in free-text notes; can learn how radiologists, pathologists, and oncologists express recurrence status as a nested hierarchy of words in the reports; and can outperform a curated rule-based system. These findings suggest that a neural network approach to the NLP task can achieve optimal performance without grammatical feature definitions, concept codes, or other predefined terms. Furthermore, these advantages make it readily generalizable to other health care systems and easily retrainable for cancer sites other than breast.

To achieve truly meaningful use of EMR data, it is crucial that the duration and quality of life after metastatic cancer recurrence be identifiable, comparable between treatment regimens and care environments, and amenable to ongoing surveillance over time. This is particularly urgent in breast cancer; for example, a recent pooled analysis of clinical trials[1] raised concern about long-term risks of metastatic recurrence, yet it lacked sufficient granularity of detail to identify correlates of this event. If applied to large, EMR-based data sets, our neural network strategy may identify specific patient characteristics, treatment regimens and toxicities that predict recurrence outcomes, informing the design of novel therapies and guiding clinical decision making.

Our study has limitations. Although the accuracy of the neural network model exceeds that of the rules-based approach, its prediction is not perfect. The most common source of error was limited documentation of metastatic recurrence in clinical notes, requiring that context-based

 

**A**

| No Recurrence Test Cases | | |
|---|---|---|
| | Rule-Based Result | Ground Truth |
| 1-cm lesion arising from the left adrenal, consistent with either a typical adenoma or **adrenal metastatic disease**. MRI showed **no definite metastatic lesions** in the lumbar spine and recommended correlation with a bone scan. Small, well-defined density within the left fifth rib, possibly representing a bone island, although given the patient's history, a tiny **metastatic focus** is difficult to exclude. The patient's abnormal bone scans are concerning for possible metastatic disease. | Recurrence = Present | Recurrence = Not present |

**B**

**Definite recurrence test cases**

N/A<NEW_NOTE>N/A interval development multiple new pulmonary nodules highly suspicious pulmonary metastases <NEW_NOTE> findings consistent metastatic breast adenocarcinoma <NEW_NOTE>

<NEW_NOTE> indication: metastatic breast carcinoma <NEW_NOTE> discussed prognosis implications metastatic disease N/A indication: metastatic breast carcinoma <NEW_NOTE> status post treatment recently found metastatic disease cerebellum <NEW_NOTE> <NEW_NOTE> december bone scan, well chest, abdomen, pelvis reportedly negative metastatic disease <NEW_NOTE> <NEW_NOTE> also march rd, chest, abdomen pelvis showed evidence metastatic disease <NEW_NOTE> <NEW_NOTE> march th, underwent craniotomy tissue diagnosis revealed metastatic adenocarcinoma compatible breast primary <NEW_NOTE> <NEW_NOTE> unfortunately, february complained headaches leading workup revealed metastatic disease cerebellum <NEW_NOTE> <NEW_NOTE> workup metastatic disease showed evidence findings outside brain <NEW_NOTE>

**No recurrence test cases**

N/A<NEW_NOTE> N/A bone marrow signal heterogeneous, especially endplates, likely representing degenerative disease, though settingbreast cancer, metastasis cannot entirely excluded <NEW_NOTE>

<NEW_NOTE> lytic osseous lesions seen current study worrisome metastases <NEW_NOTE>

**C**

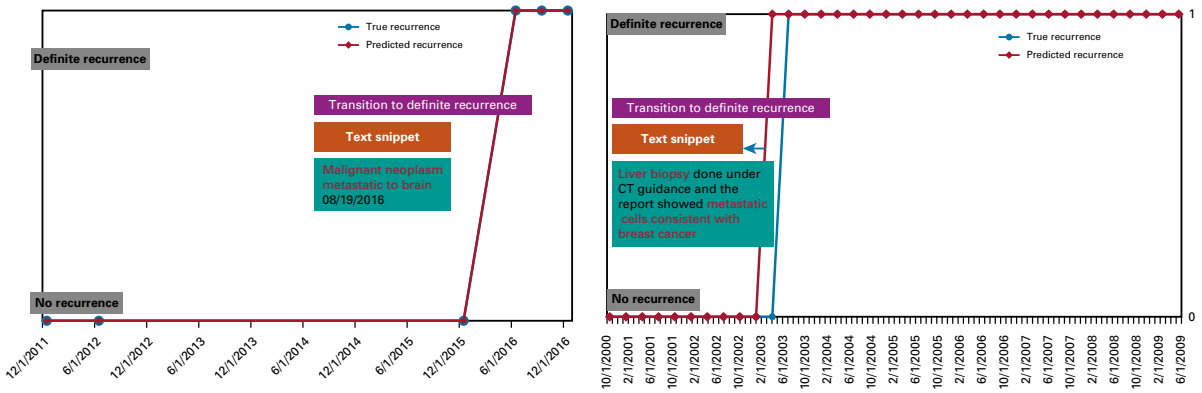


**FIG 4.** Interpretation of the natural language processing (NLP) models. (A) False-positive outcome of extended rule-based model as a result of suspicion of metastatic disease. (B) Results of sensitivity analysis of neural recurrence model: definite recurrence cases and no recurrence cases. Each cell represents sentences extracted from a quarter of the year. The darker color signifies more weight assigned by the model. <NEW_NOTE> signifies sentences extracted from a different report within the targeted quarter. (C) Two-test detection of time of recurrence: ground truth created by manual chart review (in red line) and inferred by the neural recurrence pipeline (in green) along with the text block from the quarter where recurrence status changed. Population-level analysis on the test data based on

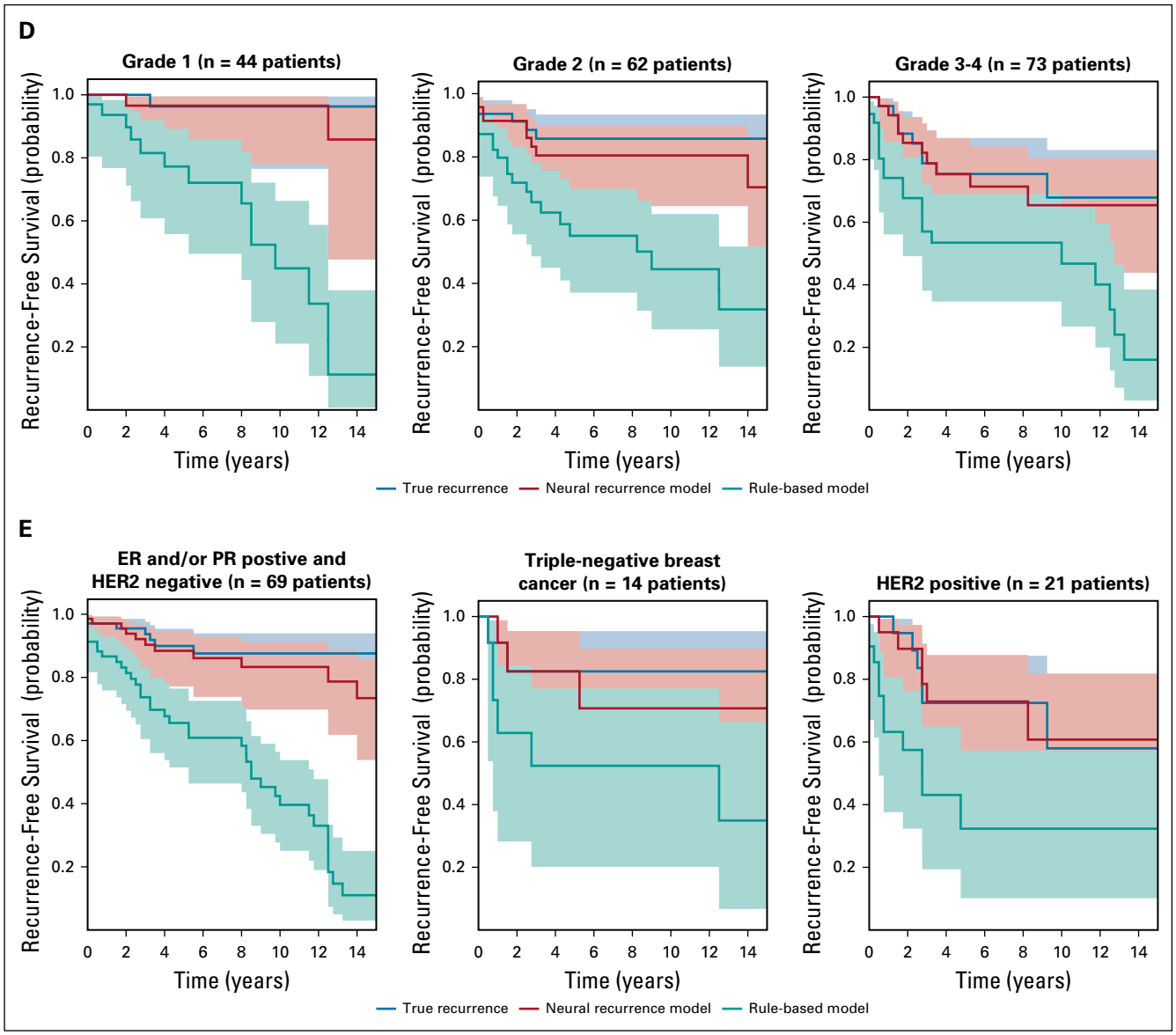


**FIG 4.** (Continued). (D) tumor grade and (E) receptor status, where the *y*-axis represents the recurrence-free survival after *t* years, with time (in years) shown on the *x*-axis. We see that the neural recurrence model's inferred recurrence rate matches for every tumor subtype with the manually curated ground truth compared with the rule-based system inferred recurrence rate. CT, computed tomography; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; MRI, magnetic resonance imaging; PR, progesterone receptor.

inferences be made. The results we present are also based on data from a single academic institution, with validation in a community-based setting (the PAMF component of OncoSHARE[23]) pending. An additional limitation is the relatively short median follow-up time of 5 years, because metastatic recurrence often occurs 5 or more years after breast cancer diagnosis.[1] In addition, understanding the basis for determinations made by neural networks is obscure, and we are in the process of exploring heatmap visualization to identify key components of the neural network model, which may enhance its face validity, trust, and intuitive appeal to clinicians.

In summary, we present a novel strategy to unlock the potential of EMR-based data regarding metastatic cancer recurrence. Important next steps include updating the neural network model as our patient follow-up time increases and validating its performance in diverse health care settings, both in the United States and internationally. To support reproducibility, we are publishing our trained model.[1] We also plan to extend our training data set size by obtaining manually curated, multi-institutional data to develop an NLP model that uses full notes as inputs instead of key sentences and to compare the performance with the current study design.

 

## AFFILIATION

[1]Stanford University School of Medicine, Stanford, CA

## CORRESPONDING AUTHOR

Imon Banerjee, PhD, Stanford University School of Medicine, 1201 Welch Rd, Stanford, CA 90305; Twitter: @ImonBanerjee6, @StanfordMed; e-mail: imon.banerjee@emory.edu.

## EQUAL CONTRIBUTION

A.W.K. and D.L.R. contributed equally to this work.



## SUPPORT

Supported by the Koo Foundation Sun Yat Sen Cancer Center, the Breast Cancer Research Foundation, the Susan and Richard Levy Gift Fund, the Suzanne Pride Bryan Fund for Breast Cancer Research, the Jan Weimer Junior Faculty Chair in Breast Oncology, the Regents of the University of California's California Breast Cancer Research Program (16OB-0149 and 19IB-0124), and the BRCA Foundation. The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; the Centers for Disease Control and Prevention's National Program of Cancer Registries, under Cooperative Agreement No. 5NU58DP006344; and the National Cancer Institute's SEER Program under Contract No. HHSN261201800032I awarded to the University of California, San Francisco, Contract No. HHSN261201800015I awarded to the University of Southern California, and Contract No. HHSN261201800009I awarded to the Public Health Institute, Cancer Registry of Greater California.


### AUTHOR CONTRIBUTIONS

**Conception and design:** Imon Banerjee, Selen Bozkurt, Allison W. Kurian, Daniel L. Rubin

**Financial support:** Allison W. Kurian

**Provision of study materials or patients:** Selen Bozkurt, Allison W. Kurian

**Collection and assembly of data:** Imon Banerjee, Selen Bozkurt, Jennifer Lee Caswell-Jin, Allison W. Kurian

**Data analysis and interpretation:** All authors

**Manuscript writing:** All authors

**Final approval of manuscript:** All authors

**Accountable for all aspects of the work:** All authors

### AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

**Allison W. Kurian**

**Research Funding:** Myriad Genetics (Inst)

**Other Relationship:** Ambry Genetics, Color Genomics, GeneDx/BioReference, InVitae, Genentech

**Daniel L. Rubin**

**Research Funding:** AstraZeneca (Inst)

**Patents, Royalties, Other Intellectual Property:** Several pending patents on artificial intelligence algorithms (Inst)

No other potential conflicts of interest were reported.

## REFERENCES


1. 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 377:1836-1846, 2017
2. Caswell-Jin JL, Plevritis SK, Tian L, et al: Change in survival in metastatic breast cancer with treatment advances: Meta-analysis and systematic review. JNCI Cancer Spectr 2:pky062, 2018
3. Unger JM, Cook E, Tai E, et al: The role of clinical trial participation in cancer research: Barriers, evidence, and strategies. Am Soc Clin Oncol Educ Book 35:185-198, 2016
4. Warren JL, Yabroff KR: Challenges and opportunities in measuring cancer recurrence in the United States. J Natl Cancer Inst 107:djv134, 2015
5. Presley CJ, Tang D, Soulos PR, et al: Association of broad-based genomic sequencing with survival among patients with advanced non-small cell lung cancer in the community oncology setting. JAMA 320:469-477, 2018
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 110:273-281, 2018
7. Hassett MJ, Uno H, Cronin AM, et al: Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. Med Care 55:e88-e98, 2017
8. Carrell DS, Halgrim S, Tran DT, et al: Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. Am J Epidemiol 179:749-758, 2014
9. 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 55:e81-e87, 2017
10. 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. Front Oncol 6:18, 2016
11. Zeng Z, Espino S, Roy A, et al: Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics 19:498, 2018 (suppl 17)
12. Soysal E, Warner JL, Denny JC, et al: Identifying metastases-related information from pathology reports of lung cancer patients. AMIA Jt Summits Transl Sci Proc 2017:268-277, 2017
13. Weber SC, Seto T, Olson C, et al: Oncoshare: Lessons learned from building an integrated multi-institutional database for comparative effectiveness research. AMIA Annu Symp Proc 2012:970-978, 2012
14. Lowe HJ, Ferris TA, Hernandez PM, et al: STRIDE: An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc 2009: 391-395, 2009
15. Goetz MP, Gradishar WJ, Anderson BO, et al: NCCN guidelines insights: Breast Cancer, Version 3.2018. J Natl Compr Canc Netw 17:118-126, 2019
16. Tamang S: CLEVER: CL-inical EVE-nt R-ecognizer. https://github.com/stamang/CLEVER
17. Strauss JA, Chao CR, Kwan ML, et al: Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Inform Assoc 20:349-355, 2013
18. Arras L, Horn F, Montavon G, et al: Explaining predictions of non-linear classifiers in NLP. https://arxiv.org/abs/1606.07298

19. 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 54:e47-e54, 2016
20. Mariotto AB, Zou Z, Zhang F, et al: Can we use survival data from cancer registries to learn about disease recurrence? The case of breast cancer. Cancer Epidemiol Biomarkers Prev 27:1332-1341, 2018
21. Kim Y: Convolutional neural networks for sentence classification. https://arxiv.org/abs/1408.5882
22. Banerjee I, Gensheimer MF, Wood DJ, et al: Probabilistic prognostic estimates of survival in metastatic cancer patients (PPES-Met) utilizing free-text clinical narratives. Sci Rep 8:10037, 2018
23. Kurian AW, Mitani A, Desai M, et al: Breast cancer treatment across health care systems: Linking electronic medical records and state registry data to enable outcomes research. Cancer 120:103-111, 2014