Detection of Metastatic Recurrence Timeline from Clinical Notes

1Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322
2Department of Biomedical Data Science, Stanford University School of Medicine, Stanford CA 94305
3Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford CA 94305
4Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford CA 94305

Introduction
While breast cancer mortality has declined over time, early-stage disease may recur as incurable distant metastases 15 years or more after initial treatment. 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. Population-based cancer registries (e.g., SEER) are funded to collect data only on the first course of cancer therapy and cannot conduct the continuous follow-up that would be 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. 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 detection either by analyzing structured electronic medical record (EMR) data, e.g., diagnostic and procedural codes or using natural language processing (NLP) approaches applied to free-text notes. 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 diagnostic and procedure codes are limited by their inflexibility. For example, Nordstrom et al. included ICD-9 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. Previous NLP approaches to detecting metastatic recurrence have been limited by use of pathology reports only or by reliance on rule-based pipelines (e.g., prior knowledge based, regex) that reduces 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 in order to make the recurrence algorithm easier to adopt and more broadly generalizable.

Method

Dataset: With the IRB approval from Stanford, we trained and validated the NLP algorithms on the Oncoshare breast cancer research database. Unstructured data consists of free-text clinician notes, such as progression notes, radiology and pathology reports. For this study, we focused on 8,956 patients treated at SHC, on whom 1,212,400 clinical notes were available to us. Among the 8,956 breast cancer patients, we selected 1,519 patients to establish a set of cases for which their recurrence status (definite recurrence or no recurrence) is known (“ground truth”) to be used for training and validation of the NLP methods. To determine whether or not these patients had recurrence, we recruited three senior medical students to undertake a chart-review of each case using a web-based in-house tool - STARR. Each of them annotated ~500 patients in additional to 60 overlapping cases between readers for computing agreement (cohen’s kappa>0.85). Subsequently, two senior oncologists removed the uncertain cases and finally 894 patients served as the ground truth dataset.

Train – Test set splitting and quarterly division: To evaluate the NLP models, we performed a patient-level separation of the 894 annotated patients, where we randomly selected 179 patients as test set (20%) and the remaining 715 patients as a training and validation set (80%). Following the NCCN Guidelines for surveillance, we defined the time of recurrence of cancer in terms of the quarter of the year during the follow-up period that breast cancer recurred, starting from the date of diagnosis. The goal of the NLP methods is to analyze all the clinical notes for a patient during each quarterly text block and use that to classify the patient as either having recurrent cancer or no recurrence of cancer. The NLP processing block is composed of basic text cleaning steps (e.g., segmentation, signature removal, punctuation removal, number to string conversion) followed by named entity tagging. On average, patients belonging to the training set contain 19 quarters [minimum 1 and maximum 78], i.e. 5 years follow up and test set contains 15 quarters [minimum 1 and maximum 67], 4 years follow up. If a patient did not have any visits on a particular quarter (mainly no pathology, radiology, and 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 two complementary dictionaries, (1) Target term list: a publicly available terminology CLEVER extended with 430
additional metastatic terms which were primarily captured by analyzing the training set, and (2) **Modifier list:** a list of modifier terms, which include clinical terms related to negations, temporality, family, anatomical locations, risk, and discussion (‘risk of’). Finally, a key-word based sentence retrieval method was applied on each quarterly text block, which selects only the sentences that contain at least one of the recurrence-related terms (terms from Target term list) as a named entity and generates a text snippet by combining the sentences extracted from the whole targeted quarter. On average, 17.16 (+/- 37.43) sentences were extracted from each quarter with 122.63 (+/-387.54) words. 

**NLP Model Development and Evaluation:** 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. The neural network model consists of an input layer (read vectorized text block), hidden layers (transforming input using non-linear activation and creating embedding of the text block), a dropout layer (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 trained a word embedding model by parsing the clinical notes from Stanford Cancer Registry database (excluding notes from test set). The dimension of the layers (number of neurons) is determined according to 2/3 of the size of vocabulary.

To compare the benefit of our approach over alternative, commonly used approaches, we created a rule-based method as a sequential NLP pipeline in order to identify the recurrent status from each quarterly text block. As domain knowledge, we supplied the candidate recurrence identification rules which were generated by consulting two oncologists and notes in the training set and the prior domain knowledge, including rules defined by previous systems.

**Result**

Table 1 summarizes the performance at the patient- and quarter-level 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). ‘Baseline rule-based’ presents a rule-based model which only includes publicly available CLEVER basic terms and context related to negations and temporality (historical or hypothetical). ‘Extended rule-based’ presents our rule-based model with the final ruleset, and ‘Neural recurrence’ presents our neural network model. As seen from the table, all of the methods, including the simple rule-based baseline model, performed equally well in identifying the “No recurrence” cases. However, the specificity for the “Definite recurrence” class of the rule-based NLP methods is low, and they generate more false-positive cases for recurrence, which would require additional manual post-processing to identify the timeline for definite recurrence.

**Table 1:** Comparison of NLP model’s performance at the quarter-level and patient-level: class-wise Sensitivity, Specificity, f1-score, and AUC-ROC. Best performance is highlighted in **bold.**

<table>
<thead>
<tr>
<th></th>
<th>Quarter-level performance of the NLP methods</th>
<th>Patient-level performance of the NLP methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definite recurrence</td>
<td>No recurrence</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Baseline rule-based</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Extended rule-based</td>
<td>0.35</td>
<td>0.88</td>
</tr>
<tr>
<td>Neural recurrence</td>
<td><strong>0.82</strong></td>
<td><strong>0.73</strong></td>
</tr>
</tbody>
</table>

The Neural recurrence model provided a better trade-off between sensitivity and specificity for “Definite recurrence” cases and outperformed the baseline and extended rule-based model, and it preformed equally well for tagging “No recurrence” cases. Table 1 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 metastatic patients from the EMR system.

**Conclusion**

We developed a fully automated approach to scan free-text EMR progress notes in addition to radiology and pathology reports, in order to generate a patient-level timeline of metastatic breast cancer recurrence for each three-month period. 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 and offer great potential to enhance understanding of real-world cancer outcomes without manual chart-review.

---