Journal of Healthcare Informatics Research (2019) 3:283–299
https://doi.org/10.1007/s41666-019-00046-3

RESEARCH ARTICLE

Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning

Zexian Zeng[1] · Liang Yao[1] · Ankita Roy[2] · Xiaoyu Li[3] · Sasa Espino[2] · Susan E Clare[2] · Seema A Khan[2] · Yuan Luo[1]

Received: 12 July 2018 / Revised: 30 November 2018 / Accepted: 7 January 2019 /
Published online: 8 April 2019

Abstract

Accurately identifying distant recurrences in breast cancer from the electronic health records (EHR) is important for both clinical care and secondary analysis. Although multiple applications have been developed for computational phenotyping in breast cancer, distant recurrence identification still relies heavily on manual chart review. In this study, we aim to develop a model that identifies distant recurrences in breast cancer using clinical narratives and structured data from EHR. We applied MetaMap to extract features from clinical narratives and also retrieved structured clinical data from EHR. Using these features, we trained a support vector machine model to identify distant recurrences in breast cancer patients. We trained the model using 1396 double-annotated subjects and validated the model using 599 double-annotated subjects. In addition, we validated the model on a set of 4904 single-annotated subjects as a generalization test. In the held-out test and generalization test, we obtained *F*-measure scores of 0.78 and 0.74, area under curve (AUC) scores of 0.95 and 0.93, respectively. To explore the representation learning utility of deep neural networks, we designed multiple convolutional neural networks and multilayer neural networks to identify distant recurrences. Using the same test set and generalizability test set, we obtained *F*-measure scores of 0.79 ± 0.02 and 0.74 ± 0.004, AUC scores of 0.95 ± 0.002 and 0.95 ± 0.01, respectively. Our model can accurately and efficiently identify distant recurrences in breast cancer by combining features extracted from unstructured clinical narratives and structured clinical data.

Keywords Breast cancer · Distant recurrence · Metastasis · NLP, EHR · Computational phenotyping · Convolutional neural networks · Multilayer perceptron

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s41666-019-00046-3) contains supplementary material, which is available to authorized users.

✉ Yuan Luo
yuan.luo@northwestern.edu

Extended author information available on the last page of the article

1 Introduction

Distant recurrences are defined as metastasis of the primary breast tumor to lymph nodes or organs beyond the loco-regional pathological field. Nodes located within the loco-regional field include ipsilateral axillary, ipsilateral internal mammary, supraclavicular, and intramammary lymph nodes [1]. Distant lymph nodes beyond the loco-regional field include cervical, contralateral axillary, and contralateral internal mammary lymph nodes. The most common sites of metastasis to organs are the bone, brain, lung, and liver [1]. It is important to distinguish between local and distant recurrences for several reasons: the categorization informs treatment decision-making and directs studies analyzing outcomes of local versus distant recurrences. Most importantly, the 10-year survival rates are much lower for distant recurrences as compared to local recurrences (56% after an isolated local recurrence as opposed to 9% after distant metastasis) [2]. The delineation can be an important prognostic marker for mortality.

The emerging cancer prognosis research has directed efforts towards identifying distant recurrence events accurately and efficiently. The National Program of Cancer Registries (NPCR) was launched to capture cancer patient information and one of its major tasks is to capture disease prognosis status for each cancer patient. However, many tumor registries fail to accurately identify cancer distant recurrences due to the significant human effort required for data maintenance [3, 4]. Manual chart review is one of the traditional methods used to identify breast cancer distant recurrences. Unfortunately, chart review is a time-consuming and costly process. It limits the number of samples available for research and is not feasible for large cohort studies. Furthermore, it is subject to human error in data analysis.

Computational phenotyping aims to automatically mine or predict clinically significant, or scientifically meaningful phenotypes from structured EHR data, unstructured clinical narratives, or combination of the two. In this study, we aim to develop a model to identify distant recurrences within a cohort of breast cancer patients. To develop the model, we utilized data collected in Northwestern Medicine Enterprise Data Warehouse (NMEDW), which is a joint initiative across the Northwestern University Feinberg School of Medicine and Northwestern Memorial HealthCare [5]. The NMEDW houses the EHR for about 6 million patients. Both structured and unstructured data are available in the NMEDW. Structured data typically capture patients' demographic information, lab values, medications, diagnoses, and encounters. Although readily available and easily accessible, studies have concluded that structured data alone are not sufficient to accurately infer phenotypes [6, 7]. For example, ICD-9 codes are mainly recorded for administrative purposes and are influenced by billing requirements and avoidance of liability [8, 9]. Consequently, these codes do not always accurately reflect a patient's underlying physiology. Furthermore, not all patient information (such as clinicians' observations and insights) is well documented in structured data [10]. As a result, using structured data alone for phenotype identification often results in low performance [7]. The limitations associated with structured data for computational phenotyping have encouraged the use of clinical narratives, which typically include clinicians' notes, observations, referring letters, specialists' reports, discharge summaries, and records of communication between doctors and patients [11]. These clinical narratives contain rich descriptions of patients' disease assessment, history, and treatments. However, the clinical narratives are not readily accessible without the use of

natural language processing (NLP). The abundance of information in the free text makes NLP an indispensable tool for text mining [12–14].

Our goal is to develop a system that combines structured EHR data and unstructured clinical narratives to accurately and efficiently identify distant recurrences in breast cancer. Such a model can be easily replicated and requires a minimum amount of human effort and input.

2 Related Work

Computational phenotyping has facilitated biomedical and clinical research across many applications, including patient diagnosis categorization, novel phenotype discovery, clinical trial screening, pharmacogenomics, drug-drug interaction (DDI) and adverse drug event (ADE) detection, and downstream genomics studies. Different NLP applications have also been developed to identify breast cancer local and distant recurrences. Carrell et al. [15] proposed a method to identify breast cancer subcohorts with ipsilateral, regional, and metastatic events using the concepts identified within the free text. The binary classification model achieved an *F*-measure scores of 0.84 and 0.82 in the training set and test set, respectively. However, the model could not distinguish a local recurrence from a distant recurrence. In addition, defining the number of hits in the system to segment the documents required substantial effort. Using morphology codes and anatomical sites from pathology reports, Strauss et al. [16] attempted to identify local and distant recurrences. However, their approach required that the pathology reports be well documented under a standard format. In addition, distant recurrence information identified from pathology report might be incomplete becasue the majority of distant recurrences in breast cancer have been diagnosed clinically rather than pathologically [17]. Haque et al. [18] applied a hybrid approach to identify breast cancer local and distant recurrences using a combination of pathology reports and EHR data. They achieved a relatively high NPV of 0.995 and a relatively low PPV of 0.65. This model also required a minimum amount of 10% manual chart review, which is still fairly time-consuming. In addition, the model was not able to distinguish between local, regional, or distant recurrences. NLP has also been applied to attempt retrieving distant recurrences for other types of cancer. Lauren et al. [19] tried to identify distant recurrences in prostate cancer from clinical notes, radiology reports, and pathology reports. They concluded that NLP could be used to identify metastatic prostate events more accurately than claim data.

Clinical narratives are known to have high-dimensional feature spaces, few irrelevant features, and sparse instance vectors [20]. These problems were found to be well addressed by SVMs [20], which also have been recognized for their generalizability and are widely used for computational phenotyping [21–24]. Carroll et al. [25] implemented a SVM model for rheumatoid arthritis identification using a set of features from clinical narratives using the Knowledge Map Concept Identifier (KMCI) [26]. They demonstrated that a SVM algorithm trained on these features outperformed a deterministic algorithm.

Recently, deep learning methods have been successfully applied to clinical text mining. Two representative deep learning models are convolutional neural networks (CNN) [27, 28] and recurrent neural networks (RNN) [29, 30]. They achieve state of

the art performances on a number of clinical text mining tasks. For instance, Gehrmann et al. [31] compared convolutional neural networks to the traditional rule-based entity extraction systems using the cTAKES and logistic regression using *n*-gram features. They tested ten different phenotyping tasks using discharge summaries. The CNN outperformed other phenotyping algorithms in the prediction of ten phenotypes, and they concluded that NLP-based deep learning methods improved the performance of patient phenotyping compared to other methods. Luo et al. applied both CNN and RNN to classify the semantic relations between medical concepts in discharge summaries from the i2b2-VA challenge data set and showed that CNN and RNN with only word embedding features can obtain similar performances compared to state-of-the-art systems by challenge participants with heavy feature engineering [32, 33]. Wu et al. [34] applied CNN using a set of pre-trained embeddings on clinical text for named entity recognition. They found that their models outperformed the baseline of conditional random fields (CRF). Jagannatha et al. [35, 36] experimented with RNN, long short-term memory (LSTM), gated recurrent units (GRU), bidirectional LSTM, combinations of LSTM with CRF, and CRF to extract clinical concepts from texts. They found that all variants of RNN outperformed the CRF baseline.

A combination of structured data and narratives for phenotyping has been found to improve model performances. DeLisle et al. [37] implemented a model to identify acute respiratory infections. They used structured data combined with narrative reports and demonstrated that the inclusion of free text improved the PPV score by 0.2–0.7 while retaining sensitivities around 0.58–0.75. In a study of the identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis, Lin et al. [38] obtained an *F*-measure of 0.83 in a performance evaluation. Liao et al. [39] implemented a penalized logistic regression as a classification algorithm to predict patients' probabilities of having Crohn's disease and achieved a PPV score of 0.98. Both Lin's and Liao's methods experimented with a combination of features from structured EHR- and NLP-processed features from clinical narratives. Their studies showed that the inclusion of NLP methods resulted in significantly improved performance.

3 Methodology

3.1 Cohort Description

Patients diagnosed with breast cancer between January 1, 2001 and December 31, 2015 were drawn from NMEDW. Patients were identified by ICD-9 codes. In total, 19,874 females were included. Within this cohort, to rule out the subjects that were not primarily diagnosed and treated in our hospital, only cases with at least one surgical pathology report documented in the desired time window were selected ($N = 7060$). Furthermore, 161 subjects with initial stage IV breast cancer were ruled out for further study. In total, 6899 subjects were identified and included in this study. The workflow to generate this data set is presented in Fig. 1.

To establish a gold standard for algorithm development, each patient was assigned a definite distant recurrence status ("yes" or "no") according to manual chart review. Only metastasis of the primary breast tumor to lymph nodes or organs beyond the loco-regional pathological field were defined as distant recurrences. In total, 1995 subjects

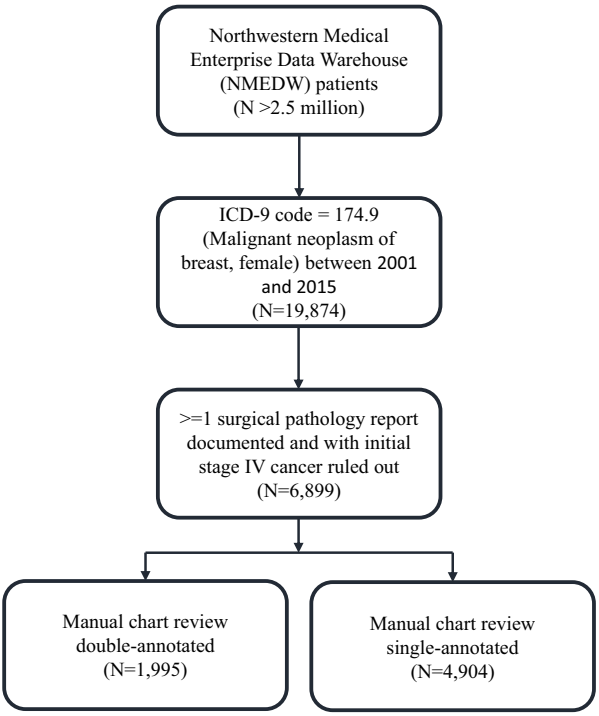


Fig. 1 Workflow to identify the cohort

were annotated twice by two annotators (co-authors, medical student AR; breast surgery fellow SE) and were included for model training and validation. The inter-rater agreements for the two annotators were measured by Cohen's kappa score, and the obtained score is 0.87 [40]. The items without agreements were resolved by a discussion between the two annotators. The other 4904 subjects were annotated once by annotators (co-authors, post-doc fellow XL; Ph.D. candidate ZZ) and were used as an independent set for model generalization test. These annotations were conducted over a span of 15 months (completed September 2017).

The 1995 double-annotated subjects were randomly split into a cross-validation set and a held-out test set according to a 7:3 ratio. In the cross-validation set, fivefold cross validation was applied with the 1396 samples. Among these 1396 samples, 138 distant recurrence events were identified; among the 599 samples in held-out test set, 55 distant recurrences were identified. In the generalization test set, 443 distant recurrences were identified among the 4904 samples. The cohort distribution is shown in Table 1.

3.2 Structured Clinical Data

Automated SQL codes were developed to query structured data from NMEDW. In total, 18 structured clinical variables were retrieved or derived. The variable names and corresponding categories or values are displayed in Table 2. Demographic data such as the age of diagnosis, race, smoking history, alcohol usage, family cancer history, and insurance type were queried. Smoking history is categorized as "yes," "no," "ex-

Table 1 Cohort distribution in the training and generalization set

	Total	Distant recurrence	Percentage (%)	Overall percentage (%)
Double-annotated set	1995	193	9.87%	9.22%
Cross-validation set	1396	138	9.89%	
Held-out test set	599	55	9.19%	
Single-annotated set	4904	443	9.03%	

smoker," or "unknown." Alcohol usage is categorized as "no," "moderate," "heavy," "former," or "unknown." Family history was self-reported all cancer histories in extended family members. Tumor characteristics and biomarkers, such as estrogen receptor (ER), progesterone receptor (PR), HER2, P53, nodal positivity, histology, tumor grade, and tumor size were retrieved. Nodal positivity is a variable to indicate whether any positive nodes were found in the axillary lymph nodes examination, and is categorized as "positive," "negative," or "unknown." The variable histology and nodal positivity were selected, because subjects with invasive ductal breast cancer or positive lymph nodes are more likely to develop a distant recurrence compared to those that have ductal in situ or negative lymph nodes [41]. IDC is invasive ductal carcinoma, DCIS is ductal carcinoma in situ, ILC is invasive lobular carcinoma, and network category is the network of patient's insurance plan. Primary surgery type is categorized as "Breast conservation surgery," "mastectomy," "no," or "unknown."

Table 2 The name and corresponding categories (values) of the 18 retrieved structured clinical variables

Variable name	Category
Age of diagnosis	Continuous
Race	White, Black, Asian, other
Smoking history	Yes, no, ex-smoker, unknown
Alcohol usage	No, moderate, heavy, former, Unknown
Family cancer history	Yes, no, unknown
Insurance type	Network category
Estrogen receptor	Positive, negative, unknown
Progesterone receptor	Positive, negative, unknown
HER2	Positive, negative, unknown
P53	Positive, negative, unknown
Nodal positivity	Positive, negative, or unknown
Histology	IDC, DCIS, ILC, unknown
Grade	Grade1, grade2, grade3, unknown
Size	0–2 cm, 2 cm–5 cm, > 5 cm, unknown
Surgery type	Mastectomy, breast conservation surgery, unknown
Deceased	Yes, no
Targeted therapy	Yes, no
Radiation	Yes, no

Additional clinical variables were derived to help identify distant recurrences. Variables of deceased, targeted therapy, and radiation were developed. The deceased variable is a binary variable to indicate whether a patient was deceased before the age of 75. Intuitively, patients with distant recurrences might have a shorter survival length compared to the women who do not have distant recurrences. After a discussion with a domain expert (co-author SK), we chose the age of 75 as the cutoff. Another variable "targeted therapy" is a binary variable created to indicate whether the patient has taken any of the following drugs: "Afinitor," "Everolimus," "Bevacizumab," "Avastin," "Ibrance," or "Palbociclib." These drugs are prescriptions for patients with distant recurrences. An additional variable "radiation" is a binary variable indicating whether the subject has received radiation treatment at the site of metastases, such as brain, lung, or bone. This variable was derived from the intuition that patients receiving radiation at a site different from the primary tumor are at a higher chance of having distant recurrences.

3.3 Clinical Narratives

We queried the NMEDW for clinical narratives generated before May 2016 (the start time of manual chart review) or the date when the patient was censored. All inpatient and outpatient notes were retrieved without any provider type restriction. The retrieved clinical narratives include progress notes, pathology reports, telephone encounter notes, assessment and plan notes, problem overview notes, treatment summary notes, radiology notes, lab notes, procedural notes, and nursing notes. Only notes generated after the diagnosis of breast cancer were retrieved. We only included the notes having at least one mention of "breast." After retrieving the narratives, we first preprocessed the corpus by removing duplicate copies, tokenizing sentences, and removing non-English symbols. Following these preprocessing steps, we annotated the medical concepts in the sentences using MetaMap, an NLP application to map the biomedical text to the UMLS Metathesaurus [42]. The surrounding semantic context was determined. CUIs that were tagged as negated by NegEx [43] were excluded (NegEx is a negation tool configured in MetaMap). If multiple CUIs were mapped, the one with maximum MMI score (a score ranked by MetaMap) was retained. In order to completely and accurately exclude negations or unrelated contextual cues, such as a differential diagnosis event, sentences with negative contextual features (e.g., "no," "rule out," "deny," "unremarkable") and uncertain contextual features (e.g., "risk," "concern," "worry," "evaluation") were also removed. This customized list of contextual features was obtained from the development corpus.

3.4 Feature Generation

To focus our NLP efforts, we identified a set of target distant recurrence concepts with the help of sample notes. We reviewed a development corpus of ten randomly selected samples' notes with distant recurrences and extracted partial sentences that were related to a breast cancer distant recurrence. These extracted partial sentences appear in Table S1. The initial set contains 20 partial sentences. These partial sentences were tagged by MetaMap, and the CUIs corresponding to each concept was obtained. The customized dictionary contains 83 CUIs (Table S2). After data preprocessing and

concept mapping, only CUIs with highest MMI score that also fall within the customized dictionary were used as features for model training. CUIs with MMI score smaller than one were filtered and excluded. Following this feature selection, there were 83 narrative-based features remaining for inclusion in the machine learning algorithm. In addition to the obtained CUI features, the 18 structured clinical variables described above were used as additional features.

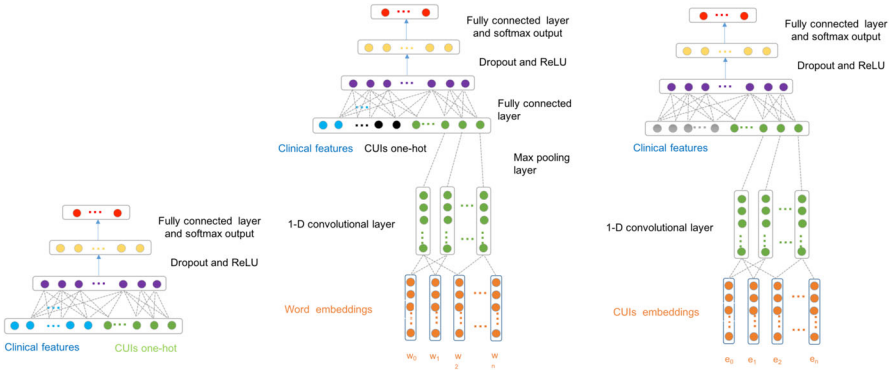
3.5 Prediction Model and Evaluation

We used support vector machine (SVM) to develop an algorithm to predict whether patients had distant recurrences. SVMs have been widely used for computational phenotyping [21, 22]. We applied linear kernel type for the SVM models. In our experiments, we trained four baseline classifiers on different feature types: a full set of medical concepts tagged by MetaMap [42], a filtered set of medical concepts tagged by MetaMap, only the structured clinical data, and a standard bag of words from clinical narratives. TfIDFVectorizer class in scikit-learn [44] was used to convert the raw documents to a matrix of TF-IDF features to assemble a bag of words. In the full MetaMap and bag of words, chi-square test was applied to select features before training the model to remove the common words that exist in clinical narratives. Only top 5% features were retained for modeling.

In the model evaluation, we chose precision, recall, macro F-measure, and area under curve (AUC) score as measurement matrix. Cross-validation performance depends on the randomly shuffled split of the training data set into multiple folds. In order to obtain robust performance statistics, each fivefold cross validation is replicated 20 times using shuffled stratified splits initialized with different random seeds.

3.6 Deep Learning Models

We also explored several deep learning models for distant recurrences prediction. As shown in Fig. 2. We used three deep learning models: (1) a Multi-Layer Perceptron (MLP) with structured clinical features and CUIs one-hot vectors as input. (2) A convolutional neural network (CNN) with clinical text, structured clinical features, and CUIs one-hot vectors. CNN is a widely used deep learning model which has been successfully applied in many text classification tasks. In this study, we used CNN to learn clinical narrative features automatically. We used all types of notes and 200 dimensional word2vec word embeddings learned from MIMIC-III clinical text as the input of the convolutional layer. (3) A CNN with CUIs embeddings and structured clinical features. We also treated CUIs sequences as words in clinical narrative text and used CNN to learn text features. We tried to add paddings between CUIs in different sentences. We used pre-trained CUIs embeddings made by De Vine et al. [45] as the input entity representations of CNN. In the three figures, w_0, w_1, w_2,..., w_n are words and e_0, e_1, e_2,..., e_n are CUIs in a record. A one-dimensional convolution layer was built on the input word embeddings or entity embeddings. We used max pooling to select the most important feature with the highest value in the convolutional feature map. We then concatenated the max pooling results of word embeddings or CUIs embeddings with structured clinical features and/or CUIs one-hot vectors. The concatenated hidden features were fed into a fully connected layer, then a dropout and ReLU activation



(a) MLP (b) Clinical features + CUIs +word CNN (c) Clinical features + CUIs CNN

Fig. 2 Deep learning architectures for distant recurrences prediction. **a** Multi-layer perceptron with clinical features and CUIs. **b** Convolutional neural network with clinical text, structured clinical features, and CUIs one-hot vectors. **c** Convolutional neural network with CUIs embeddings and structured clinical features

layer. Finally, a fully connected layer is fed to a softmax output layer, whose output is the probability distribution over labels.

We implemented our deep learning models using TensorFlow [46], a popular deep learning platform. We set the following parameters for our model: the number of convolution filters, 32; the convolution kernel size, 4; the dimension of hidden layer in the fully connected layer, 64; dropout keep probability, 0.8; learning rate, 0.001; batch size, 64; the number of learning epochs, 30. We also tried other settings of these parameters but did not find much difference. We used softmax cross entropy loss as the loss function and Adam algorithm [47] as the optimizer.

4 Experiment Results

As demonstrated in Table 3, clinical data with a significant difference between the distant recurrence group and the non-recurrence group in the double-annotated training set are presented. Compared to the non-recurrence patients, women with distant recurrences had a higher percentage of nodal positivity and higher grade of tumor, were more likely to be diagnosed with invasive ductal carcinomas, had more radiation performed at the metastasis site, had received more targeted therapies, and were more likely to die before the age of 75.

Linear kernel was applied for the SVM model (default parameters in the Python package “sklearn.svm” were used [44]). The performance of our proposed model significantly outperformed the other four baselines in the cross-validation test; the P value for Student’s t test was 0.0004 comparing our proposed model with the second-ranked model of filtered MetaMap.

We trained an SVM model on the training set (1396 samples) and then predicted labels on the held-out test set (599 samples). Comparing the predicted labels/probabilities and the annotated labels, the obtained precision, recall, F1-measure score, and AUC scores are presented in Table 4. The F1-measure score and AUC score obtained in our proposed

Table 3 Descriptive summaries of 1995 subjects' clinical data. The significance test is performed between the distant recurrence group and the non-recurrence group. Only data with *P* values less than 0.05 are presented. DR stands for distant recurrence. The mean and standard deviation are calculated for continuous variables. Numbers and percentages are presented for categorical variables. *P* values were obtained using Student's *t* test for continuous variables and chi-squared test for categorical variables

	Double-annotated set $N = 1995$	DR $N = 193$	No DR $N = 1802$	*P* value
Nodal positivity (%)	544 (27.3%)	103 (53.4%)	441 (24.5%)	1.4E-14
Histology (%)				2.6E-06
IDC	1530 (76.7%)	174 (90.2%)	1356 (75.2%)	
DCIS	279 (14.0%)	3 (1.6%)	276 (15.3%)	
ILC	155 (7.8%)	15 (7.8%)	140 (7.8%)	
Grade (%)				2.1E-10
Grade 1	458 (23.0%)	16 (8.3%)	442 (24.5%)	
Grade 2	851 (42.7%)	73 (37.8%)	778 (43.2%)	
Grade 3	665 (33.3%)	101 (52.3%)	564 (31.3%)	
Deceased (%)	157 (7.9%)	98 (50.8%)	59 (3.3%)	< 2.2E-16
Radiation (%)	67 (3.4%)	52 (26.9%)	15 (0.8%)	< 2.2E-16
Targeted therapy (%)	60 (3.0%)	44 (22.8%)	16 (0.9%)	< 2.2E-16

model were 0.78 and 0.95, respectively. The model with NLP-features, namely Filtered MetaMap, also had a notable performance with AUC of 0.93. The performance in our proposed model outperformed all the baseline models. The performance was improved when we combined the EHR data and clinical narratives as features.

In addition to our training and validation analyses, we applied our fitted model to predict labels on the generalization set, which contained 4904 single-annotated samples. In this generalization test, we obtained a precision of 0.76, a recall of 0.72, a *F*-measure of 0.74, and an AUC score of 0.93, which had a similar performance as the held-out test.

From the fitted SVM model using the 1396 samples in the training set, we retrieved the coefficient scores for each feature. The top 15 ranked coefficient scores and their corresponding variable names appear in Table 5. Three of the clinical variables (radiation, deceased, and targeted therapy) were highly ranked on the list. These three variables were treatment or outcome variables. The rest of the top-ranked features were concepts obtained from clinical narratives. Most of the CUIs were either related to

Table 4 The number of features (in parenthesis) and the precision, recall, F1-score, and AUC scores obtained in the external test using the test set (599 samples)

Model	Precision	Recall	F1-score	AUC
Filtered MetaMap + clinical data (101)	0.82	0.75	0.78	0.95
Full MetaMap (1537)	0.53	0.51	0.50	0.56
Filtered MetaMap (83)	0.72	0.62	0.67	0.93
Clinical data (18)	0.68	0.47	0.56	0.87
Bag of words (4959)	0.45	0.50	0.48	0.55

metastases events or related to the metastatic sites that breast cancer could spread to. The term "IXEMPRA" is a prescription medicine used for locally advanced breast cancer or breast cancer with distant recurrences.

Table 6 presents the distant recurrence prediction performance of deep learning models. Building a MLP on structured clinical features and filtered CUIs one-hot vectors results the best performance, with a precision of 0.81 ± 0.05, a recall of 0.78 ± 0.04, and an F1-measure score of 0.79 ± 0.02 on the 599 test set. Applying this model on the generalizability test set of 4904 samples, results in a precision of 0.77 ± 0.02, a recall of 0.70 ± 0.02, an F1-measure score of 0.74 ± 0.004, and an AUC of 0.95 ± 0.01, which again shows the effectiveness of filtered CUIs. We can also note that MLP has better predictive capabilities compared to SVM and building a MLP on structured clinical features and filtered CUIs one-hot vectors results in the best performance ($p < 0.05$ using Student's t test).

5 Discussion

In this study, we combined 83 features from unstructured clinical narratives and 18 features from structured clinical data to identify distant recurrences in breast cancer.

Table 5 Top 15 variables with their corresponding coefficients

CUI	Name	Coefficient	Partial sentences
C0153678	Secondary malignant neoplasm of pleura	1.00	Cancer metastatic to pleura metastatic cancer to pleura
Radiation	Clinical variable	0.90	
Deceased	Clinical variable	0.90	
Targeted therapy	Clinical variable	0.84	
C0153690	Secondary malignant neoplasm of bone	0.76	Metastases to bone, bone metastases
C1967552	IXEMPRA	0.71	ixempra
C0278488	Carcinoma breast stage IV	0.70	Metastatic breast cancer, breast cancer stage iv, metastatic breast carcinoma
C0494165	Secondary malignant neoplasm of liver	0.62	Liver metastases, liver metastatic disease, metastatic disease liver, metastases to the liver, liver metastases
C0220650	Metastatic malignant neoplasm to brain	0.59	Brain metastases
C1266909	Entire bony skeleton	0.39	Bone
C2939420	Metastatic neoplasm	0.27	Metastatic disease
C0036525	Metastatic to	0.25	Metastatic
C0027627	Neoplasm metastasis	0.25	Metastatic disease
C0346993	Secondary malignant neoplasm of female breast	0.23	Metastatic breast cancer to the
C1522484	Metastatic qualifier	0.22	Metastatic

Table 6 Distant recurrences prediction performances of deep learning models. We run all models ten times and report the mean plus/minus standard deviation

Model	Precision	Recall	F1-score	AUC
MLP + structured clinical features	0.6596 ± 0.0320	0.4626 ± 0.0465	0.5430 ± 0.0367	0.9250 ± 0.0071
MLP + filtered CUIs one hot	0.7526 ± 0.0263	0.7091 ± 0.0393	0.7326 ± 0.0173	0.9278 ± 0.0040
MLP + structured clinical feature + filtered CUIs one hot	0.8147 ± 0.0514	0.7782 ± 0.0418	0.7942 ± 0.0234	0.9489 ± 0.0023
CNN + filtered CUIs one-hot embeddings (no padding) + structured clinical features	0.6529 ± 0.0316	0.3782 ± 0.0839	0.4703 ± 0.0741	0.9112 ± 0.0134
CNN + filtered CUIs one-hot embeddings (padding) + structured clinical features	0.6332 ± 0.0561	0.4109 ± 0.0733	0.4925 ± 0.0535	0.9107 ± 0.0093
CNN + filtered CUIs dense embeddings (no padding) + structured clinical features	0.5939 ± 0.0509	0.3491 ± 0.0657	0.4356 ± 0.0496	0.8998 ± 0.0167
CNN + filtered CUIs dense embeddings (padding) + structured clinical features	0.6010 ± 0.0685	0.3327 ± 0.0630	0.4234 ± 0.0489	0.8849 ± 0.0167
CNN + filtered CUIs one-hot embeddings (no padding)	0.0341 ± 0.0705	0.0091 ± 0.0129	0.0140 ± 0.0191	0.4710 ± 0.0265
CNN + filtered CUIs one-hot embeddings (padding)	0.1005 ± 0.0705	0.0291 ± 0.0195	0.0443 ± 0.0291	0.5389 ± 0.0281
CNN + filtered CUIs dense embeddings (no padding)	0.0488 ± 0.0461	0.0236 ± 0.0258	0.0296 ± 0.0293	0.5258 ± 0.0244
CNN + filtered CUIs dense embeddings (padding)	0.1121 ± 0.0757	0.0327 ± 0.0254	0.0499 ± 0.0376	0.5231 ± 0.0257
Word CNN	0.7125 ± 0.0515	0.3909 ± 0.0801	0.5007 ± 0.0757	0.8953 ± 0.0145
Word CNN + filtered CUIs one-hot + structured clinical features	0.7862 ± 0.0180	0.6764 ± 0.0711	0.7255 ± 0.0455	0.9404 ± 0.0052

Clinical narratives were extracted from progress notes, pathology reports, telephone encounter notes, assessment and plan notes, problem overview notes, treatment summary notes, radiology notes, lab notes, procedural notes, and nursing notes generated after diagnosis of primary breast cancer. The clinical narratives were tagged by NLP application MetaMap to generate UMLS concepts. After filtering out concepts that were not in the customized dictionary, the remaining concepts were combined with the structured clinical data to train an SVM and deep dearning models for distant recurrence identification. We were able to identify structured clinical variables that could stratify the groups of women with and without distant recurrences. Using the SVM model, we obtained *F*-measure scores of 0.78 and 0.74, AUC scores of 0.95 and 0.93, in the test and generalizability test set, respectively. Using the deap learning model, we obtained *F*-measure scores of 0.79 ± 0.02 and 0.74 ± 0.004, AUC scores of 0.95 ± 0.002 and 0.95 ± 0.01, respectively. These results seem to suggest that convolutional neural networks and multilayer perceptron can further improve classification performance, though bearing with heavier machinery and less interpretability to some degree. Both the SVM and deep learning models achieved the best performance using filtered

MetaMap + clinical data as features. In the filtered CUIs, only the concepts that were associated with distant recurrence were retained. In a preliminary analysis, we have found that some of the clinical data features were significantly associated with distant recurrences. Since these features were independently associated with distant recurrence, when using SVM model with a linear kernel, we obtained comparable results with the deep learning models.

During the feature coefficient study in SVM model, we found that the features "secondary malignant neoplasm of pleura, radiation, deceased, targeted therapy, and secondary malignant neoplasm of bone" were the top-ranked features. Intuitively, women with distant recurrences have a higher chance of receiving radiation at the metastatic site and of receiving targeted therapy compared to those without distant recurrences. They are also more likely to have a lower survival rate. The most common sites of metastasis to organs were the bone, brain, lung, and liver [1]. In our study, we found the mentions of metastatic to bone, liver, and brain were also top ranked. The terms "metastatic" and "breast cancer" were also more likely to appear in the clinical notes of patients with distant recurrences.

In this study, modeling CUI sequences in the CNN model did not improve the model's performance. Indicating that including word orders in the text as features is not helpful. This is likely due to clinical narrative text contains noises. Excessively busy residents and senior clinicians might create notes by simply copying and pasting previous encounter notes, while making only minor updates for the most recent appointment. This results in a high similarity between notes, even though they contain different important information. The same applies to the full set of MetaMap concepts, which is similar to the bag of words. In the model using only filtered words, where only highly associated words were retained, we have obtained better results.

For the baseline models of full MetaMap and bag of words, we have applied chi-square test to select features before training the model. Only the top 5% features were retained for full MetaMap and bag of words modeling. This test might have the potential for overfitting in cross validations (AUC = 0.78, SD = 0.04 for full MetaMap, AUC = 0.82, SD = 0.024 for bag of words). Indeed, we saw a lower performance in the held-out test for full MetaMap and bag of words in the SVM model. To adjust this problem, we tested different thresholds for the chi-square test selection. However, we found 5% ended with the best results.

Identifying breast cancer distant recurrence in clinical data sets is important for clinical research and practice. Annotation of distant recurrence is difficult using standard EHR phenotyping approaches and is commonly beyond the scope of manual annotation efforts by cancer registries. A model using natural language processing, EHR data, and machine learning to identify distant recurrences in breast cancer patients allows more accurate data mining and significantly less time-consuming manual chart review. We expect that by minimally adapting the positive concept set, this study has the potential to be replicated at other institutions with a moderately sized training data set. In this study, we generated features using sentences extracted from the clinical narratives combined with structured data. The training and testing data sets were cross annotated in the process, which offered a solid ground truth for the study. Replicating this model requires minimal outside effort. We offered the customized dictionary in this study, so a user can retrieve the required notes and clinical structured data in order to replicate this study. After the rigorous manual chart review and feature retrieval, our

data set has offered a gold-standard data set with rich, validated information for further breast cancer research.

When replicating this study at another institution, there is a chance that one will not be able to find the structured clinical data in their databases. If this is the case, some of the structured data can be found from other resources. Variables of "histology" and "lymph node status" can be extracted from pathology reports using a rule-based system. For example, expressions of "total lymph nodes," "total lymph nodes number positive," "axillary lymph nodes examined," "axillary lymph nodes examined number of positive versus total" can be used to extract lymph node status from pathology report at our institution. Survival information can be found in the administrative billing system.

6 Future Work

The NLP pipeline cannot characterize the context of features. Clinical narratives contain patients' concerns, clinicians' assumptions, and patients' past medical histories. Clinicians also record diagnoses that are ruled out or symptoms that patients denied. Our next aim will be that such conditions, mentions, and feature relations will be extracted to better distinguish differential diagnoses. Generalized relation and event extraction, rather than binary relation classification, will be conducted. To this end, graph methods are a promising class of algorithms and should be actively investigated [48, 49].

In this study, we have chosen SVM model with linear kernel for interpretation convenience. On the other hand, we experimented with multiple convolutional neural networks and multilayer perceptrons and found that they may result in better performance. In the future, we will test our data with more machine learning models in order to identify the sweet spot between better performance and more interpretability.

We will also aim to address the heterogeneity problem in clinical narratives. It is a common problem in clinical narratives due to the variance in physicians' expertise and behaviors [50]. Features derived from clinical narratives included in this study were extracted from notes generated by clinicians with different specialties and professional levels of expertise. As a result, some content was not relevant to the breast cancer distant recurrence event, even though we had limited the notes to include the mention of "breast." For example, a liver cancer metastasis to the breast from a primary liver tumor would be difficult to identify. We will need to resolve the heterogeneity in clinical narratives.

7 Conclusions

We developed multiple machine learning models by combining structured clinical data and unstructured clinical narratives in order to identify distant recurrence events in breast cancer. Our model choices include SVM with linear kernels that are easy to interpret, as well as convolutional neural networks and multilayer perceptrons that are more accurate. We demonstrated the high accuracy and efficiency of our models, using cross validation, held-out test evaluation, and a further generalization set evaluation. Our proposed models allow for more accurate and efficient identification of distant

recurrences than single modality models using either clinical narratives or structured clinical data. Thus, our models offer a significantly less time-consuming and practical alternative to manual chart review. This is particularly relevant in an era when evidence-based medicine receives growing attention and there is more emphasis on computational phenotyping and data-driven discovery. This model would also be valuable and applicable to research in other medical fields beyond breast cancer.

Funding Information This project is supported in part by NIH grant R21LM012618-01.

References

1. Egner JR (2010) AJCC cancer staging manual. JAMA 304(15):1726–1727
2. Lê MG, Arriagada R, Spielmann M, Guinebretière JM, Rochard F (2002) Prognostic factors for death after an isolated local recurrence in patients with early-stage breast carcinoma. Cancer 94(11):2813–2820
3. Geiger AM, Thwin SS, Lash TL, Buist DSM, Prout MN, Wei F, Field TS, Ulcickas Yood M, Frost FJ, Enger SM, Silliman RA (2007) Recurrences and second primary breast cancers in older women with initial early-stage disease. Cancer 109(5):966–974
4. Habel LA, Achacoso NS, Haque R, Nekhlyudov L, Fletcher SW, Schnitt SJ, Collins LC, Geiger AM, Puligandla B, Acton L, Quesenberry CP (2009) Declining recurrence among ductal carcinoma in situ patients treated with breast-conserving surgery in the community setting. Breast Cancer Res 11(6):R85
5. Starren JB, Winter AQ, Lloyd-Jones DM (2015) Enabling a learning health system through a unified enterprise data warehouse: the experience of the Northwestern University Clinical and Translational Sciences (NUCATS) Institute. Clin Transl Sci 8(4):269–271
6. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF (2005) Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care 43(5):480–485
7. Singh JA, Holmgren AR, Noorbaloochi S (2004) Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. Arthritis Care Res 51(6):952–957
8. O'malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM (2005) Measuring diagnoses: ICD code accuracy. Health Serv Res 40(5p2):1620–1639
9. Hripcsak G, Albers DJ (2012) Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 20(1):117–121
10. Greenhalgh T (1999) Narrative based medicine: narrative based medicine in an evidence based world. BMJ Br Med J 318(7179):323–325
11. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM (2010) Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res 62(8):1120–1127
12. G. Chao and S. Sun, "Applying a multitask feature sparsity method for the classification of semantic relations between nominals," in *Machine Learning and Cybernetics (ICMLC),* 2012 *International Conference on*, 2012, vol. 1, pp. 72–76: IEEE
13. Luo Y et al (2017) Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf:1–15
14. Zeng Z, Deng Y, Li X, Naumann T, Luo Y (2018) Natural language processing for EHR-based computational phenotyping. IEEE/ACM Transactions on Computational Biology and Bioinformatics: 1–1
15. D. S. Carrell*, S. Halgrim, D.T. Tran, D. S. M. Buist, J. Chubak, W. W. Chapman, G. Savova*, "Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence," *American journal of epidemiology,* p. kwt441, 2014, 179, 749, 758
16. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP (2013) Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Inform Assoc 20(2):349–355
17. Bosco JL et al (2009) Breast cancer recurrence in older women five to ten years after diagnosis. Cancer Epidemiology and Prevention Biomarkers 18(11):2979–2983
18. Haque R, Shi J, Schottinger JE, Ahmed SA, Chung J, Avila C, Lee VS, Cheetham TC, Habel LA, Fletcher SW, Kwan ML (2015) A hybrid approach to identify subsequent breast cancer using pathology and automated health information data. Med Care 53(4):380–385

19. Wallner LP, Dibello JR, Li BH, Zheng C, Yu W, Weinmann S, Richert-Boe KE, Ritzwoller DP, VanDenEeden SK, Jacobsen SJ (2014) Development of an algorithm to identify metastatic prostate cancer in electronic medical records using natural language processing. Proc Am Soc Clin Oncol 32:164
20. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. Mach Learn ECML-98:137–142
21. Garla V, Taylor C, Brandt C (2013) Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 46(5):869–875
22. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M (2012) Pneumonia identification using statistical feature selection. J Am Med Inform Assoc 19(5):817–823
23. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, Fry MJ (2007) Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 14(6):736–745
24. Z. Zeng *et al.*, "Contralateral breast cancer event detection using Nature Language Processing," in *AMIA Annual Symposium Proceedings*, 2017, vol. 2017, pp. 1885–1892: American Medical Informatics Association
25. R. J. Carroll, A. E. Eyler, and J. C. Denny, "Naïve electronic health record phenotype identification for rheumatoid arthritis," in *AMIA annual symposium proceedings*, 2011, vol. 2011, p. 189: American Medical Informatics Association
26. Denny JC, Smithers JD, Miller RA, Spickard A III (2003) "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc 10(4):351–362
27. Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882,* 2014
28. N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188,* 2014
29. K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075,* 2015
30. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489
31. S. Gehrmann *et al.*, "Comparing Rule-Based and Deep Learning Models for Patient Phenotyping," *arXiv preprint arXiv:1703.08705,* 2017
32. Luo Y (2017) Recurrent neural networks for classifying relations in clinical notes. J Biomed Inform 72: 85–95
33. Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J (2017) Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc 25(1):93–98
34. Wu Y, Jiang M, Lei J, Xu H (2015) Named entity recognition in Chinese clinical text using deep neural network. Studies in health technology and informatics 216:624
35. A. N. Jagannatha and H. Yu, "Structured prediction models for RNN based sequence labeling in clinical text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing Conference on Empirical Methods in Natural Language Processing*, 2016, vol. 2016, p. 856: NIH Public Access
36. A. N. Jagannatha and H. Yu, "Bidirectional rnn for medical event detection in electronic health records," in *Proceedings of the conference Association for Computational Linguistics North American Chapter Meeting*, 2016, vol. 2016, p. 473: NIH Public Access
37. DeLisle S, Kim B, Deepak J, Siddiqui T, Gundlapalli A, Samore M, D'Avolio L (2013) Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy. PLoS One 8(8):e70944
38. Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, Braggs NS, Cagan A, Gainer V, Denny JC, Savova GK (2014) Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. J Am Med Inform Assoc 22(e1):e151–e161
39. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P, Churchill S, Kohane I (2015) Development of phenotype algorithms using electronic medical records and incorporating natural language processing. bmj 350:h1885
40. F. Galton, Finger prints. Macmillan and Company, 1892
41. Leemans CR, Tiwari R, Nauta J, Van der Waal I, Snow GB (1993) Regional lymph node involvement and its significance in the development of distant metastases in head and neck carcinoma. Cancer 71(2): 452–456
42. A. R. Aronson, "Metamap: mapping text to the umls metathesaurus," Bethesda, MD: NLM, NIH, DHHS, pp. 1–26, 2006

43. Chapman WW et al (2013) Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 192:677
44. Pedregosa F et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
45. L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, "Medical semantic similarity with a neural language model," in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 2014, pp. 1819–1822: ACM
46. M. Abadi *et al*, "Tensorflow: a system for large-scale machine learning," in *OSDI*, 2016, vol. 16, pp. 265–283
47. D. Kinga and J. B. Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015, vol. 5
48. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P (2015) Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc:ocv016
49. Luo Y, Sohani AR, Hochberg EP, Szolovits P (2014) Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc 21(5):824–832
50. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C (2013) Defining a comprehensive verotype using electronic health records for personalized medicine. J Am Med Inform Assoc 20:e232–e238

Affiliations

Zexian Zeng[1] · Liang Yao[1] · Ankita Roy[2] · Xiaoyu Li[3] · Sasa Espino[2] · Susan E Clare[2] · Seema A Khan[2] · Yuan Luo[1]

Zexian Zeng
zexian.zeng@northwestern.edu

Liang Yao
liang.yao@northwestern.edu

Ankita Roy
ankita.roy@northwestern.edu

Xiaoyu Li
xil288@mail.harvard.edu

Sasa Espino
sasa-grae.espino@northwestern.edu

Susan E Clare
susan.clare@northwestern.edu

Seema A Khan
s-khan2@northwestern.edu

[1] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[2] Department of Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[3] Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA