# Contralateral Breast Cancer Event Detection Using Nature Language Processing

**Zexian Zeng[1], Xiaoyu Li[2], Sasa Espino[3], Ankita Roy[3], Kristen Kitsch[3], Susan Clare[3], Seema Khan[3], Yuan Luo[1]***

**[1]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [2]Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [3]Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; *Corresponding author**

**Abstract**

*To facilitate the identification of contralateral breast cancer events for large cohort study, we proposed and implemented a new method based on features extracted from narrative text in progress notes and features from numbers of pathology reports for each side of breast cancer. Our method collects medical concepts and their combinations to detect contralateral events in progress notes. In addition, the numbers of pathology reports generated for either left or right side of breast cancer were derived as additional features. We experimented with support vector machine using the derived features to detect contralateral events. In the cross-validation and held-out tests, the area under curve score is 0.93 and 0.89 respectively. This method can be replicated due to the simplicity of feature generation.*

**Introduction**

Contralateral breast cancer is defined as a solid tumor developed in the opposite breast after the detection of the first primary breast cancer. Woman with a first primary breast cancer has two to six folds of increased risk to develop a contralateral breast cancer compared to the normal population[1]. Understanding the etiology of contralateral breast cancer can not only help us understand the risks associated with breast cancer development, but also help monitor the effects of treatments[2]. Efforts have been devoted to study the shared risk factors between the first and second primary breast cancer, including family history[3], environmental exposures[4], and genetic mutations[5]. These studies require us to identify the group of patients with contralateral breast cancer accurately. The prevalence of Electronic Health Records (EHR) has enabled large cohort study for different clinical problems, including the contralateral breast cancer. The abundant available information in EHR makes deep phenotyping in large cohort studies more achievable. However, in most cases, identifying contralateral events are still based on manual chart review, which is time consuming and labor intensive.

Because contralateral event is a progressive event, a patient may have been associated with risks of developing such an event for an extended period along his/her life time. The amount of work to capture and maintain pathophysiologic data along the development of risk factors and to identify new events is not trivial. On the other hand, the patient's progressive information and clinical status are well recorded in the progress notes during the course of a hospitalization or over the course of outpatient care. In addition, the progress notes are readily and prevalently available. Moreover, in most cases, every diagnostic procedure of breast cancer generates at least one pathology report. Usually, if a patient has bilateral breast cancer, the patient should have at least one pathology report generated for each side.

In this study, we proposed a new method for detecting contralateral breast cancer using the narrative text in progress notes and the numbers of pathology reports generated for each side of the breast. With such a model, users can identify the group of patients with contralateral breast cancer among a large cohort efficiently.

**Related Work**

Capturing contralateral breast cancer events is one of the major tasks for the tumor registries. However, many of the registries, including National Cancer Institute's Surveillance, did not successfully capture the contralateral events[6]. Studies heavily relied on manual chart review, which is both time consuming and labor intensive, thus not feasible for large cohort study[7]. Automated methods have been proposed to extract breast contralateral and recurrence events[8-10]. However, these studies did not distinguish breast cancer recurrence with contralateral breast cancer events, which significantly limited further cohort studies. Strauss et al. used the morphology codes and anatomic sites to detect contralateral breast cancer events[11]. However, the work required that the pathology reports are well documented in standard formats, which in reality is not true and requires special care from Natural Language Processing systems to unify the cross-institutional variations in pathology reports[12,13]. In addition, defining rules to retrieve information from

the copious pathology reports can be labor intensive. Furthermore, if the report did not state which side of breast was examined, the rule based system will have difficulty in calling a contralateral event. Efforts have been also devoted to apply claims data for contralateral event detection[14,15]. However, claims data are believed to have limited validity for inferring cancer recurrence events[16].

Motivated by the limitation from previous studies, we proposed a method to extract features from the common narrative text in progress notes, together with the numbers of pathology reports for each side of breast cancer, to detect contralateral breast cancer events. We experimented with Support Vector Machine (SVM) and quantitatively assessed the probability of a breast contralateral event.

**Study Cohort**

The Northwestern Medicine Enterprise Data Warehouse (NMEDW) is a joint initiative across the Northwestern University Feinberg School of Medicine and Northwestern Memorial HealthCare[17]. The Lynn Sage database in NWEDW was searched for women who underwent breast conservation surgery for Ductal Carcinoma in Situ (DCIS) or primary invasive breast cancer. We identified 1063 women who underwent breast conservation surgery for a new diagnosis of stage 0 to stage 3 breast cancer. Three co-authors (SE, AR, KK) performed chart review for these patients, and identified 33 contralateral events among these 1063 women. Study procedures were approved by the hospital's Institutional Review Board (IRB).

**Method**

We first randomly split the 1063 subjects into a training set and a held-out test set according to a 7:3 ratio. In the training dataset, progress notes from 15 women with contralateral breast cancer were extracted and reviewed. The sentences or partial sentences indicating the occurrence of contralateral breast cancer and cancer diagnoses related events were retrieved and summarized in Table 1. These partial sentences were then tagged by MetaMap, which is a nature language processing (NLP) tool to map the biomedical text to the Unified Medical Language System (UMLS) Metathesaurus[18]. The concept unique identifier (CUI) corresponding to each concept is obtained by parsing the MetaMap outputs. To reduce the noise, CUIs that are not related to breast cancer event is manually filtered and discarded, such as the CUIs of 'with', 'has', 'seen', and etc. were filtered. After filtering the CUIs, 42 CUIs were retained and these CUIs together represent the descriptions for contralateral breast cancer events. We refer to these 42 CUIs as a positive CUI dictionary. These 42 CUIs appear in Appendix A.

After obtaining the positive CUI dictionary, a number of pre-processing steps were performed on the progress notes. Such as removing duplicate copies, dividing the notes to sentences, and removing non-English symbols. Negation and uncertain sentences containing the words of 'no', 'risk', 'concern', 'worry', 'unremarkable', 'rule out', 'deny', 'evaluation', and their different inflections (e.g., tenses of verbs), were excluded. Following these pre-processing steps, the remaining sentences were tagged using MetaMap. Once we got the MetaMap output, the CUIs with negations were excluded. In addition, those CUIs that are not in the positive dictionary were excluded. The retained CUIs were used as features in our model. However, using single CUI as feature may not be informative enough for us to detect some of the contralateral events. For example, the sentence "Patient was first seen for right breast cancer who now has new left breast dcis." If we look at each individual CUI, we won't be able to conclude the contralateral event. We need the CUIs of 'right breast cancer' and 'new left breast dcis' to reach the conclusion. To this end, a complete combination of the CUIs in the same sentence would be able to help us discriminate contralateral events. Following this observation, additional features were generated by combining CUIs that were in the same sentence. In the above example, the feature combinations of (left, right), (new, left), (left, right, breast cancer) etc. were generated. Clearly, the feature of (left, right, breast cancer) offered clues for contralateral events. Using the sentence "Patient was first seen for right breast cancer who now has new left breast dcis" as example. CUIs 'C0007124' (Non-infiltrating Intraductal Carcinoma), 'C0006142' (Malignant neoplasm of breast), 'C0444532' (Right sided), 'C0222601' (Left breast), 'C0205314' (New) were generated. With a complete combination of these CUIs, we obtained 31 new features. For example, one of the new feature is {'C0444532' (Right sided); 'C0205314' (New); 'C0222601' (Left breast); 'C0007124' (Non-infiltrating Intraductal Carcinoma)} and we can use it to infer breast contralateral event.

Following the pre-processing of the progress notes, additional features were derived based on the number of pathology reports. For each sample, the numbers of pathology reports generated for left and right breast cancer were separately counted and used as two additional features. Intuitively, if a patient has contralateral breast cancer event, then the patient should have at least one pathology report for each side. One additional binary feature indicating whether the patient has pathology reports for both sides were derived. Ideally, every patient with contralateral event should have

pathology reports for both sides. In our experiment, checking whether the word 'left' or 'right' is contained in the report, we were able to derive such features.

These generated features were used to train a support vector machine (SVM) model for further contralateral event detection. We chose SVM because of its widely-acknowledged generalizability. To obtain a reasonable feature sample ratio and remove the redundant features, Chi-square test was applied to select features before training the model. Only top 50% features were retained for subsequent modeling.

**Table 1.** Sentences or partial sentences indicating the occurrence of contralateral events or cancer diagnoses related events

| |
|---|
| New idc in r breast. |
| Newly diagnosed contralateral ilc. |
| Now with new primary on the right breast. |
| Recently with contralateral ilc. |
| Was first seen for right breast cancer who now has new left breast dcis. |
| Right breast infiltrating ductal carcinoma stage i and left breast ductal carcinoma in situ. |
| Presents for a new right breast cancer. |
| Newly diagnosed right breast carcinoma. |
| Right breast concerning for breast cancer. |
| History of bilateral breast cancer. |
| A second primary was diagnosed in the contralateral breast. |
| Bilateral breast cancer. |
| Bilateral breast cancer with infiltrating carcinoma of the left breast. |
| With bilat breast cancer. |
| The patient later presented with a contralateral breast cancer on the right side. |
| With contralateral breast cancer on the right side. |
| Developed a contralateral breast cancer on the right side. |
| With a history of bilateral dcis. |
| Lumpectomy with radiation on both sides. |
| Had contralateral bc. |
| Had contralateral dcis. |

Five-fold cross-validation was applied on the training dataset to tune parameters for the model, which were then evaluated on the held-out test data. In our experiments, we trained four baseline classifiers on different feature types. Baseline 1 is the proposed model without the additional information from the number of pathology reports, referred to as *combined MetaMap*. Baseline 2 uses only the numbers of pathology reports, referred to as *pathology report count*. Baseline 3 uses only concepts in the positive dictionary without their combination, referred as *Positive Dictionary without Combination*. Baseline 4 uses bag of words as features, referred as *Bag of Words*. To generate the features for *bag of words*, TfIDFVectorizer class in scikit-learn was used to convert the raw documents to a matrix of TF-IDF features.

An overview of the workflow employed in this study is shown in Figure 1.

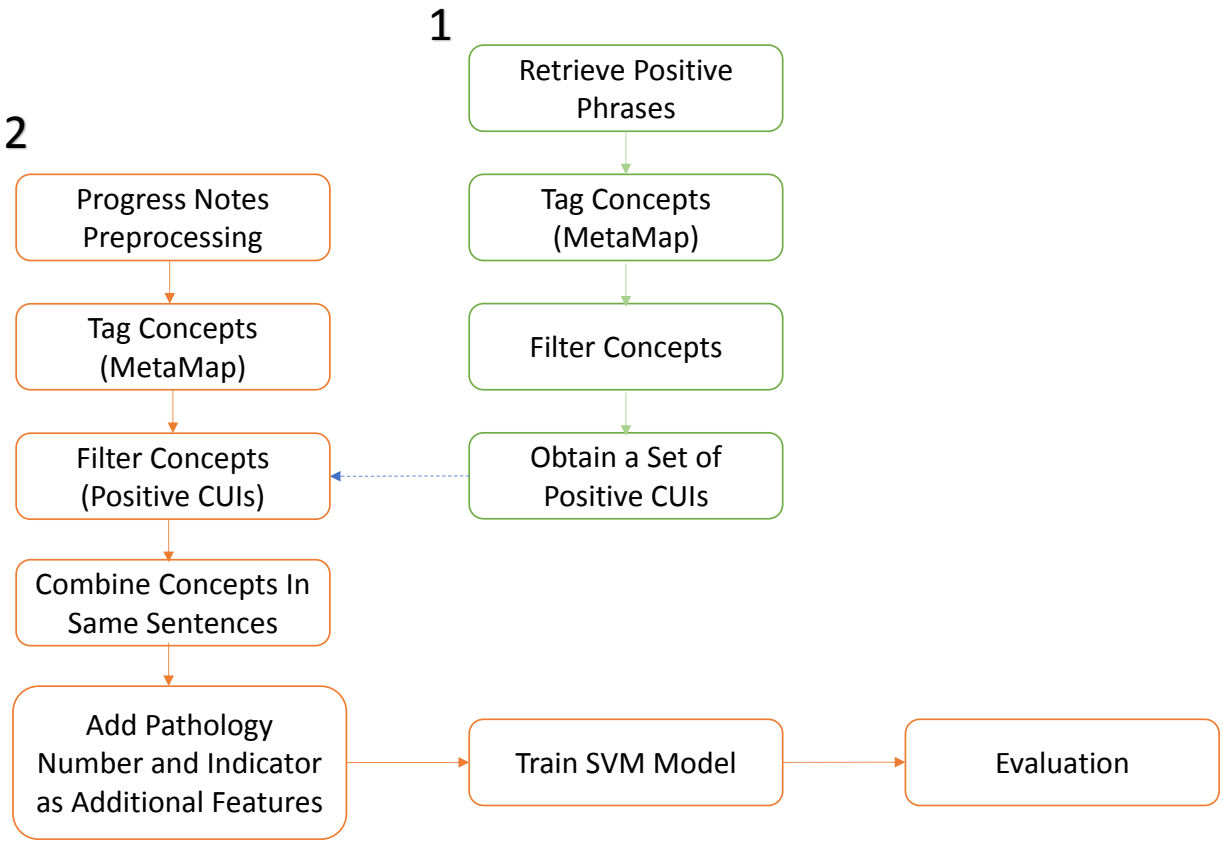


**Figure 1.** Overview of the processes employed in this study.

**Results**

Among the 1063 subjects, the average numbers of pathology reports are 3.79 with 95% Confidence Interval of ± 0.19 for left side and 3.27 with 95% Confidence Interval of ± 0.19 for right side. If a subject has pathology reports for both sides, the new feature was labeled as 1. In the training data, among the 21 subjects with contralateral breast cancer, 21 (100%) have pathology reports on both sides. Among the 724 subjects without contralateral event, 259 (35.77%) subjects have pathology reports generated for both sides.

In total, 1282 features were generated in the baseline 1 of *combined MetaMap*, which used all information from progress notes but not from pathology reports. Three features were generated in baseline 2 by using the three features derived from the numbers of pathology reports. A total of 42 features were used in baseline 3 by using positive dictionary without combination. In baseline 4, we used bag of words and 55192 features were generated.

Table 2 shows the feature numbers and cross-validation area under curve (AUC) scores of our proposed model in comparison with the other four baselines. To account for performance variability due to different split of folds, the five-fold cross-validation was repeated 10 times and the standard deviation was obtained. It is clear that *combined MetaMap* outperforms the *positive dictionary without combination* and also the *bag of words*. We compared proposed model with *combined MetaMap* using Student's t-test ($\alpha$=0.05). The difference is statistically significant with $P_{value} = 0.00027$. We see improvements on AUC score in our proposed model compared to all other four baselines. In the cross-validation, the AUC score of our proposed model is 0.93 with standard deviation equals 0.02.

Table 3 shows the feature number and AUC scores on the model for prediction in comparison with the four other methods in held-out test. The AUC score is 0.89, which outperforms all four baseline methods by a large margin.

Using the trained model upon the training set, we obtained the coefficient for each feature. The top ranked six features appeared in Table 4.

**Table 2.** Cross-validation results using different methods. Standard deviation (SD) is included in the parenthesis.

| Model | Feature Number | AUC (SD) |
|---|---|---|
| Combined MetaMap +Pathology Report Count | 1285 | **0.93 (0.02)** |
| Combined MetaMap | 1282 | 0.82 (0.07) |
| Pathology Report Count | 3 | 0.75 (0.07) |
| Positive Dictionary without Combination | 42 | 0.46 (0.05) |
| Bag of Words | 55192 | 0.66 (0.06) |

**Table 3.** Held-out test results using different methods.

| Model | Feature Number | AUC |
|---|---|---|
| Combined MetaMap +Pathology Report Count | 1285 | 0.89 |
| Combined MetaMap | 1282 | 0.68 |
| Pathology Report Count | 3 | 0.67 |
| Full MetaMap without Combination | 42 | 0.30 |
| Bag of Words | 55192 | 0.70 |

**Table 4.** Top ranked features in a coefficient study. The descriptions in parenthesis right after CUIS are UMLS concept preferred names.

| Features | Coefficient | Feature descriptions |
|---|---|---|
| {C0007097 (Carcinoma); C0449450 (Presentation)} | 0.556 | {Carcinoma; Presentation} |
| Pathology Report for Both Side Indicator | 0.374 | It is an indicator whether the patient has pathology reports generated for both sides |
| {C0205314 (New); C0222600 (Right breast)} | 0.278 | {New; Right breast} |
| {C0006141(Breast); C0007124 (Noninfiltrating Intraductal Carcinoma); C0007124 (Noninfiltrating Intraductal Carcinoma)} | 0.256 | {Breast; Noninfiltrating Intraductal Carcinoma; Noninfiltrating Intraductal Carcinoma} |
| {C0007124 (Noninfiltrating Intraductal Carcinoma); C0007124 (Noninfiltrating Intraductal Carcinoma); C1268990 (Entire breast)} | 0.256 | {Noninfiltrating Intraductal Carcinoma; Noninfiltrating Intraductal Carcinoma; Entire breast} |
| C0281267 (Bilateral breast cancer) | 0.246 | Bilateral breast cancer |

**Discussion**

In this study of detecting contralateral breast cancer events from progress notes and counts of pathology reports, the AUC score for our proposed model is 0.93 (±0.02) in cross-validation and is 0.89 for held-out test. The model is able to retrieve contralateral breast cancer events by using the combination of narrative text in progress notes and the additional features derived from the numbers of pathology reports. The proposed model outperformed all four baseline methods, demonstrating that different features offer different levels of information. For the *combined MetaMap* feature

only, the AUC is low because some progress notes do not necessarily contain the concepts that exist in the positive CUI dictionary. On the contrary, all patients with contralateral event has pathology reports for both sides of breast cancer. In addition, many patients without contralateral breast cancer event also have pathology reports for both sides, indicating that this feature will help improve recall but may lower precision. Putting these two types of features together has the potential to address the limitations of each other and increase the chance of identifying contralateral events. It is acknowledged that we have seen some jargons in the progress notes. However, considering that one patient's progress notes are often written by multiple clinicians, we still have a high chance to find well-formatted sentences in the progress notes. In the positive feature study, the derived variable 'Pathology Report for Both Side Indicator' ranked as second feature, indicating that the variable we have created is an efficient one. In addition, the top ranked features hinted us a story that if new carcinoma presents in right or left breast, or if Noninfiltrating Intraductal Carcinoma presents twice in one sentence together with breast or entire breast, the patient then have a high chance to have contralateral breast cancer. Obviously, 'bilateral breast cancer' is another indicator to be used to find the events. These semi-structured features may provide alternative perspectives that are useful in capturing contralateral recurrence. In the future, we plan to identify more semi-structured features to complement CUI-based features, where tensor modeling may provide useful tools for integrating different types of clinical features[19].

In an error analysis, one of the patient with contralateral event was not identified because of concept positions in the power set. In the patient's progress note, one sentence appears as: "this is a 61-year-old woman with right breast cancer newly diagnosed". The power set derived was {right side; breast cancer; newly diagnosed} However, in the model we have trained, we only have a position-sensitive feature of {newly diagnosed; right side; breast cancer}. The derived power set was not recognized and the event wat not identified. In the future, instead of using position-sensitive power set, we plan to use graph based representation to capture the relations between medical concepts (CUIs) with more accuracy[20,21]. In another case, we saw: "including stage 2 l breast ca, dcis r breast" in the progress notes. However, the 'left' and 'right' are both abbreviated to 'l' and 'r'. Our model is not yet complicated enough to recognize these abbreviations.

**Conclusion**

Using self-defined rule-based system, one can possibly identify the numbers of pathology reports for each side of breast cancer as supplementary features. We expect this study to generalize well across medical institutions. The easiness of replication can reduce the time-consuming manual effort to identify contralateral breast cancer events for cancer registries. Moreover, instead of binary classification, this model can provide the abstractors with the continuous probability score as confidence. This study can also be applied to retrieve other breast cancer events such as local recurrence, distant recurrence as long as the positive CUI dictionary is defined.

**References**


1. Hankey BF, Curtis RE, Naughton MD, Boice JD, Flannery JT. A retrospective cohort analysis of second breast cancer risk for primary breast cancer patients with an assessment of the effect of radiation therapy. *Journal of the National Cancer Institute.* 1983;70(5):797-804.
2. Thompson WD. Methodologic perspectives on the study of multiple primary cancers. *The Yale journal of biology and medicine.* 1986;59(5):505.
3. HORN PL, THOMPSON WD. Risk of contralateral breast cancer: associations with factors related to initial breast cancer. *American journal of epidemiology.* 1988;128(2):309-323.
4. Prior P, Waterhouse J. Incidence of bilateral tumours in a population-based series of breast-cancer patients. I. Two approaches to an epidemiological analysis. *British journal of cancer.* 1978;37(4):620.
5. Bernstein JL, Thompson WD, Risch N, Holford TR. The genetic epidemiology of second primary breast cancer. *American journal of epidemiology.* 1992;136(8):937-948.
6. Bosco JL, Lash TL, Prout MN, et al. Breast cancer recurrence in older women five to ten years after diagnosis. *Cancer Epidemiology and Prevention Biomarkers.* 2009;18(11):2979-2983.
7. Enger SM, Thwin SS, Buist DS, et al. Breast cancer treatment of older women in integrated health care settings. *Journal of Clinical Oncology.* 2006;24(27):4377-4383.
8. Gao H, Bowles EJA, Carrell D, Buist DS. Using natural language processing to extract mammographic findings. *Journal of biomedical informatics.* 2015;54:77-84.
9. Carrell DS, Halgrim S, Tran D-T, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *American journal of epidemiology.* 2014:kwt441.
10. Haque R, Shi J, Schottinger JE, et al. A hybrid approach to identify subsequent breast cancer using pathology and automated health information data. *Medical care.* 2015;53(4):380-385.

11. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *Journal of the American Medical Informatics Association.* 2013;20(2):349-355.
12. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. *J Am Med Inform Assoc.* 2014;21(5):824-832.
13. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. *J Am Med Inform Assoc.* 2015:ocv016.
14. Lamont EB, Herndon JE, Weeks JC, et al. Measuring disease-free survival and cancer relapse using Medicare claims from CALGB breast cancer trial participants (companion to 9344). *Journal of the National Cancer Institute.* 2006;98(18):1335-1338.
15. Chubak J, Yu O, Pocobelli G, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *Journal of the National Cancer Institute.* 2012;104(12):931-940.
16. Chawla N, Yabroff KR, Mariotto A, McNeel TS, Schrag D, Warren JL. Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. *Annals of epidemiology.* 2014;24(9):666-672. e662.
17. Starren JB, Winter AQ, Lloyd‐Jones DM. 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.* 2015;8(4):269-271.
18. Aronson AR. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS.* 2006:1-26.
19. Luo Y, Wang F, Szolovits P. Tensor factorization toward precision medicine. *Briefings in Bioinformatics.* 2016.
20. Luo Y, Riedlinger G, Szolovits P. Text mining in cancer gene and pathway prioritization. *Cancer Inform.* 2014(Suppl. 1):69.
21. Luo Y, Uzuner Ö, Szolovits P. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics.* 2016.

**Appendix A**

Table A1: The CUIs identified in the positive dictionary

| CUIS | CUI Preferred Name |
|---|---|
| C0006041 | Botswana |
| C0006141 | Breast |
| C0006142 | Malignant neoplasm of breast |
| C0006826 | Malignant Neoplasms |
| C0007097 | Carcinoma |
| C0007124 | Noninfiltrating Intraductal Carcinoma |
| C0011900 | Diagnosis |
| C0019665 | Historical aspects qualifier |
| C0021367 | Mammary Ductal Carcinoma |
| C0205090 | Right |
| C0205091 | Left |
| C0205225 | Primary |
| C0222600 | Right breast |
| C0222601 | Left breast |
| C0238767 | Bilateral |
| C0281267 | bilateral breast cancer |
| C0439612 | True primary (qualifier value) |
| C0439631 | Primary operation |
| C0441988 | Contralateral |
| C0443246 | Left sided |
| C0205314 | New |
| C0449450 | Presentation |
| C0567470 | Breast present |
| C0678222 | Breast Carcinoma |
| C0684010 | Rabbi |
| C0750546 | Newly |
| C0853879 | Invasive carcinoma of breast |
| C0998265 | Cancer Genus |
| C1096616 | Contralateral breast cancer |
| C1134719 | Invasive Ductal Breast Carcinoma |
| C1268990 | Entire breast |
| C1306459 | Primary malignant neoplasm |
| C1366566 | CCL27 gene |
| C1449563 | Cardiomyopathy, Familial Idiopathic |
| C1527349 | Ductal Breast Carcinoma |
| C1552822 | Table Cell Horizontal Align - left |
| C1705078 | CCL27 wt Allele |
| C1997028 | History of malignant neoplasm of breast |
| C2603358 | R prime |
| C2984916 | Best Case Imputation Technique |
| C0444532 | Right sided |
| C1387407 | Personal history of primary malignant neoplasm of breast |