Application of Text Information Extraction System for Real-Time Cancer Case Identification in an Integrated Healthcare Organization

Fagen Xie¹, Janet Lee¹, Corrine E. Munoz-Plaza¹, Erin E. Hahn¹, Wansu Chen¹
¹Department of Research and Evaluation, Kaiser Permanente Southern California Medical Group, Pasadena, CA, USA

Received: 28 July 2017   Accepted: 1 November 2017    Published: 14 December 2017

Abstract

Background: Surgical pathology reports (SPR) contain rich clinical diagnosis information. The text information extraction system (TIES) is an end-to-end application leveraging natural language processing technologies and focused on the processing of pathology and/or radiology reports. Methods: We deployed the TIES system and integrated SPRs into the TIES system on a daily basis at Kaiser Permanente Southern California. The breast cancer cases diagnosed in December 2013 from the Cancer Registry (CANREG) were used to validate the performance of the TIES system. The National Cancer Institute Metathesaurus (NCIM) concept terms and codes to describe breast cancer were identified through the Unified Medical Language System Terminology Service (UTS) application. The identified NCIM codes were used to search for the coded SPRs in the back-end datastore directly. The identified cases were then compared with the breast cancer patients pulled from CANREG. Results: A total of 437 breast cancer concept terms and 14 combinations of “breast” and “cancer” terms were identified from the UTS application. A total of 249 breast cancer cases diagnosed in December 2013 was pulled from CANREG. Out of these 249 cases, 241 were successfully identified by the TIES system from a total of 457 reports. The TIES system also identified an additional 277 cases that were not part of the validation sample. Out of the 277 cases, 11% were determined as highly likely to be cases after manual examinations, and 86% were in CANREG but were diagnosed in months other than December of 2013. Conclusions: The study demonstrated that the TIES system can effectively identify potential breast cancer cases in our care setting. Identified potential cases can be easily confirmed by reviewing the corresponding annotated reports through the front-end visualization interface. The TIES system is a great tool for identifying potential various cancer cases in a timely manner and on a regular basis in support of clinical research studies.

Keywords: Breast cancer, case identification, natural language processing, pathology reports, text information extraction system

INTRODUCTION

Recent innovations in computer technology have resulted in the exponential expansion of electronic information in various industries, including the rapid growth of electronic medical records (EMR) in health-care systems.[1,2] EMR systems capture and store patient health information in structured and unstructured formats electronically in place of paper charts. The resulting structured healthcare data have been extensively used to support clinical operations, decision-making, and biomedical research.[3] However, clinical narrative captured in clinician notes is the most natural and efficient way to capture communication between patients and clinicians, nuanced clinical details, and explanation for medical decision-making.[4] These free text notes capture substantial and rich information related to patients’ health conditions; however, the unstructured format can make it difficult to use this information directly in patient care management and medical research without further information processing.[5] To resolve these challenges, over the past several decades, the field of clinical natural

How to cite this article: Xie F, Lee J, Munoz-Plaza CE, Hahn EE, Chen W. Application of text information extraction system for real-time cancer case identification in an integrated healthcare organization. J Pathol Inform 2017;8:48.

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_55_17
Surgical pathology reports (SPRs) contain valuable medical information embedded in the narrative free text, including information on the gross and microscopic description of tissue. SPRs are a vital research resource, in particular, for cancer-related research.\[8,10,12-22\] However, extracting information from SPR is generally time-consuming, laborious, costly, and requires manual processes and significant domain expertise. Alternatively, automating the application for information extraction (IE) provides a method for radically increasing the speed and scope with which this data can be accessed quickly.\[14\] A series of NLP methods and applications have been developed to retrieve this type of information from pathology reports.\[8,10,12,15,17,18,20,21,23\]

For example, the text information extraction system (TIES) released by the University of Pittsburgh School of Medicine\[8\] was initially focused on extracting cancer information from SPR and later extended to radiology reports to support multi-center collaborative translational medical research through a federated network model.\[21\] Yip et al. developed an N-gram model for concept discovery from pathology reports.\[19\] The iterative online machine learning-based IE system (IDEAL-X) was introduced by Zheng et al.\[16\] to incrementally process pathology reports to improve the learning model. The medical knowledge analysis tool pipeline (MedKATp), a system established by the Open Health Natural Language Processing Consortium,\[23\] automatically extracts cancer-specific characteristics from unstructured clinical reports. The pathology extraction pipeline (PEP), built on MedKATp by University of California at Irvine, was also developed for extracting data elements and relationships from pathology reports.\[20\] Additional examples and the details of these NLP systems were documented in the review article by Burger et al.\[12\]

The TIES system leverages the Noble coder\[22\] and other well-known NLP methods and algorithms to process pathology and radiology reports.\[8\] One of the advantages of the TIES system is the capacity to share de-identified data and tissue through coded SPRs among a federated research network of institutions under bundle of restrict security regulations and compliances. As of 2015, this federated TIES system had been implemented in four institutions\[21\] and accumulated >5 million coded pathology reports and 25 million radiology reports between January 2003 and January 2017.\[24\] As an end-to-end application for processing both pathology and radiology reports, the TIES system is increasingly of interest to biomedical research communities and healthcare maintenance organizations to facilitate translational medical research and clinical operations.\[25,26\] The TIES system attracted our attention due to its potential for rapid cancer case ascertainment in support of research studies. Especially, it offers detailed technical documentation and real-time online support, as well as other needed functionalities (such as deidentification and restricted secured model). On the other hand, the other systems such as MedKATp, PEP, and IDEAL-X are capable of effectively processing SPRs, but they are more focused on IE rather than case identification. Additional detail documentation and appropriate technical support are important for successful implementation of these comprehensive systems.

A majority of retrospective cancer research studies rely on local or national cancer registries (CANREGs) for cohort identification. This presents a significant challenge, especially for prospective studies such as clinical trials, because CANREG data are often delayed at least several months due to the lengthy manual process involved in collecting and coding registry data. Methods to quickly identify newly diagnosed cancer patients are critically needed to expedite prospective studies. In this paper, we will demonstrate the implementation and application of the TIES system in a large integrated health maintenance organization, Kaiser Permanente Southern California (KPSC), and the performance of the TIES application for rapid case ascertainment. Furthermore, the limitations and the caveats of the system learned through our processes will be shared to shorten the learning curves of potential future users.

**Methods**

**Implementation of the text information extraction system**

**Natural language processing software**

The TIES system, an open-source system formerly known as caTIES, was originally developed for the Shared Pathology Informatics Network to enable translational research within a federated network of institutions.\[8,21\] TIES has evolved over time through many iterations with the latest available version being V5.4 at the time of our implementation.\[24\] The TIES system consists of clients, services, and datastores connected and implemented under the Globus grid service architectures with a restricted regulatory model for federated data sharing (relying on Institutional Review Board (IRB) protocols and honest brokers). The unique TIES components used in our study are described below.

1. **Back-end Datastores:** The private datastore is the recipient of data and stores the original free text reports with identifiers while the research datastore contains de-identified free text reports for the NLP Pipeline Services, which creates and stores annotations for each de-identified report.

2. **Data preparation services:** The HL7 data importer loads the identified reports with HL7 specification format into the private datastore. The de-identifier recognizes and removes the identifiers protected by the Health Insurance Portability and Accountability Act. De-identification was achieved by using the built-in MGH scrubber. The concept coding service performs a sequence of NLP processing\[8\] to produce conceptual annotations and codes based on free-text reports. The indexing service creates an index for quick access to reports based on the text and conceptual codes being searched.
3. Information retrieval services: The data access and integration service provide web service which interacts with data sources. The search service communicates the user entered search criteria to the TIES server.

4. The restrict security enforcement layers guard the resources and authorize access based on roles of users.

The detailed components, functionalities and evaluations of the TIES system have been described in the published paper of the system[20] and the TIES official website.[24] The developers of the TIES system created a helpful user manual[27] and built a TIES community forum for technical discussions, support, user feedback, logs of issues and improvements.[28]

Deployment of the text information extraction system

Our initial implementation of TIES version 4.01 on a Window-based server started in 2008. Due to the rapidly increasing volumes of SPRs, we adopted version 5.4 in 2016, the latest version at the time of implementation. The new version was implemented on a Linux server. After a lengthy installation and configuration process described below, the implementation was a success. Although the TIES system has the capability for integration into a federated research network, our deployment was restricted to our local network without any outside communication activities.

Installation and configuration

The required hardware was prepared and software was downloaded according to the TIES system’s installation guideline.[27] The components included: (1) configuration and installment of the Linux server and disk space allocation for data storage; (2) third-party software installed were JAVA, MySQL database server, Tomcat web server, Apache Ant, and NCIMetathesaurus; and (3) The zip executable files for TIES application, TIES Java Message Services (JMS) services for supporting parallel coding, and ActiveMQ required by the TIES JMS services.[28] All downloaded software packages described above were first unzipped and installed into the designed folders in the Linux server. The configuration files of each installed package were then examined and adjusted to the corresponding installation settings as required.

Launching text information extraction system application, MySQL server, text information extraction system java message services, and ActiveMQ

A configured and executed script was stored at the configuration or common bin folder for each service. These scripts were submitted through a command line to launch a series of processes that brought up the TIES application, the MYSQL server, TIES JMS services, and ActiveMQ. The logs of the launching processes and the execution statuses were stored in the corresponding log files. It is critical to check the log files to make sure the launching processes are error-free. The launching process could also be performed through a job scheduler.

Operation of the text information extraction system at Kaiser Permanente Southern California

Data source

The CoPathPlus system, an interactive and comprehensive system of Cerner Corp, manages accessioning and handles specimens at the Kaiser Permanente national anatomic pathology laboratories. In the past few years, the system has processed more than a half million specimens annually.
Pathology reports are generated for all processed surgical accessions. The final diagnosis, based on the pathologist assessment, is in the form of free text and is a mandated section of any report. All the pathology reports are loaded into the KPSC Research Data Warehouse (RDW) on a daily basis, and these reports are further loaded into the TIES system.
Architecture of the Kaiser Permanente Southern California text information extraction system and data loading

The architecture of the TIES system within the KPSC research environment is illustrated in Figure 5. The column to the left demonstrates the flow of data extraction from CopathPlus, to RDW, and finally to the reports formatted in the HL7 specification. The middle part consists of the TIES application server, MySQL database server and web server, whereas the right side shows client requests through the TIES’s web-based client interfaces or querying directly against the back-end datastores.

First, the historic pathology reports were retrieved and converted into HL7 specification format and transferred into the HL7 data folder in the TIES server where the HL7 importer automatically loaded these reports into the back-end datastores of the TIES system. Second, a process was set up to extract and load the pathology reports on a daily basis. To handle large daily volumes of pathology reports (~4000/day), the TIES JMS service was implemented to process the reports in parallel. Currently, the KPSC TIES system contains over 3 million pathology reports dating back to 2013. For security control, we followed our internal operational procedures and the TIES built-in security protocol to manage and monitor the authorized access and other activities through the web-based administrative interface.

Searching cases by using a flexible time window

Instead of allowing users to specify time windows on a daily basis (e.g., 1/2/2015–1/8/2015), the front-end window of installed TIES system only supports the search of time windows on an annual basis at this point. This limitation prevents the effective use of the TIES system for timely cancer identification. For example, if a user is interested in identifying breast cancer patients evidenced by pathology reports last week, he/she is not able to rely on the front-end application to do so. Therefore, we developed a query process to search against the TIES back-end datastore through a batch mode (referred to as the direct method) to support our study application described below.
Application of the text information extraction system

We conducted a validation study to examine the ability of the TIES system to identify potential breast cancer cases diagnosed in December of 2013, using the KPSC CANREG as the gold standard. This is part of a large-scale internal initiative to develop efficient processes to prospectively identify newly diagnosed cancer patients for enrollment into clinical studies. The methods and results described below are limited to the validation study only. The KPSC CANREG contains information on patients who were diagnosed or received at least part of their first course of treatment for cancer at KPSC hospitals for all reportable cancers. The diagnosis date in CANREG was based on multiple sources, including diagnostic mammograms; thus, it could be potentially different from the reported date of a pathology report. Therefore, the TIES searching time window was extended to the end of 2014 from December 2013. Because the TIES system is based on the NCIM, we first used the online Unified Medical Language System terminology service (UTS) web application\(^\text{[29]}\) to search all potential NCIM codes describing breast cancer and the terms described in the Buckley et al. study\(^\text{[19]}\) to create the initial list of breast cancer terms and corresponding NCIM codes. This list was then reviewed and finalized by the study team. These selected NCIM terms and codes were then used to search against the TIES back-end datastores through the direct method described above. The search was limited to only the final diagnosis section, assuming such a section could be detected by the TIES system; otherwise, the search was applied to the entire pathology reports (the TIES system defines it as report section). Potential reasons for the failure of identifying final diagnosis sections are provided in the Discussion Section. After discovering that a majority of reports negative for breast cancer only contained negated cancer concepts (these are identified by the codes starting with the character “N”), only concepts that were not negated (the codes starting with the character “A”) were included for query extractions.

The potential breast cancer cases identified by the TIES were manually examined and compared against the CANREG data. Error analysis was performed. The study was approved by KPSC IRB.

RESULTS

A total of 437 specific NCIM terms and codes associated with breast cancer were identified through the UTS application. These specific terms contained the combination of the anatomical location (lobular or duct), malignant disorder (carcinoma or cancer), and severity (infiltrating or metastatic). The detailed list was included in Appendix 1. Some positive pathology reports were not identified through breast cancer concepts because breast and cancer were mentioned separately in different sentences or places. Therefore, an additional list was created, including 14 cancer concepts that were required to occur along with the breast concept [Appendix 2].

There were a total of 249 breast cancer cases diagnosed in December 2013 based on CANREG records. Of these, 241 cases (457 pathology reports) were found by using the preidentified concept codes listed in Appendices 1 and 2. Out of the eight false negative cases, negated terms were found in the pathology reports for 3 cases. Table 1 displays the number of the identified patients and the number of pathology reports by the concept codes used. A total of 13 preidentified concepts were found to have 10 or more reports while the “other terms” (all concepts with <10 reports combined) had 46 reports. Among the concepts for which 10 or more reports were found, the top three were “Ductal Breast Carcinoma in situ” (274 reports, and 171 patients), “Invasive Ductal Carcinoma, NOS” (197 reports and 124 patients), and “Carcinoma in situ” (156 reports and 119 patients), respectively. Nearly 56% of the diagnosis dates and pathology report dates were within 1 month, and the median of the difference of these two dates was within one and half months. The differences between the report sign-off dates and the diagnosis dates derived from the CANREG were shown in Table 2.

The TIES system also identified additional 277 potential cases with the report sign-off dates within December of
Due to the positive findings reported above, the reported algorithm was implemented in an ongoing research study to prospectively identify newly diagnosed cancer patients for recruitment.

**Discussions**

The TIES system is an end-to-end clinical medical NLP application which can be used to support single or multiple institutional collaborative cancer research. It has evolved over time, resulting in numerous versions since it was introduced a decade ago. The system has garnered particular attention from the NLP cancer-focused research community since the publication of the caTIES system (its previous version) in 2010[8] and the establishment of the TIES Cancer Research Network among four research institutions.[21]
Validation with patients diagnosed in December 2013 extracted from the CANREG demonstrated that the TIES system has the capability to identify 96.8% (241 of 249) of breast cancer cases. Of the eight false negative cases, one case lacked any pathology report. The misclassification of the remaining seven false negative cases was due to the following reasons: (1) The pathology reports lacked any of the study’s predefined breast cancer concepts ($n = 1$). (2) The pathology reports contained concepts being searched; however, they were either negated or appeared in the other report sections such as (addendum or gross description) instead of the final diagnosis ($n = 6$). The TIES system also identified additional potential breast cancer cases by searching pathology reports within December 2013. However, a majority of these cases (86%) were diagnosed in 2012 or 2013. The KPSC CANREG was based on multiple sources including diagnostic mammograms, biopsy pathology reports, etc. Therefore, there was a potential time lag between the diagnosed date and biopsy reporting date, which revealed that the diagnosis date could be the earlier date, while the biopsy date was the later confirmed date. However, the median of the date difference was within one and half months.

The developers of the TIES system provide a useful web-based online forum for knowledge sharing and issue discussion, as well as system support for a limited time. However, the implementation or migration of the TIES system will continue to face challenges without a better understanding of the fundamental mechanism and framework of the system. First, trouble-shooting the errors requires a thorough understanding of the working mechanism, process flow and corresponding computer and NLP technologies. Second, the TIES default coder pipeline only runs in a single process. Thus, the performance is reasonable for small data volume but could be deteriorated as the data volume increases significantly. At KPSC, when we loaded the historic data into the TIES system, the coding process was relatively sluggish. After checking available information in the TIES online support, we realized that the TIES JMS component (optional component not included in the single download package) could be downloaded to speed up the coding process. Third, although the TIES system is an open source tool, it remains difficult to customize the design and improve the system without a deep understanding of the system and advanced technical expertise. Fourth, as the TIES system stands today, it lacks the proper components for disaster and error recovery. One error could potentially result in a full system failure, requiring a complete rebuild. In addition, although we noted that the TIES web application session can be timed out automatically after a certain period without activities, the front-end web interface lacks a user signoff/logout button. Given these identified challenges, we recommend that new versions focus on developing solutions to address these potential issues.

The study identified several limitations that could be considered for future enhancement. First, the search time window can only be specified annually. However, in real life, a more flexible time window is required. For example, if a user intends to identify cancer patients real-time using the system, he/she will need a narrower time window (e.g., month, week, or day). Second, the de-identified pipeline incorrectly de-identified some words or terms. For example, we noted that the word “mass” was constantly de-identified and resulted in an erroneous report. Third, when the TIES system codes the pathology report, a small percentage of reports erroneously combined all sections into a single “report section” rather than keeping the specific sections, such as “Clinical history,” “Final diagnosis,” “Gross description,” etc. In this instance, the TIES system searched the entire context of the report and was unable to limit to the truly “Final diagnosis” section. Such a misclassification could result in either false negative cases or false positive cases when a user searches with a specific section in a pathology report. As a result, three historical cases identified from the actual “Clinical history” section were not identified as historic cases and therefore misclassified as current potential cases. This type of misclassification could be avoided by using the section detection functionality provided by the TIES to properly configure sections with new section headers. Fourth, it seems the TIES engine is unable to accurately exclude the historic cases when the history term and breast cancer concept term are located in different sentences. For example, the system failed

### Table 3: Distribution of pathology reports identified by text information and extraction system with the predefined breast cancer concepts and codes for pathologist signoff date within December, 2013 for whom the corresponding records were not found in the subset of breast cancer patients diagnosed in December of 2013 based on the cancer registry

| NCIM code and term | Total of pathology reports* | Total of patients* |
|--------------------|----------------------------|--------------------|
| C0007124 - ductal breast carcinoma in situ | 178 | 173 |
| C1134719 - invasive ductal carcinoma, NOS | 110 | 108 |
| C0279563 - lobular breast carcinoma in situ | 43 | 43 |
| C0007099 - carcinoma in situ | 37 | 37 |
| C1334274 - invasive carcinoma | 26 | 26 |
| C0442835 - atypical lobular breast hyperplasia | 25 | 25 |
| C1384494 - metastatic carcinoma | 18 | 18 |
| C1176475 - ductal carcinoma | 15 | 15 |
| C0678222 - breast carcinoma | 15 | 15 |
| C0206692 - lobular breast carcinoma | 12 | 12 |
| C0278488 - Stage IV breast cancer | 10 | 10 |
| C0853879 - invasive breast carcinoma | 10 | 10 |
| Other terms | 32 | 32 |

*One patient may have multiple reports and one report may contain one or more NCIM codes. The total of unique report was 287, which belonged to 277 unique patients. NCIM: National Cancer Institute Metathesaurus, NOS: Not otherwise specified.
to recognize the historical nature when “History of” appeared in the first sentence, and “breast carcinoma” appeared in the following sentence. Finally, there were cases in which the negation should be applied to two or more conditions while the TIES system can only negate the condition close to the negation term. For example, the negation term “negative for” in the description “negative for dysplasia and malignancy” should be negated for the conditions of “dysplasia” and “malignancy”. However, the TIES system only highlighted “dysplasia” as negative diagnosis condition. Despite these limitations, our study demonstrated that the TIES system offers a robust and precise breast cancer case identification. In addition, the TIES system has the great potential to be easily applied to search for other cancer types, either single or multiple, with minimal development work.

CONCLUSIONS
We have successfully implemented the TIES system to import and process pathology reports on a daily basis. The validated results demonstrated that the TIES system can effectively identify the potential breast cancer cases within our care setting. All identified potential cases can be easily confirmed by reviewing the corresponding annotated reports through the front-end visualization interface. The TIES system is a useful NLP tool to identify cancer cases in a timely and efficient manner to support research studies and operational care management.

Acknowledgment
We would like to thank Philip Morales for the installation of the Linux server and the TIES system.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES
1. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, D.C.: Institute of Medicine; 2001.
2. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009;78:22-31.
3. Cadarette SM, Wong L. An introduction to health care administrative data. Can J Hosp Pharm 2015;68:232-7.
4. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000;270-4.
5. Rosenblum ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: A perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18:181-6.
6. Chapman B, Chapman WW, Dayton G, Mowery D. Python Implementation of the ConText Algorithm. Available from: https://www.pyipi.python.org/pyipi/pyConTextNLP. Last accessed on 2017 Jun 09.
7. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507-13.
8. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. cTIES: A grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010;17:253-64.
9. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations; 2014.
10. Zheng S, Lu JJ, Appin C, Brat D, Wang F. Support patient search on pathology reports with interactive online learning based data extraction. J Pathol Inform 2015;6:51.
11. Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: A unified system architecture overview. Methods Mol Biol 2014;1168:275-94.
12. Burger G, Abu-Hanna A, de Keizer N, Corent R. Natural language processing in pathology: A scoping review. J Clin Pathol 2016; pii:jclinpath-2016-203872.
13. Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annu Symp Proc 2003:584-8.
14. Mitchell KJ, Becich MJ, Berman JJ, Chapman WW, Gilbertson J, Gupta D, et al. Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. Stud Health Technol Inform 2004;107:663-7.
15. Schlangen D, Stede M, Bontas EP. Feeding OWL: Extracting and Representing the Content of Pathology Reports, in Proceedings of the Workshop on NLP and XML (NLPPXM-2004): RDF/RDFS and OWL in Language Technology. Association for Computational Linguistics; 2004. p. 43-50.
16. Carrell D, Miglioretti D, Smith-Bindman R. Coding free text radiology reports using the Cancer Text Information Extraction System (cTIES). Proc AMIA Symp 2007;889.
17. Coden A, Savoya G, Sominsky I, Tenenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009;42:937-49.
18. Yip V, Mete M, Topaloglu U, Koeckara S. Concept discovery for pathology reports using an N-gram model. AMIA Jt Summits Transl Sci Proc 2010;2010:43-7.
19. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23.
20. Ashish N, Dahm L, Boicey C. University of California, Irvine-Pathology Extraction Pipeline: The pathology extraction pipeline for information extraction from pathology reports. Health Informatics J 2014;20:288-305.
21. Jacobson RS, Becich MJ, Bollag RJ, Chavan G, Corrigan J, Dhir R, et al. A federated network for translational cancer research using clinical data and biospecimens. Cancer Res 2015;75:5194-201.
22. Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, Jacobson RS, et al. NOBLE - flexible concept recognition for large-scale biomedical natural language processing. BMC Bioinformatics 2016;17:32.
23. MedKATp. Available from: http://www.ohnlp.org/index.php/MedKAT/p. [Last accessed on 2017 Oct 12].
24. Text Information Extraction System. Available from: http://www.ties.dbmi.pitt.edu/. [Last accessed on 2017 Jun 09].
25. Evans MH, Rohm BW, Schultz FA, Kroth PJ. Using cTIES as a case-finding tool in tissue repositories: System challenges and lessons learned. National Cancer Institute Cancer Biomedical Informatics Grid (cBIG) Annual Meeting, Washington, DC; 2009.
26. Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. Am J Epidemiol 2014;179:749-58.
27. TIES Dictionaries. Available from: http://www.ties.upmc.edu/doc/TIES%20v5%20User%20Manual.pdf. [Last accessed on 2017 Mar 20].
28. TIES Forum. Available from: https://www.sourceforge.net/p/caties/discussion/626701/. [Last accessed on 2017 Jun 09].
29. UMLS Terminology Services. Available from: https://www.uts.nlm.nih.gov/home.html. [Last accessed on 2017 Oct 12].
APPENDICES

Appendix 1: The breast cancer term lists and National Cancer Institute Metathesaurus codes identified by the Unified Medical Language System Terminology Service Metathesaurus browser web application

| Term                                                                 | Code     |
|---------------------------------------------------------------------|----------|
| Malignant breast neoplasm                                           | C0006142 |
| Female breast carcinoma                                             | C0007104 |
| Ductal breast carcinoma in situ                                     | C0007124 |
| Malignant neoplasm of nipple and areola of female breast            | C0024621 |
| Paget disease of the breast                                         | C0030185 |
| Malignant neoplasm of central part of female breast                 | C0153549 |
| Malignant neoplasm of upper-inner quadrant of female breast        | C0153550 |
| Malignant neoplasm of lower-inner quadrant of female breast        | C0153551 |
| Malignant neoplasm of upper-outer quadrant of female breast        | C0153552 |
| Malignant neoplasm of lower-outer quadrant of female breast        | C0153553 |
| Malignant neoplasm of axillary tail of female breast               | C0153554 |
| Malignant neoplasm of other specified sites of female breast       | C0153555 |
| Malignant neoplasm of nipple and areola of male breast              | C0153558 |
| Malignant neoplasm of other and unspecified sites of male breast   | C0153559 |
| Stage 0 breast cancer                                               | C0154084 |
| Lobular breast carcinoma                                            | C0206692 |
| Malignant neoplasm of female breast                                 | C0235653 |
| Male breast carcinoma                                               | C0238033 |
| Malignant neoplasm of male breast                                   | C0242787 |
| Stage I breast cancer AJCC v6                                       | C0278485 |
| Stage II breast cancer                                              | C0278486 |
| Stage III breast cancer AJCC v6                                     | C0278487 |
| Stage IV breast cancer                                              | C0278488 |
| Stage IIIA breast cancer                                            | C0278489 |
| Recurrent breast carcinoma                                         | C0278493 |
| Stage IIIB breast cancer                                            | C0278513 |
| Inflammatory breast carcinoma                                       | C0278601 |
| Invasive ductal breast carcinoma with predominant intraductal component | C0279556 |
| Lobular breast carcinoma in situ                                    | C0279563 |
| Invasive lobular breast carcinoma with predominant in situ component | C0279564 |
| Invasive lobular breast carcinoma                                   | C0279565 |
| Paget disease and intraductal carcinoma of the breast               | C0279566 |
| Paget disease of the breast with invasive ductal carcinoma          | C0279567 |
| Cellular diagnosis, breast cancer                                   | C0279855 |
| Bilateral breast carcinoma                                          | C0281267 |
| Secretory breast carcinoma                                          | C0334371 |
| Intraductal papillary breast carcinoma                              | C0334372 |
| Intraductal papillary adenocarcinoma with invasion                  | C0334373 |
| Intracystic papillary breast carcinoma                              | C0334376 |

Contd...

Appendix 1: Contd...

| Term                                                                 | Code     |
|---------------------------------------------------------------------|----------|
| Ductal breast carcinoma in situ and lobular carcinoma in situ       | C0334383 |
| Invasive ductal and lobular carcinoma in situ                       | C0334384 |
| Scirrhous breast carcinoma                                          | C0346151 |
| Cancer en cuirasse                                                  | C0346152 |
| Hereditary breast carcinoma                                         | C0346153 |
| Malignant breast phyllodes tumor                                    | C0346154 |
| Primary malignant neoplasm of skin of breast                       | C0346742 |
| Breast melanoma                                                     | C0346787 |
| Primary malignant neoplasm of nipple of male breast                 | C0346857 |
| Primary malignant neoplasm of areola of male breast                 | C0346858 |
| Malignant neoplasm of ectopic site of male breast                   | C0346860 |
| Primary malignant neoplasm of nipple of female breast               | C0346861 |
| Primary malignant neoplasm of areola of female breast               | C0346862 |
| Malignant neoplasm of ectopic site of female breast                 | C0346865 |
| Secondary malignant neoplasm of skin of breast                      | C0346986 |
| Metastatic malignant neoplasm in the breast                         | C0346993 |
| Carcinoma in situ of skin of breast                                | C0347152 |
| Other carcinoma in situ                                            | C0348409 |
| Malignant neoplasm, overlapping lesion of breast                     | C0348912 |
| Breast sarcoma                                                      | C0349667 |
| Breast lymphoma                                                      | C0349669 |
| Atypical lobular breast hyperplasia                                  | C0442835 |
| Malignant neoplasm: Nipple and areola                               | C0496806 |
| Malignant neoplasm of central portion of breast                     | C0496807 |
| Malignant neoplasm of breast upper inner quadrant                   | C0496808 |
| Malignant neoplasm of breast lower inner quadrant                   | C0496809 |
| Malignant neoplasm of breast upper outer quadrant                   | C0496810 |
| Malignant neoplasm of breast lower outer quadrant                   | C0496811 |
| Malignant neoplasm of axillary tail of breast                       | C0496812 |
| Carcinoma of axillary tail of breast                               | C0559063 |
| Metastasis to breast of unknown primary                             | C0563510 |
| Carcinoma of breast - upper, inner quadrant                        | C0564706 |
| Carcinoma of breast - lower, inner quadrant                        | C0564707 |
| Carcinoma of breast - upper, outer quadrant                        | C0564708 |
| Carcinoma of breast - lower, outer quadrant                        | C0564709 |
| Breast carcinoma                                                    | C0678222 |
| Malignant melanoma of skin of breast                               | C0684503 |
| Carcinoma in situ of female breast                                 | C0686288 |
| Carcinoma in situ of nipple of female breast                       | C0686292 |
| Secondary malignant neoplasm of nipple of female breast             | C0686293 |
| Carcinoma in situ of areola of female breast                       | C0686296 |
| Secondary malignant neoplasm of areola of female breast             | C0686297 |
| Carcinoma in situ of central portion of female breast               | C0686300 |
| Secondary malignant neoplasm of central portion of female breast    | C0686301 |
| Carcinoma in situ of upper inner quadrant of female breast         | C0686304 |
### Appendix 1: Contd...

| Term                                                                 | Code    |
|----------------------------------------------------------------------|---------|
| Secondary malignant neoplasm of upper inner quadrant of female breast | C0686305 |
| Carcinoma *in situ* of lower inner quadrant of female breast        | C0686308 |
| Secondary malignant neoplasm of lower inner quadrant of female breast | C0686309 |
| Carcinoma *in situ* of upper outer quadrant of female breast        | C0686312 |
| Secondary malignant neoplasm of upper outer quadrant of female breast | C0686313 |
| Carcinoma *in situ* of lower outer quadrant of female breast        | C0686316 |
| Secondary malignant neoplasm of lower outer quadrant of female breast | C0686317 |
| Carcinoma *in situ* of axillary tail of female breast               | C0686320 |
| Secondary malignant neoplasm of axillary tail of female breast       | C0686321 |
| Carcinoma *in situ* of ectopic female breast tissue                 | C0686324 |
| Primary malignant neoplasm of ectopic female breast tissue          | C0686325 |
| Secondary malignant neoplasm of ectopic female breast tissue        | C0686326 |
| Carcinoma *in situ* of male breast                                   | C0686328 |
| Secondary malignant neoplasm of male breast                         | C0686329 |
| Carcinoma *in situ* of nipple of male breast                        | C0686332 |
| Secondary malignant neoplasm of nipple of male breast                | C0686333 |
| Carcinoma *in situ* of areola of male breast                        | C0686336 |
| Secondary malignant neoplasm of areola of male breast                | C0686337 |
| Carcinoma *in situ* of ectopic male breast tissue                   | C0686340 |
| Primary malignant neoplasm of ectopic male breast tissue            | C0686341 |
| Secondary malignant neoplasm of ectopic male breast tissue          | C0686342 |
| Invasive breast carcinoma                                            | C0853879 |
| Recurrent inflammatory breast carcinoma                              | C0853968 |
| Stage IIIB inflammatory breast carcinoma                             | C0853971 |
| Stage IV inflammatory breast carcinoma                               | C0853972 |
| Breast cancer aggravated                                             | C0856130 |
| Breast lump (malignant)                                              | C0857005 |
| Slow growing lung and soft tissue metastases from cancer breast     | C0857220 |
| Breast adenocarcinoma                                                | C0858252 |
| Malignant nipple neoplasm                                            | C0859086 |
| Colloid breast carcinoma                                             | C0860579 |
| Medullary breast carcinoma                                           | C0860580 |
| Mucinous breast cancer                                               | C0860581 |
| Lobular neoplasia                                                    | C0861352 |
| Breast adenocarcinoma recurrent                                      | C0861355 |
| Colloid breast carcinoma recurrent                                   | C0861357 |
| Lobular carcinoma recurrent                                          | C0861358 |
| Medullary carcinoma of breast recurrent                              | C0861359 |
| Mucinous breast cancer recurrent                                     | C0861360 |
| Mucinous ductal breast carcinoma recurrent                           | C0861361 |
| Breast adenocarcinoma Stage I                                       | C0861362 |
| Colloidal breast carcinoma Stage I                                   | C0861364 |
| Lobular breast carcinoma Stage I                                     | C0861366 |
| Lobular carcinoma Stage I                                            | C0861367 |
| Medullary carcinoma of breast Stage I                                | C0861368 |
| Mucinous breast cancer Stage I                                       | C0861369 |
| Mucinous ductal breast carcinoma Stage I                             | C0861370 |
| Breast adenocarcinoma Stage II                                       | C0861371 |
| Colloid breast carcinoma Stage II                                    | C0861373 |
| Ductal breast carcinoma Stage II                                     | C0861374 |
| Lobular breast carcinoma Stage II                                    | C0861375 |
| Lobular carcinoma Stage II                                           | C0861376 |
| Medullary carcinoma of breast Stage II                               | C0861377 |
| Mucinous breast cancer Stage II                                      | C0861378 |
| Mucinous ductal breast carcinoma Stage II                            | C0861379 |
| Colloid breast carcinoma Stage III                                   | C0861382 |
| Ductal breast carcinoma Stage III                                    | C0861383 |
| Lobular breast carcinoma Stage III                                   | C0861384 |
| Lobular carcinoma Stage III                                          | C0861385 |
| Medullary carcinoma of breast Stage III                              | C0861386 |
| Mucinous breast cancer Stage III                                     | C0861387 |
| Mucinous ductal breast carcinoma Stage III                           | C0861388 |
| Breast adenocarcinoma Stage IV                                       | C0861389 |
| Colloid breast carcinoma Stage IV                                    | C0861391 |
| Ductal breast carcinoma Stage IV                                     | C0861392 |
| Lobular breast carcinoma Stage IV                                    | C0861393 |
| Lobular carcinoma Stage IV                                           | C0861394 |
| Medullary carcinoma of breast Stage IV                               | C0861395 |
| Mucinous breast cancer Stage IV                                      | C0861396 |
| Mucinous ductal breast carcinoma Stage IV                            | C0861397 |
| Breast adenocarcinoma                                                | C0861398 |
| Lobular neoplasia                                                    | C0861399 |
| Female malignant nipple neoplasm                                     | C0861400 |
| Contralateral breast cancer                                          | C1096616 |
| Squamous cell breast cancer cancer                                   | C1112794 |
| Invasive ductal carcinoma, NOS                                        | C1134719 |
| Overlapping malignant neoplasm of female breast                     | C1263794 |
| Overlapping malignant neoplasm of male breast                       | C1263804 |
| Carcinoma *in situ* of other site of breast                         | C1263808 |
| Local recurrence of malignant tumor of breast                       | C1282471 |
| Primary malignant neoplasm of breast lower outer quadrant            | C1298788 |
| Primary malignant neoplasm of breast upper outer quadrant            | C1298924 |
| Primary malignant neoplasm of breast upper inner quadrant            | C1298925 |
| Primary malignant neoplasm of breast lower inner quadrant            | C1298926 |
| Primary malignant neoplasm of axillary tail of breast                | C1299235 |
| Secondary malignant neoplasm of axillary tail of breast              | C1299236 |
| Primary malignant neoplasm of breast                                | C1299258 |
| Localized skin involvement by breast carcinoma                      | C1304482 |
| Primary malignant neoplasm of female breast                         | C1304708 |
| Primary malignant neoplasm of central portion of female breast       | C1305893 |
### Appendix 1: Contd...

| Term                                                                 | Code     |
|----------------------------------------------------------------------|----------|
| Primary malignant neoplasm of upper quadrant of female breast        | C1306024 |
| Primary malignant neoplasm of lower quadrant of female breast        | C1306025 |
| Primary malignant neoplasm of upper outer quadrant of female breast  | C1306026 |
| Primary malignant neoplasm of lower outer quadrant of female breast  | C1306027 |
| Primary malignant neoplasm of male breast                            | C1306469 |
| Tubular breast carcinoma                                              | C1328544 |
| Tubular breast cancer Stage I                                         | C1328545 |
| Tubular breast cancer Stage III                                       | C1328547 |
| Tubular breast cancer Stage IV                                        | C1328548 |
| Tubular breast cancer metastatic                                      | C1328549 |
| Adenoid cystic breast carcinoma                                       | C1332167 |
| Apocrine breast carcinoma in situ                                     | C1332315 |
| Apocrine breast carcinoma                                             | C1332316 |
| Breast adenocarcinoma with squamous metaplasia                        | C1332613 |
| Breast angiosarcoma                                                   | C1332614 |
| Breast carcinoma metastatic in the bone                               | C1332623 |
| Breast carcinoma metastatic in the brain                              | C1332624 |
| Breast carcinoma metastatic in the liver                              | C1332625 |
| Breast carcinoma metastatic in the lung                               | C1332626 |
| Breast fibrosarcoma                                                   | C1332630 |
| Breast leiomyosarcoma                                                 | C1332631 |
| Breast liposarcoma                                                    | C1332632 |
| Breast mucosa-associated lymphoid tissue lymphoma                     | C1332633 |
| Breast rhabdomyosarcoma                                               | C1332637 |
| Breast small cell carcinoma                                           | C1332638 |
| Ductal breast carcinoma with squamous metaplasia                      | C1333319 |
| Grade 1 invasive breast carcinoma                                     | C1333832 |
| Grade 2 invasive breast carcinoma                                     | C1333838 |
| Grade 3 invasive breast carcinoma                                     | C1333843 |
| Hereditary female breast carcinoma                                    | C1333986 |
| Hereditary male breast carcinoma                                      | C1333988 |
| High grade ductal breast carcinoma in situ                            | C1334002 |
| High grade mucoepidermoid breast carcinoma                            | C1334006 |
| Intermediate grade ductal breast carcinoma in situ                    | C1334206 |
| Intraductal cribriform breast adenocarcinoma                          | C1334248 |
| Intraductal micropapillary breast carcinoma                           | C1334249 |
| Intraductal noncomedo breast adenocarcinoma                           | C1334250 |
| Invasive apocrine breast carcinoma                                     | C1334272 |
| Invasive breast carcinoma by histologic grade                         | C1334273 |
| Invasive cribriform breast carcinoma                                  | C1334275 |
| Invasive ductal and invasive lobular breast carcinoma                 | C1334276 |
| Invasive ductal and lobular carcinoma                                 | C1334277 |
| Invasive papillary breast carcinoma                                   | C1334280 |
| Low grade ductal breast carcinoma in situ                             | C1334413 |
| Low grade mucoepidermoid breast carcinoma                             | C1334417 |
| Malignant breast adenomyoepithelioma                                  | C1334564 |
| Malignant breast eccrine spiradenoma                                  | C1334565 |
| Metaplastic breast carcinoma                                          | C1334708 |
| Metastatic signet ring cell breast carcinoma                          | C1334740 |
| Metastatic squamous cell breast carcinoma                             | C1334743 |
| Mucinous breast carcinoma                                             | C1334807 |

### Appendix 1: Contd...

| Term                                                                 | Code     |
|----------------------------------------------------------------------|----------|
| Mucoepidermoid breast carcinoma                                       | C1334813 |
| Nipple carcinoma                                                     | C1334966 |
| Nipple duct carcinoma                                                | C1334967 |
| Breast extraskeletal osteosarcoma                                     | C1335149 |
| Non-Hodgkin breast lymphoma                                          | C1335489 |
| Breast T-cell Non-Hodgkin lymphoma                                   | C1335493 |
| Signet ring cell breast carcinoma                                    | C1335964 |
| Solid papillary breast carcinoma                                     | C1336027 |
| Sporadic breast carcinoma                                           | C1336076 |
| Squamous cell breast carcinoma                                       | C1336079 |
| Squamous cell carcinoma in situ of the nipple                        | C1336080 |
| Stage IIa breast cancer                                               | C1336156 |
| Stage IIb breast cancer                                               | C1336178 |
| Tubular breast cancer Stage II                                        | C1504470 |
| Adenosquamous breast carcinoma                                       | C1510796 |
| Breast adenocarcinoma with spindle cell metaplasia                   | C1511281 |
| Breast burkitt lymphoma                                              | C1511286 |
| Breast carcinoma with choriocarcinomatous features                   | C1511302 |
| Breast carcinoma with melanotic features                             | C1511303 |
| Breast carcinoma with osteoclastic giant cells                       | C1511304 |
| Breast columnar mucinous carcinoma                                   | C1511305 |
| Breast diffuse large B-cell lymphoma                                 | C1511306 |
| Breast follicular lymphoma                                           | C1511311 |
| Breast large cell neuroendocrine carcinoma                           | C1511316 |
| Breast mucinous cystadenocarcinoma                                   | C1511318 |
| Glycogen-rich, clear cell breast carcinoma                           | C1512224 |
| Mixed epithelial/mesenchymal metastatic breast carcinoma              | C1513365 |
| Pleomorphic breast carcinoma                                         | C1514169 |
| Postradiotherapy breast angiosarcoma                                  | C1514246 |
| Synchronous bilateral breast carcinoma                              | C1515107 |
| Acinic cell breast carcinoma                                         | C1515868 |
| Invasive mixed breast carcinoma                                      | C1517577 |
| Lipid-rich breast carcinoma                                          | C1517894 |
| Low grade adenosquamous breast carcinoma                             | C1518013 |
| Malignant breast myoepithelioma                                      | C1518167 |
| Oncocytic breast carcinoma                                           | C1518574 |
| Sebaceous breast carcinoma                                            | C1519207 |
| Squamous cell breast carcinoma, acantholytic variant                 | C1519485 |
| Squamous cell breast carcinoma, large cell keratinizing variant       | C1519486 |
| Squamous cell breast carcinoma, spindle cell variant                 | C1519487 |
| Ductal breast carcinoma                                              | C1527349 |
| Hormone receptor positive malignant neoplasm of breast                | C1562029 |
| Breast adenocarcinoma metastatic                                     | C1697918 |
| Paget disease of the nipple                                           | C1704323 |
| Breast carcinoma with chondroid metaplasia                           | C1704704 |
| Paget disease of the breast without invasive carcinoma                | C1709447 |
| Unilateral breast carcinoma                                          | C1710547 |
| Breast carcinoma with osseous metaplasia                             | C1711312 |
| Atypical medullary breast carcinoma                                  | C1879758 |
| Ductal breast carcinoma in situ, solid type                          | C1880424 |
| Invasive lobular breast carcinoma, signet ring variant                | C1883029 |
### Appendix 1: Contd...

| Term                                                                 | Code   |
|----------------------------------------------------------------------|--------|
| Her2 positive breast carcinoma                                        | C1960398 |
| Stage I breast cancer                                                 | C2216695 |
| Malignant neoplasm of breast staging                                  | C2216702 |
| Human epidermal growth factor 2 negative carcinoma of breast          | C2316304 |
| Metastatic ductal breast carcinoma                                    | C2698203 |
| Metastatic lobular breast carcinoma                                   | C2698204 |
| Microinvasive breast carcinoma                                        | C2732473 |
| Lobular carcinoma in situ with microinvasion                           | C2733298 |
| Ductal carcinoma in situ with microinvasion and involving nipple skin | C2733413 |
| Lobular neoplasia Type A                                              | C2826777 |
| Lobular neoplasia Type B                                              | C2826778 |
| Malignant neoplasm of nipple and areola, female                       | C2842076 |
| Malignant neoplasm of nipple and areola, right female breast          | C2842077 |
| Malignant neoplasm of nipple and areola, left female breast           | C2842078 |
| Malignant neoplasm of nipple and areola, unspecified female breast    | C2842079 |
| Malignant neoplasm of nipple and areola, male                         | C2842080 |
| Malignant neoplasm of nipple and areola, right male breast            | C2842081 |
| Malignant neoplasm of nipple and areola, left male breast             | C2842082 |
| Malignant neoplasm of nipple and areola, unspecified male breast      | C2842083 |
| Malignant neoplasm of central portion of right female breast          | C2842084 |
| Malignant neoplasm of central portion of left female breast           | C2842085 |
| Malignant neoplasm of central portion of unspecified female breast    | C2842086 |
| Malignant neoplasm of central portion of breast, male                 | C2842087 |
| Malignant neoplasm of central portion of right male breast            | C2842088 |
| Malignant neoplasm of central portion of left male breast             | C2842089 |
| Malignant neoplasm of central portion of unspecified male breast      | C2842090 |
| Malignant neoplasm of upper-inner quadrant of right female breast     | C2842091 |
| Malignant neoplasm of upper-inner quadrant of left female breast      | C2842092 |
| Malignant neoplasm of upper-inner quadrant of unspecified female breast| C2842093 |
| Malignant neoplasm of upper-inner quadrant of breast, male            | C2842094 |
| Malignant neoplasm of upper-inner quadrant of right male breast       | C2842095 |
| Malignant neoplasm of upper-inner quadrant of left male breast        | C2842096 |
| Malignant neoplasm of upper-inner quadrant of unspecified male breast | C2842097 |
| Malignant neoplasm of lower-inner quadrant of right female breast     | C2842098 |
| Malignant neoplasm of lower-inner quadrant of left female breast      | C2842099 |

Contd...
### Appendix 1: Contd...

| Term                                                                 | Code       |
|----------------------------------------------------------------------|------------|
| Malignant neoplasm of overlapping sites of breast, male              | C2842130   |
| Malignant neoplasm of overlapping sites of right male breast        | C2842131   |
| Malignant neoplasm of overlapping sites of left male breast         | C2842132   |
| Malignant neoplasm of overlapping sites of unspecified male breast  | C2842133   |
| Malignant neoplasm of breast of unspecified site, female             | C2842134   |
| Malignant neoplasm of breast of unspecified site of right female breast | C2842135   |
| Malignant neoplasm of unspecified site of left female breast        | C2842136   |
| Malignant neoplasm of unspecified site of unspecified female breast | C2842137   |
| Malignant neoplasm of breast of unspecified site, male               | C2842138   |
| Malignant neoplasm of unspecified site of right male breast         | C2842139   |
| Malignant neoplasm of unspecified site of left male breast          | C2842140   |
| Malignant neoplasm of unspecified site of unspecified male breast   | C2842141   |
| Lobular carcinoma *in situ* of right breast                        | C2865371   |
| Lobular carcinoma *in situ* of left breast                         | C2865372   |
| Intraductal carcinoma *in situ* of right breast                    | C2865374   |
| Intraductal carcinoma *in situ* of left breast                     | C2865375   |
| Unspecified type of carcinoma *in situ* of right breast            | C2865380   |
| Unspecified type of carcinoma *in situ* of left breast             | C2865381   |
| Pleomorphic lobular carcinoma *in situ*                            | C2919327   |
| Classic lobular carcinoma *in situ*                                | C2919427   |
| Estrogen receptor positive breast cancer                           | C2938924   |
| Lobular carcinoma *in situ* of unspecified breast                  | C2976799   |
| Intraductal carcinoma *in situ* of unspecified breast              | C2976800   |
| Other specified type of carcinoma *in situ* of left breast         | C2976801   |
| Lobular breast carcinoma *in situ*                                 | C2976802   |
| Other specified type of carcinoma *in situ* of unspecified breast  | C2976803   |
| Other specified type of carcinoma *in situ* of right breast        | C2976804   |
| Unspecified type of carcinoma *in situ* of unspecified breast      | C2976805   |
| Stage IIIC breast cancer                                           | C2980042   |
| Breast carcinoma by AJCC v6 stage                                  | C2983712   |
| Breast carcinoma by AJCC v7 stage                                  | C2984094   |
| Multifocal breast carcinoma                                        | C2986662   |
| Multicentric breast carcinoma                                      | C2986664   |
| Early-stage breast carcinoma                                       | C2986665   |
| Stage III breast cancer                                            | C3146271   |
| Node-positive breast cancer                                         | C3160887   |
| Node-negative breast cancer                                         | C3160889   |
| Sarcoma of axillary tail of female breast                          | C3163806   |
| Sarcoma lower inner quadrant of female breast                      | C3163865   |
| Sarcoma of central portion of female breast                        | C3163866   |
| Sarcoma of male breast                                             | C3164299   |

### Appendix 1: Contd...

| Term                                                                 | Code       |
|----------------------------------------------------------------------|------------|
| Sarcoma upper outer quadrant of female breast                       | C3164606   |
| Sarcoma of female breast                                            | C3164849   |
| Sarcoma of upper inner quadrant of female breast                    | C3164883   |
| Sarcoma of lower outer quadrant of female breast                    | C3165073   |
| Infiltrating duct carcinoma of female breast                        | C3165106   |
| Invasive lobular breast carcinoma, alveolar variant                 | C3273215   |
| Invasive lobular breast carcinoma, pleomorphic variant              | C3273216   |
| Invasive lobular breast carcinoma, solid variant                    | C3273217   |
| Invasive lobular breast carcinoma, tubulolobular variant            | C3273218   |
| Breast solid neuroendocrine carcinoma                               | C3273727   |
| Contralateral breast carcinoma                                      | C3274709   |
| Advanced breast cancer                                              | C3495917   |
| Locally advanced breast cancer                                      | C3495949   |
| Triple-negative breast carcinoma                                    | C3539878   |
| Breast carcinoma by gene expression profile                         | C3642344   |
| Luminal A breast carcinoma                                          | C3642345   |
| Luminal B breast carcinoma                                          | C3642346   |
| Basal-like breast carcinoma                                         | C3642347   |
| Normal breast-like subtype of breast carcinoma                     | C3642471   |
| Papillary breast carcinoma                                          | C3812899   |
| Tubulolobular carcinoma                                             | C3838879   |
| Invasive micropapillary breast carcinoma                            | C3838947   |
| Intraductal papilloma with ductal carcinoma *in situ*               | C3839576   |
| Solid papillary carcinoma *in situ*                                | C3839648   |
| Childhood breast carcinoma                                          | C3897071   |
| Mixed lobular and ductal breast carcinoma                           | CL007210   |
| Ductal breast carcinoma *in situ* and invasive lobular carcinoma    | CL018755   |
| Intraductal and lobular carcinoma                                   | CL028597   |
| Hormone receptor/growth factor receptor-negative breast cancer      | CL412277   |
| Hormone receptor/growth factor receptor-positive breast cancer      | CL412278   |
| Estrogen receptor-negative breast cancer                            | CL412279   |
| Progesterone receptor-negative breast cancer                        | CL412281   |
| Progesterone receptor-positive breast cancer                        | CL412282   |
| HER2-negative breast cancer                                         | CL412283   |
| Hormone-resistant breast cancer                                     | CL412374   |
| Stage IA breast cancer                                              | CL413891   |
| Stage IB breast cancer                                              | CL413892   |
| Invasive lobular breast carcinoma recurrent                        | CL449694   |
| Mucinous breast carcinoma recurrent                                 | CL449695   |
| Premenopausal breast cancer                                         | CL449698   |
| Mixed ductal lobular breast carcinoma infiltrating                  | CL453394   |
| Tubular breast cancer                                               | CL497426   |
| Primary malignant neoplasm of female right breast                   | CL498822   |
| Intraductal carcinoma *in situ* of bilateral breasts               | CL500272   |
| Infiltrating duct carcinoma of left female breast                   | CL500273   |
| Infiltrating duct carcinoma of right female breast                  | CL500274   |
| Infiltrating duct carcinoma of bilateral female breasts            | CL500275   |
| Carcinoma of central portion of breast                             | CL500661   |
| Mucinous carcinoma of breast                                        | C1334807   |
| Infiltrating ductal carcinoma of breast, Stage 1                    | C1827104   |
Appendix 1: Contd...

| Term                                                                 | Code       |
|----------------------------------------------------------------------|------------|
| infiltrating ductal and lobular carcinoma of breast                   | C2076522   |
| Infiltrating ductal carcinoma of breast, Stage 3                      | C1827241   |
| Infiltrating ductal carcinoma of breast, Stage 2                      | C1827300   |
| Infiltrating ductal carcinoma of breast, Stage 4                      | C1828351   |

HER2: Human epidermal growth factor receptor 2, NOS: Not otherwise specified, AJCC: American joint committe on cancer

Appendix 2: The combination list of the terms and National Cancer Institute Metathesaurus codes of “breast” and “cancer”

| Term 1          | Code 1       | Term 2                        | Code 2       |
|-----------------|--------------|-------------------------------|--------------|
| Breast          | C0006141     | Malignant neoplasm            | C0006826     |
| Breast          | C0006141     | Carcinoma in situ             | C0007099     |
| Breast          | C0006141     | Ductal carcinoma              | C1176475     |
| Breast          | C0006141     | Metaplastic carcinoma         | C1266089     |
| Breast          | C0006141     | Invasive carcinoma            | C1334274     |
| Breast          | C0006141     | Mucinous adenocarcinoma       | C0007130     |
| Breast          | C0006141     | Ductal carcinoma in situ      | C1302731     |
| Breast          | C0006141     | Papillary carcinoma           | C0007133     |
| Breast          | C0006141     | Tubular adenocarcinoma        | C0205645     |
| Breast          | C0006141     | Medullary carcinoma, NOS      | C0206693     |
| Breast          | C0006141     | Cribriform carcinoma          | C0205643     |
| Breast          | C0006141     | Infiltrating carcinoma with ductal and lobular features | C2732747 |
| Breast          | C0006141     | Metastatic carcinoma          | C1384494     |

NOS: Not otherwise specified