American Journal of Epidemiology

Vol. 179, No. 6
DOI: 10.1093/aje/kwt441
Advance Access publication: January 30, 2014

Practice of Epidemiology

Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence

David S. Carrell*, Scott Halgrim, Diem-Thy Tran, Diana S. M. Buist, Jessica Chubak, Wendy W. Chapman, and Guergana Savova

* Correspondence to Dr. David S. Carrell, Group Health Research Institute, Metropolitan Park East, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: carrell.d@ghc.org).

Initially submitted April 2, 2013; accepted for publication September 16, 2013.

The increasing availability of electronic health records (EHRs) creates opportunities for automated extraction of information from clinical text. We hypothesized that natural language processing (NLP) could substantially reduce the burden of manual abstraction in studies examining outcomes, like cancer recurrence, that are documented in unstructured clinical text, such as progress notes, radiology reports, and pathology reports. We developed an NLP-based system using open-source software to process electronic clinical notes from 1995 to 2012 for women with early-stage incident breast cancers to identify whether and when recurrences were diagnosed. We developed and evaluated the system using clinical notes from 1,472 patients receiving EHR-documented care in an integrated health care system in the Pacific Northwest. A separate study provided the patient-level reference standard for recurrence status and date. The NLP-based system correctly identified 92% of recurrences and estimated diagnosis dates within 30 days for 88% of these. Specificity was 96%. The NLP-based system overlooked 5 of 65 recurrences, 4 because electronic documents were unavailable. The NLP-based system identified 5 other recurrences incorrectly classified as nonrecurrent in the reference standard. If used in similar cohorts, NLP could reduce by 90% the number of EHR charts abstracted to identify confirmed breast cancer recurrence cases at a rate comparable to traditional abstraction.

breast cancer recurrence; chart abstraction; natural language processing

Abbreviations: COMBO, Commonly Used Medications and Breast Cancer Recurrence; cTAKES, Clinical Text Analysis and Knowledge Extraction System; EHR, electronic health record; NLP, natural language processing.

Editor's note: *An invited commentary on this article appears on page 759, and the authors' response is published on page 762.*

Medical records have long been an important source of information for epidemiologic studies, and adoption of electronic health record (EHR) systems is increasingly making patients' charts available digitally. Structured EHR data, such as diagnosis and procedure codes, are used extensively in population-based research but capture some conditions unreliably (1). They are inadequate for many important outcomes, such as breast cancer recurrence, that are documented only in unstructured chart notes and reports (2–5). Manual abstraction, the traditional method for extracting information from EHR charts, is time-consuming and expensive and poses inherent risks to patient privacy, limiting the quantity of information available for research.

To address this limitation, natural language processing (NLP)—computational methods for analyzing machine-readable unstructured text—has been used for more than a decade as an alternative or adjunct to manual chart abstraction (6, 7). Some successful applications of NLP include abstracting findings from imaging (8, 9) and pathology reports (10, 11), identifying individuals due for cancer screening (12), clinical trial recruitment (13), identifying postoperative surgical complications (14), and conducting pharmacogenomics and translational science research (15–20). Recent success has also been

reported using NLP to identify breast and prostate malignancies described in pathology reports (21). In some cases, NLP-based algorithms perform as well as, or better than, manual review (9, 12, 22). NLP has been used in epidemiology research to identify statin-induced rhabdomyolysis cases undocumented by structured diagnosis codes, augmenting by 20% the number of cases ascertained (1) and potentially reducing bias. Genetics research consortia, including the Electronic Medical Records and Genomics Network (23, 24) and the PharmacoGenomics Research Network (25), are applying NLP in phenotype definitions.

We hypothesized that an NLP-based system could substantially reduce manual abstraction efforts in large-scale population-based studies when used to identify patients for cohort inclusion without meaningful losses in sensitivity. We explored this hypothesis in a challenging test case: ascertainment of recurrent breast cancer diagnoses for women with early stage invasive breast cancers. This is an outcome of common interest in breast cancer studies and one where case ascertainment relies heavily on manual chart abstraction. Others are exploring structured data algorithms for identifying such recurrences (26, 27). Successfully reducing abstraction burden requires an algorithm with sufficient specificity to obviate manual review of a substantial number of true-negative charts while maintaining sufficient sensitivity to minimize the number of true cases excluded from manual review. Higher sensitivity means greater ascertainment; higher specificity means more efficiency. We report here the results of an NLP-based system we designed to make manual abstraction of breast cancer recurrence status more efficient.

METHODS

Study cohort, outcome, and clinical documents

Our study cohort and outcome definitions were adopted from the Commonly Used Medications and Breast Cancer Recurrence (COMBO) Study (28) conducted at Group Health, an integrated health care delivery system in the Pacific Northwest, from 2007 to 2012. This study used data collection protocols from the Optimizing Breast Cancer Outcomes Study (29, 30) and the Breast Cancer Treatment Effectiveness in Older Women Study (30, 31). It included 2,951 women diagnosed between 1990 and 2009 with early stage (I or II) first primary invasive breast cancer who received care at Group Health. The COMBO investigators abstracted paper charts from 1990 to 2004 and EHRs for 2005–2012 following a written protocol (Web Appendix 1, available at http://aje.oxfordjournals.org/) to identify recurrent and second primary breast cancers. Though no manual abstraction process is infallible, we accepted COMBO data as our patient-level reference standard to develop and evaluate our NLP-based system. The Group Health institutional review board approved this study.

We defined breast cancer recurrence as an ipsilateral, regional, or metastatic breast cancer diagnosis during a follow-up period. Our follow-up period started 120 days after the primary cancer diagnosis date (obtained from the Surveillance, Epidemiology, and End Results Program (32)) and ended on the date of death, disenrollment from the health plan, or completion of COMBO chart review, whichever came first (29–31). "Second primary" breast cancers, defined as cancers in the contralateral breast, were not considered recurrences. The recurrence diagnosis date was defined as the date of definitive pathology findings or, if unavailable, either the date of a radiology report providing evidence of recurrence or the date of a progress note in which a practitioner's diagnosis first appeared. COMBO data also identified whether the cancer was pathologically or clinically confirmed. Clinical confirmation means the diagnosis was based on clinical judgment, often aided by radiology studies, but without positive pathology findings.

We excluded all 562 COMBO subjects with primary breast cancers prior to January 1, 1995, leaving 2,389 available for NLP-based system development or evaluation. We chose this cutoff date because electronic progress notes and radiology reports were not available until November 1998, and median time to clinically confirmed recurrence in our training set was 3.0 years. Thus, excluding women with primary breast cancers before 1995, as opposed to late 1998, allowed us to include more patients and reduce the chance of clinically confirmed recurrences occurring during the pre-EHR period.

We divided our cohort into 3 groups. A training set of 544 was used to develop our NLP system. A test set of 928 was used, once, for system evaluation (reported here). The remaining 917 charts were reserved for future work. So that we could begin NLP-based system development while chart reviews were underway, we selected the training set in random samples over several months, during which time 32%–69% of COMBO chart reviews were completed. We froze the training set at 544 charts after NLP system performance (described below) plateaued, and no new variations in language usage were observed. Because the COMBO Study conducted chart reviews chronologically by primary cancer diagnosis date, women who were diagnosed early in the study period, when care was less likely to be EHR-documented, were overrepresented in the NLP training set. We sampled the test set after COMBO reviews were completed. All sampling was stratified by recurrence status (yes/no) and type (pathologically vs. clinically confirmed).

Clinical documents for all patients were obtained from Group Health EHR systems. They included all available machine-readable pathology reports, progress notes, and radiology reports during patient-specific follow-up periods. Paper reports and reports scanned as images in the EHR were available to the COMBO human review and so are reflected in our reference standard. However, they could not be processed by our NLP system.

NLP software and resources

We developed our system with the open-source Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) (33, 34), Python (35), and SQL (36). cTAKES is an NLP platform with components specifically trained on clinical text. We used cTAKES modules for document sectioning, concept coding, and assertion status annotation. Sectioning identifies headings such as "impression" and "assessment" that provide organization and meaning to document content. For example, mentions of "breast cancer" in a

progress note's "family history" and "assessment" sections have different meanings. Our section annotator was based on that developed by Savova et al. (37). Concept coding uses electronic dictionaries to identify terms of interest, such as "metastatic breast cancer," and associates with each term a standardized ontology/terminology code, such as "C0278488," that can be easily referenced in algorithms. Assertion status annotation determines whether a coded concept is negated (e.g., "no evidence of recurrence") and whether it expresses uncertainty (e.g., "may suggest recurrence"), a historical condition of the patient, or a historical condition of a family member. Assertion status defaults to affirmative if not negated, uncertain, or historical.

We created a cTAKES custom dictionary for all terms and phrases we determined to be relevant to diagnoses of recurrent breast cancer. Its initial entries were gathered from a review of the training corpus and the National Cancer Institute's online vocabulary services (38). We augmented these entries by deriving synonyms, permutations, and abbreviations. For example, we added "br ca" as an abbreviation of "breast cancer" and added "cancer [of the] breast" as a permutation. Our cTAKES dictionary had 1,360 entries for pathology findings (Web Appendix 2 and Web Table 1) and 4,891 for clinical findings (Web Appendix 3 and Web Table 2).

EXPERIMENTAL METHODS

Generally, the approach for developing NLP-based systems with additional phenotyping logic, and the one we followed, is an iterative process informed by theory, experimentation, logic, and domain knowledge. At each iteration, a candidate NLP system is applied to a reference standard training corpus, and its results are evaluated. Evaluation is based on standard measures of performance, including sensitivity and specificity. Discrepancies between the NLP system and the reference standard are investigated through error analysis. The system is then modified in an attempt to reduce errors. Iterative development continues until performance reaches a high level or plateaus. The final system is tested, once, on a test corpus reserved solely for that purpose (39). For additional information and illustrations of applications in other settings, see Nadkarni et al. (6), Savova et al. (40), and Pestian et al. (39).

Breast cancer recurrence NLP system architecture

Modular designs are common in NLP, allowing separate components to be tailored for specific information extraction tasks. Our NLP system's modules reflect the strategy used by COMBO abstractors: they first reviewed pathology reports for evidence of recurrence and then, if not found, reviewed radiology and progress notes. Pathology reports appear infrequently in charts, provide the strongest evidence of recurrence, and are usually linguistically simpler than other clinical text. Radiology reports, and especially progress notes, can be copious, topically diverse, and linguistically nuanced, often including uncertain, hypothetical, and/or historical references. We therefore designed our NLP-based system with a pathology module for processing pathology notes and a clinical module for processing radiology reports and progress notes (combined). We processed all available documents for all subjects and then combined both modules' output to assign each patient a recurrence status and, if pertinent, a diagnosis date (Figure 1).

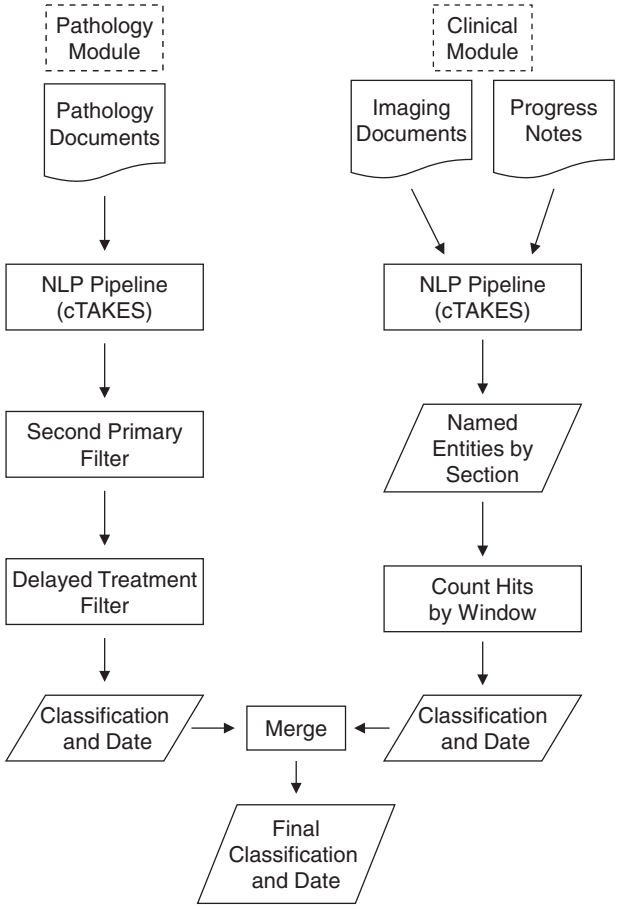

Figure 1. Architecture of the natural language processing (NLP) system used to identify recurrent breast cancer diagnoses among Group Health patients, Pacific Northwest, 1995–2012. cTAKES, Clinical Text Analysis and Knowledge Extraction System.

Pathology module

The pathology module classified a report as positive for recurrence if it met 3 criteria. First, it had to contain the following 3 elements describing breast cancer: an anatomical location related to the breast, such as breast or duct; a malignant disorder, such as cancer or adenocarcinoma; and an indication of severity, such as infiltrating or metastatic. Reports lacking any of these elements were considered negative for recurrence. Second, the report could not refer to the contralateral breast. For example, if the primary was on the right, a report referring to the left breast was presumed to be describing a second primary (discussed above) and considered negative for recurrence. Third, if the pathology report was within 210 days of the primary diagnosis date and mentioned a definitive surgery, such as mastectomy or reexcision, it was

presumed to be discussing delayed treatment of the primary cancer and was considered negative for recurrence. Surgery to treat recurrent disease 210 or more days after primary diagnosis was rare in the training set. The 210-day rule was chosen because it addressed most false positives due to delayed surgeries in the training set, while minimizing false negatives. Reports meeting all 3 criteria were considered positive for recurrence, and the report date of the earliest positive report was assigned as the recurrence diagnosis date.

Clinical module

The clinical module considered affirmative mentions of breast cancer recurrence—including metastatic disease—in progress notes and radiology reports to determine whether and when recurrence was diagnosed. Because clinical diagnosis often requires assimilating indirect evidence, such as change in imaging findings over time, language describing it can be speculative, a common challenge in clinical NLP. We therefore designed the clinical module's status annotator to recognize a broad range of uncertainty cues appearing in the training corpus (e.g., the cue "question" in the sentence "The question is whether or not this could represent recurrence of breast cancer."). However, charts of some women with clinically confirmed recurrence contained only mentions qualified by uncertainty. To avoid misclassifying such charts as nonrecurrent, we iteratively performed experiments on the training set to identify uncertainty-qualified mentions that were semantically equivalent to affirmative mentions. The phrase "suggests metastatic breast cancer" is illustrative. Though expressing uncertainty, it typically appears when clinical evidence tips in favor of diagnosis. These changes improved sensitivity but, as expected, degraded specificity, a frequent trade-off during NLP development.

We used 2 common strategies to address this degraded specificity. First, we restricted valid mentions to specific combinations of dictionary terms, document sections, and document types. For example, "metastatic breast cancer" in the "diagnosis" section of a progress note was considered valid, but the same mention in a progress note's "subjective" section was not, because only the former was likely to correspond to an actual diagnosis. We call these valid combinations of terms, sections, and document types "hits." Hits were derived empirically from the training corpus. A complete list of hits is provided in Appendix 3 and Web Tables 2–5.

Our second strategy involved splitting the chart into 30-day windows defined by rolling 15-day offsets throughout the follow-up period. Our analysis of the training corpus confirmed the intuitive: a pronounced increase in frequency of hits occurred near the diagnosis date, as illustrated in Figure 2. Therefore, we classified a chart as recurrent only if it had at least 3 hits in a 30-day window. This rule achieved the best balance between sensitivity and specificity. The module assigns the date of the earliest document with a hit in a window as the recurrence date. If no window during follow-up had at least 3 hits, the module classified the chart as nonrecurrent.

Rule-based NLP system development typically starts by defining the broadest rules and then proceeds to improve results by adding rules addressing more specific cases. One such case for us involved the word "metastasis." Women in our cohort had stage I or II primary cancers, which do not involve distant metastases. Accordingly, chart mentions of metastases overwhelmingly indicated recurrence. Occasionally, a stage II primary cancer was described as metastatic to regional lymph nodes. In such charts, subsequent metastasis mentions may refer to the primary cancer or a recurrence. We addressed this ambiguity as follows: if a chart contained any mention of positive lymph node involvement, then mentions of metastasis in a note's "history" section were excluded from the set of hits. This rule improved recurrence status classification in the training set.

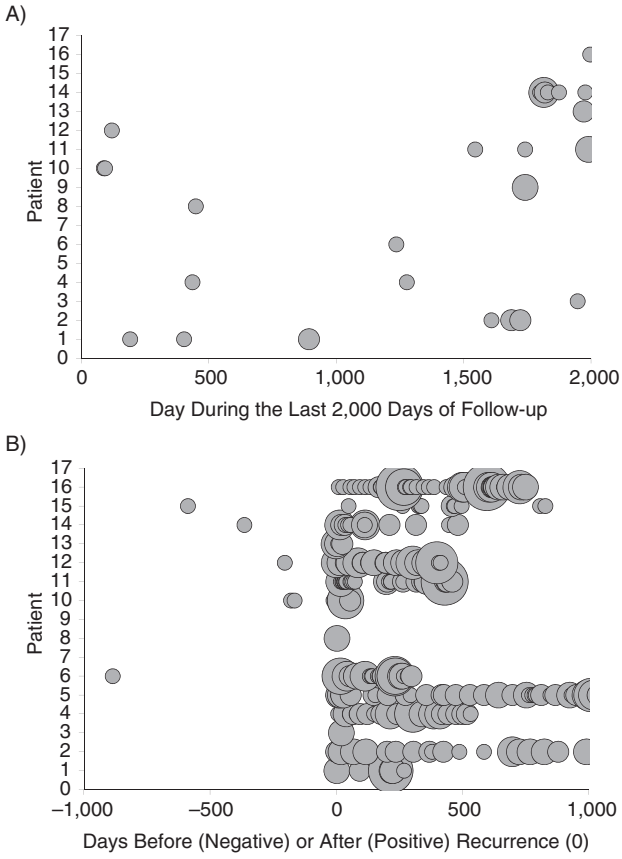

Figure 2. Breast cancer recurrence hits per calendar day for Group Health patients, Pacific Northwest, 1995–2012, found by the natural language processing system's clinical module in electronic charts for random samples of A) 16 breast cancer patients without recurrence, and B) 16 breast cancer patients with clinically confirmed recurrence.

Assigning overall recurrence status and date

In a final step called postprocessing, we combined the output of the pathology and clinical modules to produce an overall recurrence status and date for each chart. When neither module was positive for recurrence, the chart was classified as nonrecurrent. When either or both modules were positive, the chart was classified as recurrent; the recurrence date was established by the pathology module, if available, and the

clinical module otherwise. The pathology module's date was preferred because pathological evidence is more definitive.

Evaluation. Our main performance measures were specificity and sensitivity. Specificity conveys the potential benefit of our approach, indicating the percentage of nonrecurrent charts that will not require manual review. From sensitivity, we calculated the percentage of truly recurrent cases that would be missed if the system determined which charts were reviewed (1.00 − sensitivity). From positive predictive value (PPV), we calculated the percentage of cases that would be reviewed manually and be negative for recurrence (1.00 − PPV). Whereas false positives will be reclassified as negative upon human review, false negatives represent true cases that would be overlooked entirely, potentially introducing bias.

We calculated the *F* score, which is the harmonic mean of PPV and sensitivity [(2 × PPV × sensitivity) / (PPV + sensitivity)], because it provides a balanced measure of overall performance and is a common evaluation metric in the NLP field (41). We also computed a date-weighted *F* score, which is like the *F* score but weighs each true positive by the proximity of the system-assigned recurrence date to the reference standard date (Web Appendix 4).

We evaluated the NLP-based system against the reference standard 1) as documented by chart abstractors, and 2) as modified by an independent review of some charts where the system and the reference standard were discordant. Discordant recurrence statuses and dates were reviewed as follows. First, presuming the reference standard was correct, an informaticist conducted error analysis to categorize the NLP system's error. Then, charts with discordances not so resolved were independently reviewed by a trained abstractor following the COMBO protocol, yielding a recurrence status, date, and explanation for the NLP and/or reference standard errors (as relevant). We used the original reference standard to evaluate the system's ability to replicate human abstraction and the modified one to assess NLP performance relative to human abstraction. We did not review concordant charts because of the high cost and because errors therein were less likely and would not affect measures of relative performance.

RESULTS

Women in the training and test sets were comparable with respect to primary cancer diagnostic characteristics (cancer stage, cancer grade, node positivity) and treatment characteristics (surgery, radiation therapy, chemotherapy, hormone therapy; Table 1). Women in the test set tended to have primary cancers later in the study period because of the pragmatic sampling scheme explained above, corresponding to a later median recurrence date, shorter follow-up, and lower recurrence rate in the test set (Table 1). This also accounted for the test set's reduced likelihood of having any clinical notes in the pre-EHR era (Table 2). Nine percent of women in the test set had some follow-up in the period when electronic radiology and progress notes were unavailable, creating the possibility that some chart-documented recurrences would be impossible to discover via NLP (Table 2).

Confusion matrices for results of the NLP-based system are shown in Table 3; performance statistics are shown in Table 4. In the training set, the system achieved 0.93 sensitivity, 0.95 specificity, 0.76 PPV, and a date-weighted *F* score of 0.67. Against the original reference standard test set, the system achieved 0.92 sensitivity, 0.96 specificity, 0.59 PPV, and date-weighted *F* score of 0.59. Reduced PPV and *F* scores in the test set are at least partly a function of lower prevalence of recurrence in the test set compared with the training set (6.47% vs. 13.97%, Table 1). Assigned recurrence dates were within 14 days of the actual recurrence date for 76.1% and 74.5% of the true positives in the training and (original) test sets, respectively; 80.1% and 83.6% were within 30 days, respectively. If this system had been used to select charts for manual abstraction in the COMBO Study, the number of charts reviewed would have been reduced from 928 to 93, a reduction of 90%. This benefit would come at the expense of an 8% (5 of 60) loss of true cases based on the original reference standard.

The system performed better against the adjudicated reference standard than the original reference standard (Table 4). Adjudication resulted in 5 nonrecurrent charts in the original reference standard being reclassified as recurrent. In 4 of these, the original human abstractors had overlooked evidence of recurrence that the NLP-based system detected. After adjudication, the NLP-based system's false negative errors were equal in number to those overlooked by human review (5 each). Adjudication also adjusted the recurrence date of 2 charts for which the original reference standard's dates were off by more than 30 days from the adjudicated date as opposed to a difference of 1 or fewer days on the part of the NLP system (Web Appendix 5 and Web Table 6).

Error analysis of results against the adjudicated reference standard showed that 4 of the system's 5 false negatives were due to relevant clinical documents not being available electronically; the fifth occurred because the pathology module's dictionary failed to include cancer grading terms (e.g., "poorly differentiated") as indications of disease severity.

Against the adjudicated reference standard, the system had 32 false positive errors, which reduce efficiency but do not introduce bias when used to identify charts for manual review. Many of these ($n = 12$, 38%) were due to status annotation errors, such as failing to recognize rule-out diagnoses, differential diagnoses, or distant negation cues. Some other false positives were caused by misinterpreting metastases or recurrences of cancers from nonbreast primaries ($n = 3$).

DISCUSSION

We designed this study to test whether an NLP-based system could be used to increase the efficiency of EHR-based research. We chose recurrent breast cancer as a test case because it is a scientifically important outcome traditionally ascertained at substantial cost by manual review. Our results demonstrate that NLP can identify recurrent breast cancer diagnoses with high sensitivity and specificity. Used to identify charts requiring manual review, NLP could substantially reduce reviewer burden and patient privacy risk with minimal loss of true cases. Our experiment achieved a 90% reduction in charts requiring review with an 8% loss of true cases, and the number of recurrent cases misclassified by NLP was equal to the number misclassified by manual review. Four of the 5 NLP false negative errors were due to documents

Table 1. Characteristics of Group Health Study Patients and Their Breast Cancer Events by Natural Language Processing Training Set and Test Set Assignment, 1995–2012

Characteristic	All ($n = 1,472$)		NLP Training Set ($n = 544$)		NLP Test Set ($n = 928$)	
	No.	%	No.	%	No.	%
Age at primary diagnosis, years	63.0 (13.3)[a]		62.0 (13.2)[a]		63.6 (13.3)[a]	
Primary diagnosis date	March 2005[b]		February 2003[b]		February 2006[b]	
Follow-up duration, years	5.2 (3.3–7.2)[c]		6.6 (4.3–7.6)[c]		4.6 (3.0–6.3)[c]	
Breast cancer recurrence during follow-up[d]	136	9.24	76	13.97	60	6.47
Year of primary diagnosis						
1995–1998	191	13.0	103	18.9	88	9.5
1999–2005	965	65.6	440	80.9	525	56.6
2006–2009	316	21.5	1	0.2	315	33.9
Stage[e]						
Local	861	58.5	312	57.4	549	59.2
Regional	611	41.5	232	42.7	379	40.8
Grade[e]						
Well differentiated	366	24.9	136	25.0	230	24.8
Moderately differentiated	562	38.2	209	38.4	353	38.0
Poorly differentiated	422	28.7	142	26.1	280	30.1
Undifferentiated	17	1.2	6	1.1	11	1.2
Unknown	105	7.1	51	9.4	54	5.8
Positive node[e]						
No mention	130	8.8	43	7.9	87	9.4
No nodes examined	421	28.6	209	38.4	212	22.8
All negative	546	37.1	141	25.9	405	43.6
Positive	375	25.5	151	27.7	224	24.1

Table continues

not being available electronically, a problem that may diminish as EHR adoption increases.

The acceptability of an NLP-based system's error rates depends on the research application. If the NLP-based system is intended to enhance the efficiency of manual review, false negative error rates comparable to those for manual review (e.g., 8% in the COMBO Study) are acceptable, and false positive errors are less concerning because their only "cost" is reducing efficiency by increasing the number of charts reviewed. For other purposes, such as estimating event rate trends, error tolerance may be lower. The tradeoffs between sensitivity and specificity made when developing an NLP system have implications for potential biases and should be carefully considered in a given application (42).

We believe the NLP methods illustrated here may be useful for identifying other clinical outcomes, including recurrence of other cancers. Further adoption of EHRs, coupled with improvements in NLP technologies, may make it feasible to incorporate recurrence outcomes into population-based surveillance efforts such as the Surveillance, Epidemiology, and End Results program. In such contexts NLP may be used as a substitute for some manual abstraction tasks and as an adjunct to others requiring manual confirmation.

Algorithms like ours can be tailored to meet particular study design objectives (42). For example, to support fully automated EHR-driven genomics research (24, 43), where high PPV is required but sensitivity is optional, a modified system could require cases to have more hits in smaller time windows and controls to have 0 hits throughout follow-up. Alternatively, a system incorporating manual review and needing very high sensitivity could require fewer hits in longer time windows.

Several limitations of this work should be noted. First, our NLP modules may require adaptation to accommodate language usage and document sectioning in other institutional settings. Second, NLP development costs limit its applicability to large or repeated tasks where it is cost effective relative to 100% manual abstraction. Third, NLP requires access to machine-readable clinical text and does not work with print documents or their scanned copies. Although expanding EHR adoption may reduce this limitation in the future, assessing document availability is an important early step in any NLP project. Fourth, our study cohort was limited to women with early stage (I or II) breast cancers; the algorithm has not been tested for recurrence in women with initial late stage disease or ductal carcinoma in situ (stage 0). Fifth,

Table 1. Continued

Characteristic	All ($n = 1,472$)		NLP Training Set ($n = 544$)		NLP Test Set ($n = 928$)	
	No.	%	No.	%	No.	%
Surgery[e]						
No surgery	2	0.1	0	0.0	2	0.2
Lumpectomy	914	62.1	343	63.1	571	61.5
Mastectomy	556	37.8	201	37.0	355	38.3
Radiation therapy[e]						
No	145	10.4	63	12.1	82	9.4
Yes	948	67.9	361	69.2	587	67.2
Unknown	303	21.7	98	18.8	205	23.5
Chemotherapy[e]						
No	628	42.7	220	40.4	408	44.0
Yes	544	37.0	205	37.7	339	36.5
Unknown	300	20.4	119	21.9	181	19.5
Hormone therapy[e]						
No	449	30.5	166	30.5	283	30.5
Yes	869	59.0	323	59.4	546	58.8
Unknown	154	10.5	55	10.1	99	10.7

Abbreviation: NLP, natural language processing.
[a] Value expressed as mean (standard deviation).
[b] Value expressed as median month and year.
[c] Value expressed as median (interquartile range).
[d] Based on the original reference standard classification for breast cancer recurrence.
[e] Ascertained by data from the Surveillance, Epidemiology, and End Results Program.

Table 2. Quantity and Availability of Machine Readable Clinical Documents for Group Health Study Patients by Training Set and Test Set Assignment, 1995–2012

Clinical Document	NLP Training Set ($n = 544$)			NLP Test Set ($n = 928$)		
	No.	%	Median (IQR)	No.	%	Median (IQR)
Pathology reports						
No. of reports	1,991			2,409		
Patients with any reports	430	79		696	75	
Reports per patient[a]			4 (2–6)			3 (1–4)
Progress notes						
No. of notes	104,017			200,163		
Patients with any notes	534	98		916	98	
Notes per patient[a]			156.5 (77–266)			177 (110–281)
Radiology reports						
No. of reports	21,973			39,080		
Patients with any reports	532	98		914	98	
Reports per patient[a]			37 (25–53)			40 (27–54)
Progress notes and radiology reports						
No. of notes/reports	125,990			239,243		
Patients with any notes/reports	536	99		916	98	
Notes/reports per patient[a]			194 (107–314)			220 (144–332)
Patients with follow-up before 1989[b]	102	19		85	9	

Abbreviations: IQR, interquartile range; NLP, natural language processing.
[a] Based on patients with any documents.
[b] Prior to November 1998, progress notes and radiology reports were not available in machine-readable format.

Table 3. Confusion Matrices for Natural Language Processing System Results Classifying Group Health Patients as Recurrent or Nonrecurrent in the NLP Training Set, the NLP Test Set According to the Reference Standard, and the NLP Test Set According to the Corrected Reference Standard, 1995–2012

NLP System Result	NLP Training Set			NLP Test Set					
				Reference Standard			Corrected Reference Standard[a]		
	Recurrent No.	Nonrecurrent No.	Total No.	Recurrent No.	Nonrecurrent No.	Total No.	Recurrent No.	Nonrecurrent No.	Total No.
Recurrent	71	22	93	55	38	93	60	32	92
Nonrecurrent	5	446	451	5	830	835	5	830	835
Total	76	468	544	60	868	928	65	862	927

Abbreviation: NLP, natural language processing.

[a] The corrected reference standard reflects corrections to the original reference standard based on an independent review of charts where the original reference standard and the NLP system produced discordant results for patients' recurrence status and/or recurrence date.

reference standard corrections were limited to the review of charts where NLP and the reference standard were discordant. Reviews of concordant charts may have revealed additional errors. However, because NLP and the reference standard agreed on these charts, any errors so discovered would not affect assessment of NLP system performance relative to manual abstraction.

Future work should include testing the NLP-based system's performance in other institutional settings and incorporating machine learning methods to enhance accuracy of status annotations (e.g., negation, uncertainty). Future work should also explore combining NLP-based methods with structured data algorithms based on diagnosis, procedure, and medication codes (26). Such hybrid approaches are successful in other domains (23, 24) and may reduce errors. For example, if clinical documents describing an outcome are ambiguous or incomplete, structured data may clarify the situation.

CONCLUSION

The NLP-based recurrent breast cancer detection system we present demonstrates the feasibility of improving the efficiency of research efforts requiring manual abstraction from patient charts. This work contributes to the converging evidence that NLP can be used to determine which charts should receive expensive and time-consuming manual review, and NLP can scale to large sample sizes at very low additional

Table 4. NLP System Results for Classifying and Assigning Dates to Group Health Patients with Recurrent Breast Cancer in the NLP Training Set, the NLP Test Set According to the Reference Standard, and the NLP Test Set According to the Corrected Reference Standard, 1995–2012

Patient Set	Sensitivity	Specificity	PPV	*F* Score	Date-Weighted *F* Score[a]	Recurrence Date Within 14 Days		Recurrence Date Within 30 Days		Manual Review Reduction[c]		True Recurrence Cases Lost[e]	
						Ratio[b]	%	Ratio[b]	%	Ratio[d]	%	Ratio[f]	%
NLP training set	0.93	0.95	0.76	0.84	0.67	54/71	76.1	57/71	80.1	451/544	82.9	5/76	6.6
NLP test set, original reference standard	0.92	0.96	0.59	0.72	0.59	41/55	74.5	46/55	83.6	835/928	90.0	5/60	8.3
NLP test set, corrected reference standard[g]	0.92	0.96	0.66	0.76	0.66	47/60	78.3	53/60	88.3	835/927	90.0	5/65	7.7

Abbreviations: NLP, natural language processing; PPV, positive predictive value.

[a] Date-weighted *F* score is calculated like *F* score but weighs each true positive by the proximity of the NLP system–assigned date to the reference standard date (as described in Web Appendix 4).

[b] This ratio is the number of patients for which the NLP system assigned a date within the specified time period (numerator) divided by the number of patients correctly classified by the NLP system as recurrent (denominator, from Table 3).

[c] The number of patients for which manual review could be avoided if the NLP system were used in a comparable study to select patients for cohort inclusion.

[d] This ratio is the number of patients for which manual review could be avoided if the NLP system were used in a comparable study (numerator) divided by the number of patients otherwise eligible for study inclusion (denominator).

[e] The number of true recurrence cases that would be overlooked if the NLP system were used in a comparable study to select patients for cohort inclusion.

[f] This ratio is the number of true recurrence cases that would be overlooked if the NLP system were used in a comparable study (numerator) divided by the number of true recurrence cases in the study cohort (denominator).

[g] The corrected reference standard reflects corrections to the original reference standard based on an independent review of charts where the original reference standard and the NLP system produced discordant results for patients' recurrence status and/or recurrence date.

cost. Using freely available open-source NLP software to process free-text pathology reports, radiology reports, and progress notes, we assigned breast cancer recurrence status and dates to patients with known primary cancers. Our results indicate that NLP, in conjunction with manual review, could identify confirmed cases of recurrent breast cancer at a rate comparable to traditional abstraction with up to a 90% reduction in the number of charts requiring manual review.

ACKNOWLEDGMENTS

Author affiliations: Group Health Research Institute, Seattle, Washington (David S. Carrell, Scott Halgrim, Diem-Thy Tran, Diana S.M. Buist, Jessica Chubak); Division of Biomedical Informatics, Department of Medicine, University of California San Diego, San Diego, California (Wendy W. Chapman); Informatics Program, Boston Children's Hospital, Boston, Massachusetts (Guergana Savova); Department of Pediatrics, Harvard Medical School, Harvard University, Boston, Massachusetts (Guergana Savova); and Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington (Jessica Chubak).

This work was supported by the National Cancer Institute of the National Institutes of Health (grants RC1 CA146917, R01 CA093772, and R01 CA120562) and the American Cancer Society (grant CRTG-03-024-01-CCE).

We thank Dr. Michael VonKorff for his advice throughout this project. We also thank Dr. Denise Boudreau, Principal Investigator of the COMBO Study, for sharing with us COMBO data that provided our patient-level reference standard.

Conflict of interest: Guergana K. Savova is on the Advisory Board of Wired Informatics, LLC, which provides services and products for clinical natural language processing applications.

REFERENCES

1. Floyd JS, Heckbert SR, Weiss NS, et al. Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. *JAMA*. 2012;307(15):1580–1582.
2. Dean BB, Lam J, Natoli JL, et al. Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev*. 2009;66(6): 611–638.
3. Hicks J. *The Potential of Claims Data to Support the Measurement of Health Care Quality. Policy Analysis*. Santa Monica, CA: RAND Graduate School; 2003:272.
4. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128–144.
5. Jha AK. The promise of electronic records: Around the corner or down the road? *JAMA*. 2011;306(8):880–881.
6. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544–551.
7. Friedman C, Hripcsak G. Natural language processing and its future in medicine. *Acad Med*. 1999;74(8):890–895.
8. Hripcsak G, Austin JH, Alderson PO, et al. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology*. 2002;224(1):157–163.
9. Chapman WW, Fizman M, Chapman BE, et al. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. *J Biomed Inform*. 2001;34(1): 4–14.
10. Crowley RS, Castine M, Mitchell K, et al. caTIES: 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(3):253–264.
11. Coden A, Savova G, Sominsky I, et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *J Biomed Inform*. 2009;42(5):937–949.
12. Denny JC, Choma NN, Peterson JF, et al. Natural language processing improves identification of colorectal cancer testing in the electronic medical record. *Med Decis Making*. 2012; 32(1):188–197.
13. Seyfried L, Hanauer DA, Nease D, et al. Enhanced identification of eligibility for depression research using an electronic medical record search engine. *Int J Med Inform*. 2009;78(12):e13–e18.
14. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306(8):848–855.
15. Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther*. 2011;89(3):379–386.
16. Perlis RH, Iosifescu DV, Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med*. 2012; 42(1):41–50.
17. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2010;62(8):1120–1127.
18. Savova GK, Olson JE, Murphy SP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. *J Am Med Inform Assoc*. 2012;19(e1):e83–e89.
19. Kullo IJ, Fan J, Pathak J, et al. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc*. 2010;17(5):568–574.
20. Cheng LT, Zheng J, Savova GK, et al. Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging*. 2010;23(2): 119–132.
21. Strauss JA, Chao CR, Kwan ML, et al. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J Am Med Inform Assoc*. 2013;20(2): 349–355.
22. Hripcsak G, Friedman C, Alderson PO, et al. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*. 1995;122(9):681–688.
23. Peissig PL, Rasmussen LV, Berg RL, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc*. 2012; 19(2):225–234.
24. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011;3(79):79re1.

25. Pharacogenomics Research Network. Pharmacogenomics of rheumatoid arthritis therapy. http://pgrn.org/display/pgrnwebsite/PhRAT+Profile. Updated April 2011. Accessed August 23, 2013.
26. Chubak J, Yu O, Pocobelli G, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *J Natl Cancer Inst*. 2012;104(12):931–940.
27. Hassett MJ, Ritzwoller DP, Taback N, et al. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using 2 large contemporary cohorts [published online ahead of print December 6, 2012]. *Med Care*.
28. Wirtz HS, Buist DS, Gralow JR, et al. Frequent antibiotic use and second breast cancer events. *Cancer Epidemiol Biomarkers Prev*. 2013;22(9):1588–1599.
29. Buist DS, Chubak J, Prout M, et al. Referral, receipt, and completion of chemotherapy in patients with early-stage breast cancer older than 65 years and at high risk of breast cancer recurrence. *J Clin Oncol*. 2009;27(27):4508–4514.
30. Chubak J, Buist DS, Boudreau DM, et al. Breast cancer recurrence risk in relation to antidepressant use after diagnosis. *Breast Cancer Res Treat*. 2008;112(1):123–132.
31. Enger SM, Thwin SS, Buist DS, et al. Breast cancer treatment of older women in integrated health care settings. *J Clin Oncol*. 2006;24(27):4377–4383.
32. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. http://seer.cancer.gov/. Accessed August 23, 2013.
33. The Apache Software Foundation. Apache cTAKES 3.0 component use guide. https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.0+Component+Use+Guide. Accessed August 23, 2013.
34. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507–513.
35. Python Software Foundation. Python documentation. http://www.python.org/doc/. Accessed August 23, 2013.
36. Hotek M. Microsoft SQL Server 2008. Redmond, WA: Microsoft Press; 2009.
37. Savova GK, Fan J, Ye Z, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. *AMIA Annu Symp Proc*. 2010;2010:722–726.
38. National Cancer Institute. National Cancer Institute enterprise vocabulary services. http://evs.nci.nih.gov/. Accessed August 23, 2013.
39. Pestian JP, Deleger L, Savova GK, et al. Natural language processing—the basics. In: Hutton JJ, ed. *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*. New York, NY: Springer; 2012:149–172.
40. Savova GK, Deleger L, Solti I, et al. Natural language processing: applications in pediatric research. In: Hutton JJ, ed. *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*. New York, NY: Springer; 2012:173–192.
41. Jurafsky D, Martin JH. *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Pearson Prentice Hall; 2009.
42. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol*. 2012;65(3):343–349.e2.
43. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet*. 2011;12(6):417–428.