Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm

Justin A Strauss,[1] Chun R Chao,[1] Marilyn L Kwan,[2] Syed A Ahmed,[3] Joanne E Schottinger,[4] Virginia P Quinn[1]

[1] Kaiser Permanente Southern California, Research and Evaluation, Pasadena, California, USA
[2] Kaiser Permanente Northern California, Division of Research, Oakland, California, USA
[3] Southern California Permanente Medical Group, Genetic Services, Riverside, California, USA
[4] Southern California Permanente Medical Group, Quality and Clinical Analysis, Pasadena, California, USA

Correspondence to
Dr Virginia P Quinn, 100 S. Los Robles, 4th Floor, Pasadena, CA 91101, USA; virginia.p.quinn@kp.org

The HMO Cancer Research Network (CRN) consists of the research programs, enrollee populations and databases of 14 HMO members of the HMO Research Network. The overall goal of the CRN is to conduct collaborative research to determine the effectiveness of preventive, curative, and supportive interventions for major cancers that span the natural history of those cancers among diverse populations and health systems. The 14 health plans, with nearly 11 million enrollees, are distinguished by their longstanding commitment to prevention and research, and collaboration among themselves and with affiliated academic institutions.

Received 2 March 2012
Accepted 26 June 2012
Published Online First
2 August 2012

ABSTRACT

Objective Significant limitations exist in the timely and complete identification of primary and recurrent cancers for clinical and epidemiologic research. A SAS-based coding, extraction, and nomenclature tool (SCENT) was developed to address this problem.

Materials and methods SCENT employs hierarchical classification rules to identify and extract information from electronic pathology reports. Reports are analyzed and coded using a dictionary of clinical concepts and associated SNOMED codes. To assess the accuracy of SCENT, validation was conducted using manual review of pathology reports from a random sample of 400 breast and 400 prostate cancer patients diagnosed at Kaiser Permanente Southern California. Trained abstractors classified the malignancy status of each report.

Results Classifications of SCENT were highly concordant with those of abstractors, achieving κ of 0.96 and 0.95 in the breast and prostate cancer groups, respectively. SCENT identified 51 of 54 new primary and 60 of 61 recurrent cancer cases across both groups, with only three false positives in 792 true benign cases. Measures of sensitivity, specificity, positive predictive value, and negative predictive value exceeded 94% in both cancer groups.

Discussion Favorable validation results suggest that SCENT can be used to identify, extract, and code information from pathology report text. Consequently, SCENT has wide applicability in research and clinical care. Further assessment will be needed to validate performance with other clinical text sources, particularly those with greater linguistic variability.

Conclusion SCENT is proof of concept for SAS-based natural language processing applications that can be easily shared between institutions and used to support clinical and epidemiologic research.

INTRODUCTION

The identification of primary and recurrent cancer diagnoses is critical to clinical and epidemiologic research. Despite its importance, however, there is substantial lag between the time of primary cancer diagnoses and complete information capture by cancer registries. Additionally, many registries do not track cancer recurrences. Consequently, researchers frequently rely on manual chart review and medical claims data, such as International Classification of Diseases (ICD) codes, to identify primary and recurrent cancers. Chart review, while accurate, is often not feasible in large-scale studies. Conversely, claims data are inexpensive and scalable to large studies but are unreliable in terms of accuracy.

To address ongoing needs for improved identification of cancer diagnoses, a SAS-based coding, extraction, and nomenclature tool (SCENT) was developed at Kaiser Permanente Southern California (KPSC). KPSC is an integrated healthcare organization that provides medical services to a diverse membership of more than 3.5 million people throughout Southern California. Research conducted at KPSC directly impacts practice guidelines and the medical care that patients receive. Each year, approximately 20 000 new cancer cases are diagnosed at KPSC. Given the threat to patients' health and quality of life, as well as the impact on their families, prevention and treatment of cancer are high priorities for clinicians and researchers. SCENT was developed to support these critical efforts.

BACKGROUND AND SIGNIFICANCE

In 1999, Warren *et al* analyzed the medical claims data of 6784 Medicare patients and concluded that such data have limited value in accurately identifying breast cancer cases.[1] More recently, Lamont *et al* achieved high levels of sensitivity and specificity using claims data to identify cancer recurrence.[2] However, that validation used a small sample of patients (N=45) and the accuracy of claims data in identifying cancer recurrence has yet to be well established.

The use of electronic medical record (EMR) systems among healthcare providers has increased significantly over the past decade. Additionally, a provision in the 2009 American Recovery and Reinvestment Act may accelerate the adoption of electronic systems by providing financial incentives for the meaningful use of technology in healthcare delivery.[3] According to an annual survey by the Centers for Disease Control and Prevention, the use of EMRs by office-based physicians increased from 18% to 57% between 2001 and 2011.[4] Use of EMR systems complete with all basic functionality also rose from 11% to 34% between 2006 and 2011. The transition from paper to EMRs not only reduces medical errors and improves communication between providers, but also increases the value and feasibility of medical informatics applications.

Using text from EMRs, natural language processing (NLP) has the potential to supplement or replace manual chart review and electronic diagnosis codes in identifying primary and recurrent cancers. A number of studies have already demonstrated the utility of NLP in coding and extracting information from clinical text.[5–7] However, few epidemiologic studies have employed the technology. Slow adoption cannot be entirely attributed to the technology's recentness, as working implementations have existed for nearly two decades.[8 9] More likely, adoption has been limited by requirements for integration with clinical data systems, technical complexity, and habitual use of medical claims data.

In 2011, Chapman *et al* noted that despite continued improvements in NLP performance, the technology is rarely employed in clinical research settings.[10] This is attributed, in part, to a lack of focus on the end-user during development, specifically as it relates to implementation costs and customizability. Furthermore, NLP functionality in EMR and other clinical systems software currently provides limited value to researchers due to lack of customizability and inability to be readily shared between collaborating institutions. SCENT was developed to increase the attractiveness of NLP to clinical and epidemiologic researchers by reducing implementation barriers and ensuring accessibility in collaborative multi-site research.

MATERIALS AND METHODS

Overview

SCENT uses functionality from SAS Base (V.9.2, SAS Institute) and does not require the Text Miner add-on module. Components of SCENT consist primarily of SAS macro libraries and collections of Excel support files. An overview of the processes employed by SCENT can be seen in figure 1. Hierarchical classification rules are used by SCENT to analyze electronic pathology report text. Rules-based NLP approaches have been used effectively in previous studies[11 12] and Informatics for Integrating Biology and the Bedside challenges.[13–15] While SCENT has the flexibility to assign codes to electronic text using a number of different coding systems, SNOMED 3.x was initially selected due to its use by the CoPathPlus (V.3.2; Cerner DHT, Inc., Waltham, Massachusetts, USA) laboratory information system at KPSC. CoPathPlus features synoptic reporting, a process by which pathologists provide structured results using predefined cancer checklists.[16] Results are assigned SNOMED 3.x codes by CoPathPlus, which can be reviewed and subsequently modified by reporting pathologists.

For clinical concepts to be matched by SCENT, their keywords must be found, without negation, in proximate distance to each other. SCENT assigns SNOMED codes associated with matched concepts to examined pathology reports. Words within reports that contribute to concept matches are tagged with relevant information, such as disease extent and SNOMED code. SCENT examines text surrounding concept matches to differentiate clinical suspicions from conclusive findings and to determine disease extent (eg, non-invasive, invasive, or metastatic). Additional diagnostic information, including tumor stage and Gleason score, is also extracted.

Clinical concepts

SCENT relies on a dictionary of approximately 1000 clinical concepts and associated SNOMED 3.x codes related to morphology, anatomic site, and procedural type. The malignancy potential of each morphology concept was classified by up to four physicians with expert pathology or oncology knowledge. Dictionary concepts are tokenized into their component words and enhanced with regular expression logic to account for synonyms and plural words. For example, the 'intraductal papillary adenocarcinoma' concept is broken into three distinct words and 'intraductal' is replaced with '((intra)?duct(al)?)' to match mentions of intraductal, ductal, or duct. In addition to regular expressions that are manually added to

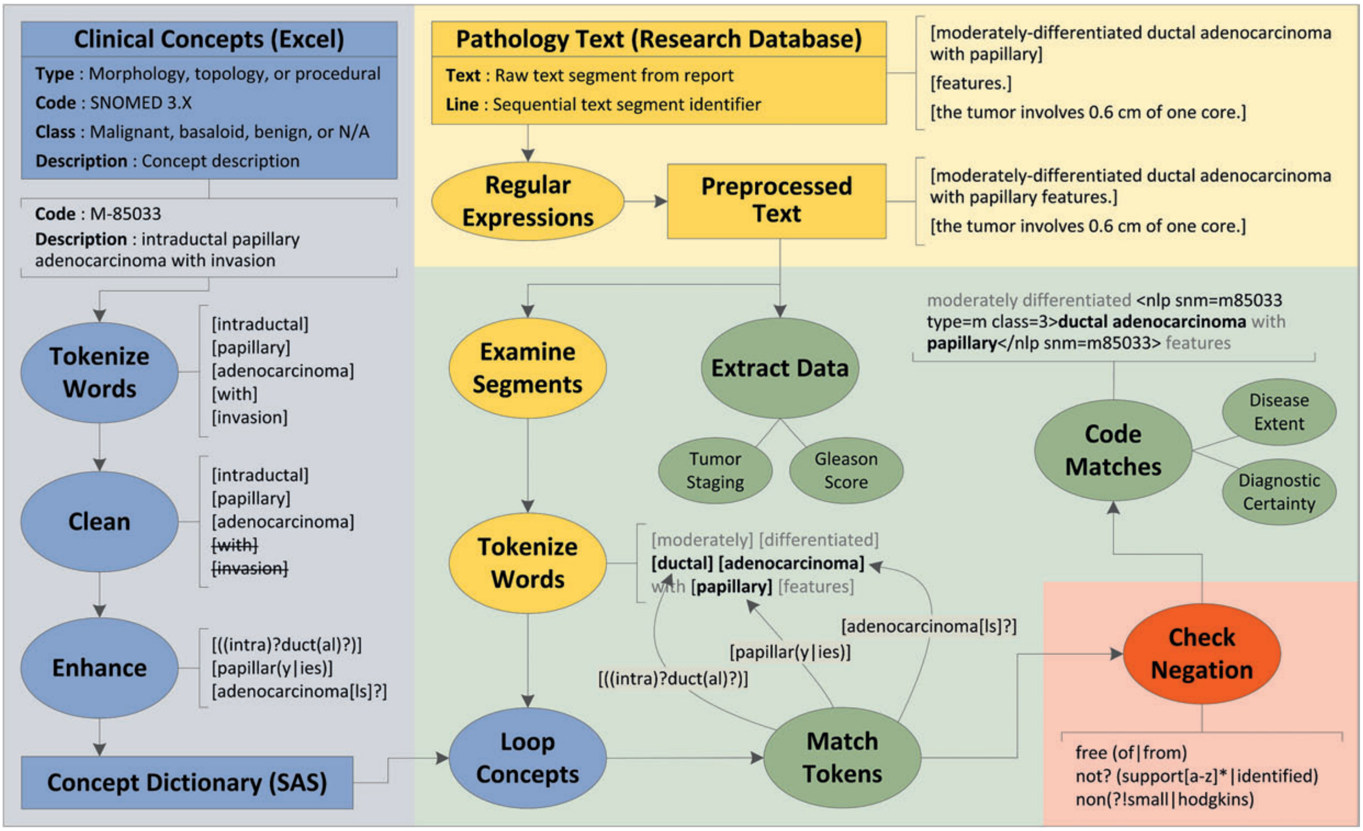

Clinical Concepts (Excel)
Type : Morphology, topology, or procedural
Code : SNOMED 3.X
Class : Malignant, basaloid, benign, or N/A
Description : Concept description
Code : M-85033
Description : intraductal papillary adenocarcinoma with invasion
Tokenize Words
[intraductal]
[papillary]
[adenocarcinoma]
[with]
[invasion]
Clean
[intraductal]
[papillary]
[adenocarcinoma]
~~[with]~~
~~[invasion]~~
Enhance
[((intra)?duct(al)?)]
[papillar(y|ies)]
[adenocarcinoma[ls]?]
Concept Dictionary (SAS)
Pathology Text (Research Database)
Text : Raw text segment from report
Line : Sequential text segment identifier
[moderately-differentiated ductal adenocarcinoma with papillary]
[features.]
[the tumor involves 0.6 cm of one core.]
Regular Expressions
Preprocessed Text
[moderately-differentiated ductal adenocarcinoma with papillary features.]
[the tumor involves 0.6 cm of one core.]
Examine Segments
Extract Data
Tumor Staging
Gleason Score
moderately differentiated <nlp snm=m85033 type=m class=3>ductal adenocarcinoma with papillary</nlp snm=m85033> features
Code Matches
Disease Extent
Diagnostic Certainty
Tokenize Words
[moderately] [differentiated]
[ductal] [adenocarcinoma]
with [papillary] [features]
[adenocarcinoma[ls]?]
[papillar(y|ies)]
[((intra)?duct(al)?)]
Check Negation
Loop Concepts
Match Tokens
free (of|from)
not? (support[a-z]*|identified)
non(?!small|hodgkins)

Figure 1 Process diagram for a SAS-based coding, extraction, and nomenclature tool (SCENT).

individual concepts, some transformations are uniformly applied to the dictionary, including partial reverse lemmatization to account for plural words. Exclusionary words are also included for some concepts to prevent matches from occurring in unrelated contexts. The 'neck' anatomic site concept, for instance, includes 'bladder' as an exclusionary word to prevent matching when 'bladder neck' is encountered.

Preprocessing

To prepare electronic pathology report text for analysis, SCENT performs a number of preprocessing tasks. In the pathology database at KPSC, report text spanning multiple records is often split at locations other than sentence boundaries. To facilitate matching of clinical concepts, SCENT uses approximately 70 regular expressions to reconstruct reports for analysis. In addition to recreating sentence boundaries, special characters are removed and common abbreviations replaced. Sample pathology report text, in original and preprocessed forms, can be seen in figure 2.

Matching

Using a dictionary of clinical concepts, SCENT examines preprocessed pathology report text for concept matches. In preparation, a clinical concept dictionary is read into memory and stored as a SAS hash object for future use. The text of each report is split by sentence and stored in arrays of character variables. Sentences are then tokenized and their component words stored in arrays. Additional hash objects are used to record intermediate information, such as match results of individual concept words, match positioning, and negation status.

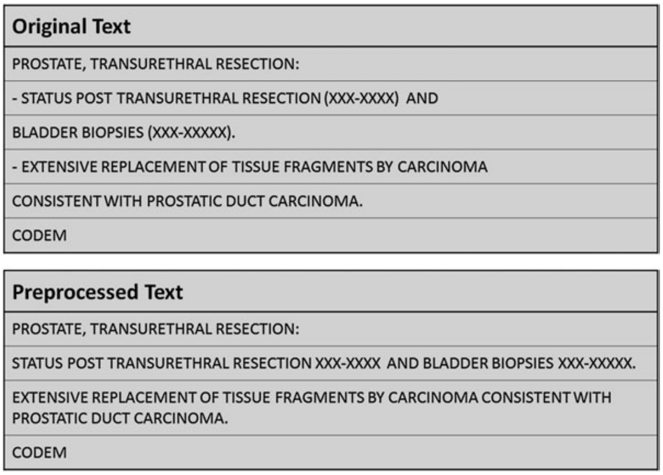
After preparing the necessary arrays and hash objects, SCENT sequentially examines the sentences of each pathology report. The clinical concept dictionary is processed separately for each new sentence and a search for the words of each concept is performed. Unmatched words are recorded to prevent unnecessary subsequent searches with concepts containing previously unmatched words. The text surrounding matches is examined for potential negation and the positions of non-negated matches are recorded. Dictionary concepts are considered fully matched if all words are found in proximate distance to each other. This distance is a token constant, defined as 10 words in the current study, and ignores prepositions, articles, and certain other words. SNOMED codes associated with matched concepts are assigned to the report from which they were matched. Surrounding text is examined to classify disease extent (eg, non-invasive, invasive, or metastatic). As shown in figure 3, words from reports that contribute to concept matches are tagged with relevant information, such as disease extent and SNOMED code. SCENT applies a single tag to multiple match words if they are adjacent or separated by only common prepositions and articles.

Using codes from matched clinical concepts and hierarchical decision rules, SCENT classifies the overall status of reports as either benign, borderline, basaloid, or malignant. The disease extent of assigned morphology codes and anatomic sites are used to differentiate between new primary and recurrent malignancies. To facilitate the process, anatomic site classifications are consolidated into categories. For example, the sternum and clavicle sites are considered part of the bone category. Sites relating to regional disease spread, such as the neck and groin, are consolidated into the lymph nodes category. In the case of identified cancer metastases, SCENT attempts to determine both origin and metastatic sites.

Negation and uncertainty

SCENT uses a collection of regular expressions to investigate potential negation of matched concepts. This approach is similar to that of the NegEx algorithm, which has been used successfully in multiple studies.[17–19] SCENT contains approximately 40 negation expressions, each assigned values limiting the acceptable distance between themselves and concept matches. For example, the expression 'free (of|from)' will negate concept matches appearing up to 10 words before or after the expression. In contrast, the negation word 'without' is valid up to only four words before concept matches and zero following them. The maximum number of words permitted between concept matches and negation expressions is modified by the presence of certain words, such as coordinating conjunctions (eg, 'and').

To differentiate clinical suspicions from conclusive findings, SCENT examines text surrounding matches for uncertainty cues, such as 'suspicious' and 'uncertain.' This process shares some conceptual similarity with the implementation of speculation cues by Clark *et al* in their MITRE system.[20] The scope of uncertainty by SCENT, however, is determined using fixed token distances rather than analysis of linguistic structure and statistical classifiers.



Original Text

PROSTATE, TRANSURETHRAL RESECTION:

- STATUS POST TRANSURETHRAL RESECTION (XXX-XXXX) AND

BLADDER BIOPSIES (XXX-XXXXX).

- EXTENSIVE REPLACEMENT OF TISSUE FRAGMENTS BY CARCINOMA

CONSISTENT WITH PROSTATIC DUCT CARCINOMA.

CODEM

Preprocessed Text

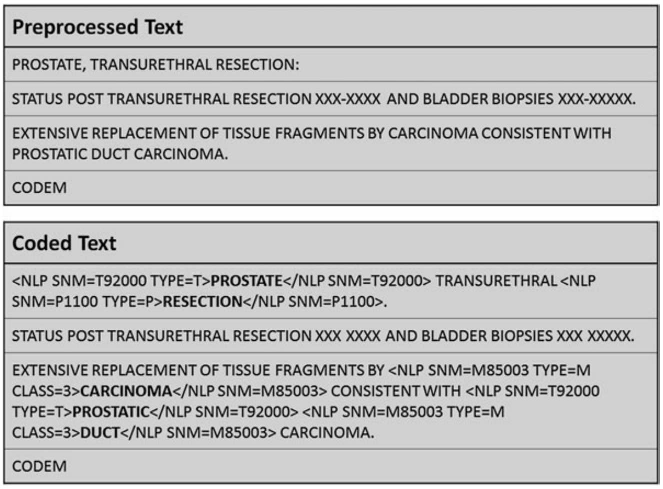
PROSTATE, TRANSURETHRAL RESECTION:

STATUS POST TRANSURETHRAL RESECTION XXX-XXXX AND BLADDER BIOPSIES XXX-XXXXX.

EXTENSIVE REPLACEMENT OF TISSUE FRAGMENTS BY CARCINOMA CONSISTENT WITH PROSTATIC DUCT CARCINOMA.

CODEM

Figure 2 Sample pathology report text following preprocessing by a SAS-based coding, extraction, and nomenclature tool (SCENT).



Preprocessed Text

PROSTATE, TRANSURETHRAL RESECTION:

STATUS POST TRANSURETHRAL RESECTION XXX-XXXX AND BLADDER BIOPSIES XXX-XXXXX.

EXTENSIVE REPLACEMENT OF TISSUE FRAGMENTS BY CARCINOMA CONSISTENT WITH PROSTATIC DUCT CARCINOMA.

CODEM

Coded Text

<NLP SNM=T92000 TYPE=T>**PROSTATE**</NLP SNM=T92000> TRANSURETHRAL <NLP SNM=P1100 TYPE=P>**RESECTION**</NLP SNM=P1100>.

STATUS POST TRANSURETHRAL RESECTION XXX XXXX AND BLADDER BIOPSIES XXX XXXXX.

EXTENSIVE REPLACEMENT OF TISSUE FRAGMENTS BY <NLP SNM=M85003 TYPE=M CLASS=3>**CARCINOMA**</NLP SNM=M85003> CONSISTENT WITH <NLP SNM=T92000 TYPE=T>**PROSTATIC**</NLP SNM=T92000> <NLP SNM=M85003 TYPE=M CLASS=3>**DUCT**</NLP SNM=M85003> CARCINOMA.

CODEM

Figure 3 Sample pathology report text following preprocessing and code assignment by a SAS-based coding, extraction, and nomenclature tool (SCENT).

Validation

A validation study was conducted under institutional review board approval using EMR records of breast and prostate cancer patients at KPSC. Electronic pathology reports were selected for validation due to their availability and significance in diagnosing most cancers. To assess the accuracy of SCENT, its classifications of pathology reports were compared to those of experienced chart abstractors. Classifications based on codes assigned by CoPathPlus were also compared.

Breast cancer cases were diagnosed from 2000 to 2007 with American Joint Committee on Cancer (AJCC) stage 0–III tumors, and had no prior history of cancer. All prostate cancer cases diagnosed from 2000 to 2005 were included irrespective of stage and previous cancer history. To address needs for improved recurrence identification in ongoing research at KPSC, validation focused on the period following cancer diagnosis and treatment. Of the 400 patients randomly selected for each cancer type, 206 breast and 186 prostate cancer patients had one or more pathology reports during the period beginning 6 months after diagnosis and subsequent to primary treatment(s) through the end of 2008. In total, 490 breast and 425 prostate cancer patient pathology reports were reviewed by two trained abstractors, one for each cancer group.

Abstractor reviews of pathology reports served as the gold standard. Each report was classified according to malignancy status. In the case of malignant and suspicious findings, abstractors also recorded anatomic site and any mention of metastasis. Non-melanoma skin malignancies were considered benign by abstractors, SCENT, and classifications based on CoPathPlus. Independent reviewers made final classifications for the small number of cases about which abstractors were uncertain. Reviewers also adjudicated the few cases of discordant classifications between SCENT and abstractors.

To reduce the time and cost of the validation process, we used SCENT to output pathology text, as can be seen in figure 4. Concepts relating to anatomic site were highlighted in green, malignancies in shades of blue according to disease extent, and suspicious findings in orange. Abstractors were instructed to fully review the text of each output report. To investigate the potential for bias stemming from SCENT highlights, a sample of pathology reports was independently reviewed by two abstractors. One set of reports contained highlighted text, the other did not. A total of 73 reports were reviewed from the first 30 breast cancer patients in the random sample. Abstractor classifications were found to be identical. Performance of SCENT and classifications based on CoPathPlus were assessed using standard evaluation metrics, including: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Cohen's κ.

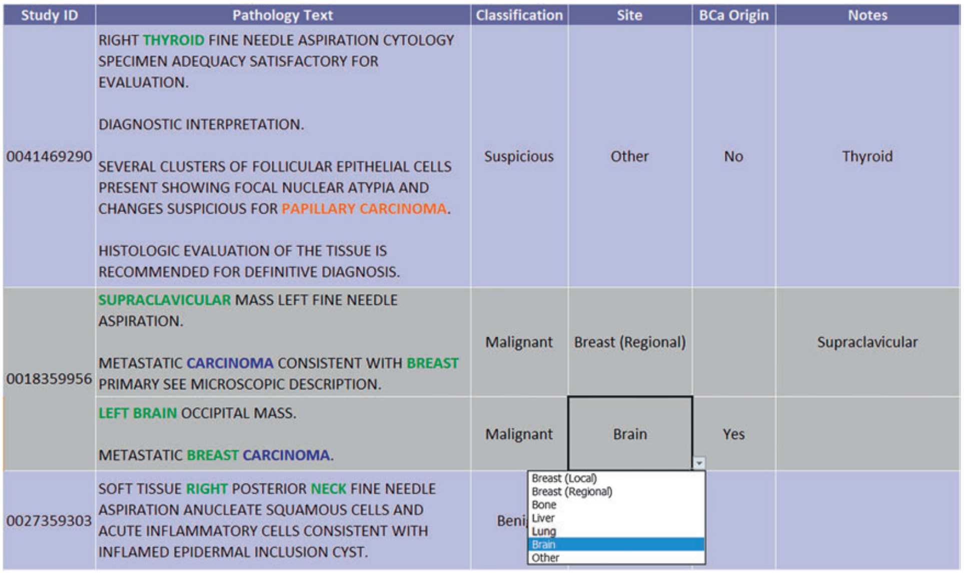
RESULTS

As shown in table 1, SCENT was highly concordant with abstractor classifications, achieving κ values of 0.96 and 0.95 in the breast and prostate cancer groups, respectively. SCENT identified 51 of 54 new primary and 60 of 61 recurrent cancer cases across both groups, with only three false positives in 792 true benign cases. Agreement was moderate for classifications based on the codes assigned by CoPathPlus, with κ values of 0.72 and 0.65 in the breast and prostate cancer groups, respectively. To calculate overall performance metrics, results were consolidated into two categories: benign/suspicious and primary/recurrent cancer. Within both the breast and prostate cancer groups, SCENT reached levels of or above 94% for sensitivity, specificity, PPV, and NPV (table 2). Specificity for classifications based on CoPathPlus exceeded 98% in both groups. PPV was moderately high in both the breast (90%) and prostate (88%) cancer groups. Sensitivity for classifications based on CoPathPlus in detecting true positive malignancies was moderate in the breast (74%) and prostate (71%) cancer groups.

A secondary assessment of SCENT was conducted to evaluate its ability to extract AJCC tumor staging and Gleason score information. SCENT identified and accurately extracted all tumor (T), lymph node (N), and metastasis (M) staging information from 19 pathology reports across both cancer groups. Within the prostate cancer group, SCENT identified all Gleason scores and accurately extracted the tumor pattern scores from each of the 20 reports. There were no instances in which SCENT failed to identify reports containing tumor staging or Gleason score information.

DISCUSSION

The favorable validation results in this study suggest that SAS-based NLP can be used to accurately identify and extract



Study ID	Pathology Text	Classification	Site	BCa Origin	Notes
0041469290	RIGHT THYROID FINE NEEDLE ASPIRATION CYTOLOGY SPECIMEN ADEQUACY SATISFACTORY FOR EVALUATION. DIAGNOSTIC INTERPRETATION. SEVERAL CLUSTERS OF FOLLICULAR EPITHELIAL CELLS PRESENT SHOWING FOCAL NUCLEAR ATYPIA AND CHANGES SUSPICIOUS FOR PAPILLARY CARCINOMA. HISTOLOGIC EVALUATION OF THE TISSUE IS RECOMMENDED FOR DEFINITIVE DIAGNOSIS.	Suspicious	Other	No	Thyroid
0018359956	SUPRACLAVICULAR MASS LEFT FINE NEEDLE ASPIRATION. METASTATIC CARCINOMA CONSISTENT WITH BREAST PRIMARY SEE MICROSCOPIC DESCRIPTION.	Malignant	Breast (Regional)		Supraclavicular
	LEFT BRAIN OCCIPITAL MASS. METASTATIC BREAST CARCINOMA.	Malignant	Brain	Yes	
0027359303	SOFT TISSUE RIGHT POSTERIOR NECK FINE NEEDLE ASPIRATION ANUCLEATE SQUAMOUS CELLS AND ACUTE INFLAMMATORY CELLS CONSISTENT WITH INFLAMED EPIDERMAL INCLUSION CYST.	Beni			

Figure 4 Sample chart review form used by abstractors to classify the pathology reports of breast and prostate cancer patients.

Table 1 Pathology report classifications of a SAS-based coding, extraction, and nomenclature tool (SCENT) and Cerner's CoPathPlus coding, as compared with abstractor review

	Benign		Cancer recurrence		Other primary cancer		Suspicious		
	%	N	%	N	%	N	%	N	κ Value
Breast cancer (abstractor)		(436)		(32)		(18)		(4)	
SCENT									
Benign	99.8	435	–	–	–	–	25.0	1	0.96
Cancer recurrence	–	–	100.0	32	–	–	–	–	
Other primary cancer	0.2	1	–	–	100.0	18	50.0	2	
Suspicious	–	–	–	–	–	–	25.0	1	
CoPathPlus									
Benign	97.2	424	12.5	4	22.2	4	50.0	2	0.72
Cancer recurrence	0.7	3	84.4	27	–	–	–	–	
Other primary cancer	–	–	–	–	55.6	10	25.0	1	
Suspicious	2.1	9	3.1	1	22.2	4	25.0	1	
Prostate cancer (abstractor)		(356)		(29)		(36)		(4)	
SCENT									
Benign	99.4	354	–	–	5.6	2	–	–	0.95
Cancer recurrence	–	–	96.6	28	2.8	1	–	–	
Other primary cancer	0.6	2	3.4	1	91.7	33	–	–	
Suspicious	–	–	–	–	–	–	100.0	4	
CoPathPlus									
Benign	96.1	342	10.3	3	33.3	12	50.0	2	0.65
Cancer recurrence	0.3	1	72.4	21	5.6	2	25.0	1	
Other primary cancer	1.1	4	13.8	4	52.8	19	–	–	
Suspicious	2.5	9	3.4	1	8.3	3	25.0	1	

Includes the pathology reports of randomly sampled breast and prostate cancer patients from 6 months after diagnosis and subsequent to primary treatment. Contralateral breast cancers were considered to be recurrences.

information from electronic clinical text. Using the pathology reports of patients previously diagnosed and treated for breast and prostate cancer, SCENT successfully identified 51 of 54 primary and 60 of 61 recurrent cancers. Additionally, there were only three false positive identifications within 793 true benign results. Performance was similar for both the breast and prostate cancer groups, with measures of sensitivity, specificity, PPV, and NPV of or above 94%.

Classifications using codes assigned by CoPathPlus were moderately successful in identifying incident and recurrent cancers within both cancer groups. Sensitivity in the breast and prostate cancer groups was 74% and 71%, respectively, while PPV was 90% and 88%. To confirm the accuracy of abstractor classifications, an oncologist reviewed pathology reports associated with discordant CoPathPlus cases. This review identified only one abstractor classification error, which was subsequently corrected. Codes assigned to discordant CoPathPlus cases were reviewed to better understand the sources of classification error. Primary sources of error for classifications based on CoPathPlus were: coding of historical malignancies, assignment of overly general codes, and lack of code assignment for unknown reasons. Additionally, there were three false positives relating to the coding of 'residual' malignancy, which had been excluded a priori from study definitions of primary and recurrent cancer.

The widespread adoption of SAS in clinical analysis and research settings ensures that SCENT is highly accessible. Integration of SAS with relevant data systems has already been established in these settings, allowing electronic text to be readily extracted for analysis. By avoiding the need for additional software installation or systems integration, technical and administrative barriers to NLP implementation are reduced. Additionally, SCENT does not require annotated training data and can be used by staff without specialized informatics or machine learning knowledge.

Primary and recurrent cancers are frequently among the main study outcomes in epidemiologic cancer research. Despite the importance of accurately identifying these events, electronic

Table 2 Validation metrics for a SAS-based coding, extraction, and nomenclature tool (SCENT) and Cerner's CoPathPlus coding, as compared with abstractor review of pathology reports

	Sensitivity*	Specificity*	PPV*	NPV*
Breast cancer				
SCENT	1.00 (0.93 to 1.00)	0.99 (0.98 to 1.00)	0.94 (0.85 to 0.98)	1.00 (0.99 to 1.00)
CoPathPlus	0.74 (0.60 to 0.84)	0.99 (0.98 to 1.00)	0.90 (0.77 to 0.96)	0.97 (0.95 to 0.98)
Prostate cancer				
SCENT	0.97 (0.89 to 0.99)	0.99 (0.98 to 1.00)	0.97 (0.89 to 0.99)	0.99 (0.98 to 1.00)
CoPathPlus	0.71 (0.59 to 0.80)	0.98 (0.96 to 0.99)	0.88 (0.77 to 0.95)	0.95 (0.92 to 0.97)

Includes the pathology reports of randomly sampled breast and prostate cancer patients from 6 months after diagnosis and subsequent to primary treatment.
*Shown with Wilson's 95% CI.
NPV, negative predictive value; PPV, positive predictive value.

diagnosis codes often lack accuracy and cancer registry data are not always available. SCENT achieved high levels of sensitivity and specificity in identifying pathologically diagnosed malignancies among patients previously diagnosed with breast or prostate cancer. In addition to its utility in retrospective research, SCENT can be used in prospective research and population care management. Cancer registry data, when available, are commonly delayed for months after patient diagnosis. It is often impractical to rely on these data for intervention studies relating to cancer treatment. SCENT is currently being implemented in one such study aimed at improving adherence to adjuvant hormonal therapy among breast cancer patients.

Beyond its value in identifying pathologically diagnosed primary and recurrent cancers, SCENT has the potential for numerous other clinical applications. Questions investigated by clinical analysts are often central to medical leadership and decisions relating to patient care. Analysts face issues similar to those of prospective research studies, relying on electronic diagnosis codes and available cancer registry data to assess the quality of cancer care. SCENT has the potential to enhance the availability and depth of data and could be used to identify appropriate patient populations for evaluation. For example, while treatment recommendations for certain precancerous lesions may dictate watchful waiting at present, SCENT could be used to identify affected patients if those recommendations were to change.

Resource requirements for chart review of uncoded electronic text can vary greatly between studies. Extraction of even small numbers of basic data elements can be infeasible in large sample studies due to labor costs and time requirements. As described in the Materials and methods section and illustrated in figure 4, SCENT can be used to highlight desired clinical concepts within electronic text. This functionality has the potential to dramatically reduce the time and costs associated with chart review. The benefits of SCENT in expediting chart review will be further explored and quantified in future work.

While SCENT performed well in classifying and extracting information from electronic pathology reports, its performance with other clinical text sources is currently untested. These sources will likely necessitate modifications to the preprocessing and decision rules employed by SCENT. Modifications to preprocessing rules may also be needed prior to implementing SCENT outside of KPSC. Additionally, text sources such clinical progress notes are less structured and have greater linguistic variability. Statistical NLP approaches are likely to provide superior performance with respect to these sources. SCENT does not use formal part-of-speech tagging and is therefore limited in its ability to disambiguate and contextualize identified clinical concepts. The performance of SCENT relative to that of a general purpose NLP solution will need to be assessed across multiple text sources to identify performance gaps and inform appropriate usage.

Future development and validation efforts for SCENT will include identification of incident and recurrent cancer cases that are diagnosed without pathological testing. Radiology reports and clinical progress notes are expected to be the primary sources for identifying non-pathologically diagnosed cancer cases. These sources pose additional challenges due to their diagnostic uncertainty and linguistic variability. SCENT may also have generalized utility extracting information outside of the oncology domain, such as standard scoring of cognitive impairment from the Mini–Mental State Examination (MMSE) and depression from the Patient Health Questionnaire (PHQ-9).

Functionality was developed to assign confidence scores to individual SCENT dictionary concepts according to their false-positive rates from chart review results. These scores can be used in conjunction with specified minimum confidence thresholds to fit the sensitivity requirements and false-positive tolerance of individual studies. This process will be updated to use Bayesian methods for incorporating additional evidence from subsequent chart reviews. The feasibility of adopting additional statistical methods to enhance the rules-based classification approach of SCENT will also be assessed.

CONCLUSION

SCENT was highly successful in identifying and extracting information on primary and recurrent cancers from electronic pathology reports. This functionality has the potential to provide significant value to clinical and epidemiologic researchers, particularly when statistical NLP is infeasible due to resource or other constraints. SCENT is proof of concept for SAS-based NLP applications that can be easily shared between institutions and used to support clinical and epidemiologic research.

Acknowledgments The authors would like to thank the following individuals for their contributions: Jianyu Rao and Jiaoti Huang from the University of California Los Angeles Ronald Reagan Medical Center for help in classifying morphology concepts; Melissa Preciado, Mayra Martinez, Nirupa Ghai, and Michelle McGuire from Kaiser Permanente Southern California for assistance in the chart abstraction process; Jeff Slezak from Kaiser Permanente Southern California for statistical and methodological advice; Lawrence Kushi from Kaiser Permanente Northern California, Debra Ritzwoller from Kaiser Permanente Colorado, and Joan Warren from the National Cancer Institute for sharing knowledge related to cancer recurrence identification.

Contributors JAS conceptualized and developed the natural language processing system described in this paper, led the validation study, and drafted the manuscript. CRC provided prostate cancer research expertise, assisted with data acquisition, contributed to validation study design, and interpreted results. MLK provided breast cancer research expertise, contributed to validation study design, and interpreted results. SAA and JES provided pathology and oncology expertise for construction of the clinical concept dictionary, contributed to validation study design, and adjudicated pathology report classifications. VPQ provided breast cancer research expertise, assisted with data acquisition, contributed to validation study design, interpreted results, and had significant input to the manuscript. All authors contributed to and approved the final manuscript.

Funding The validation study and development of both classification algorithms was supported by Kaiser Permanente Southern California Community Benefit and the National Cancer Institute funded Cancer Research Network (U19CA79689 to E Wagner). Tertiary funding for the algorithm related to CoPathPlus was provided by the National Institutes of Health (R01CA136743-01A2 to R Haque).

Competing interests None.

Ethics approval Ethics approval was provided by the Institutional Review Board of Kaiser Permanente Southern California.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement SCENT is freely available for non-commercial use and modification. Program source code and requisite support files may be downloaded from: http://www.kp-scalresearch.org/research/tools_scent.aspx

REFERENCES

1. **Warren JL,** Feuer E, Potosky AL, *et al*. Use of Medicare hospital and physician data to assess breast cancer incidence. *Med Care* 1999;**37**:445–56.
2. **Lamont EB,** Herndon JE 2nd, Weeks JC, *et al*. Measuring disease-free survival and cancer relapse using Medicare claims from CALGB breast cancer trial participants (companion to 9344). *J Natl Cancer Inst* 2006;**98**:1335–8.
3. **Office of the National Coordinator for Health Information Technology (ONC),** Department of Health and Human Services. Health information technology: Revisions to initial set of standards, implementation specifications, and certification criteria for electronic health record technology. Interim final rule with request for comments. *Fed Regist* 2010;**75**:62686–90.

4. **Hsiao CJ,** Hing E, Socey TC, *et al*. *Electronic Health Record Systems and Intent to Apply for Meaningful Use Incentives Among Office-based Physician Practices: United States, 2001—11*. Hyattsville, MD: National Center for Health Statistics, 2011.
5. **Carrell D,** Miglioretti D, Smith-Bindman R. Coding free text radiology reports using the Cancer Text Information Extraction System (caTIES). *AMIA Annu Symp Proc* 2007:889.
6. **Hazlehurst B,** Sittig DF, Stevens VJ, *et al*. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. *Am J Prev Med* 2005;**29**:434—9.
7. **Savova GK,** Ogren PV, Duffy PH, *et al*. Mayo clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc* 2008;**15**:25—8.
8. **Moore GW,** Berman JJ. Automatic SNOMED coding. *Proc Annu Symp Comput Appl Med care* 1994:225—9.
9. **Sager N,** Lyman M, Nhan NT, *et al*. Automatic encoding into SNOMED III: a preliminary investigation. *Proc Annu Symp Comput Appl Med Care* 1994:230—4.
10. **Chapman WW,** Nadkarni PM, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;**18**:540—3.
11. **Al-Haddad MA,** Friedlin J, Kesterson J, *et al*. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *HPB (Oxford)* 2010;**12**:688—95.
12. **Friedlin J,** Overhage M, Al-Haddad MA, *et al*. Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annu Symp Proc* 2010;**2010**:237—41.
13. **Childs LC,** Enelow R, Simonsen L, *et al*. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. *J Am Med Inform Assoc* 2009;**16**:571—5.
14. **Mishra NK,** Cummo DM, Arnzen JJ, *et al*. A rule-based approach for identifying obesity and its comorbidities in medical discharge summaries. *J Am Med Inform Assoc* 2009;**16**:576—9.
15. **Ware H,** Mullett CJ, Jagannathan V. Natural language processing framework to assess clinical conditions. *J Am Med Inform Assoc* 2009;**16**:585—9.
16. **Mohanty SK,** Piccoli AL, Devine LJ, *et al*. Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist. *BMC Cancer* 2007;**7**:144.
17. **Chapman WW,** Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;**34**:301—10.
18. **Uzuner O,** Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. *J Am Med Inform Assoc* 2009;**16**:109—15.
19. **Mitchell KJ,** Becich MJ, Berman JJ, *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.
20. **Clark C,** Aberdeen J, Coarr M, *et al*. MITRE system for clinical assertion status classification. *J Am Med Inform Assoc* 2011;**18**:563—7.