original report

Obtaining Knowledge in Pathology Reports Through a Natural Language Processing Approach With Classification, Named-Entity Recognition, and Relation-Extraction Heuristics

Tomasz Oliwa, PhD[1]; Steven B. Maron, MD, MSc[2]; Leah M. Chase[3]; Samantha Lomnicki[3]; Daniel V.T. Catenacci, MD[3]; Brian Furner, MS[1]; and Samuel L. Volchenboum, MD, PhD[1,3]

abstract

PURPOSE Robust institutional tumor banks depend on continuous sample curation or else subsequent biopsy or resection specimens are overlooked after initial enrollment. Curation automation is hindered by semistructured free-text clinical pathology notes, which complicate data abstraction. Our motivation is to develop a natural language processing method that dynamically identifies existing pathology specimen elements necessary for locating specimens for future use in a manner that can be re-implemented by other institutions.

PATIENTS AND METHODS Pathology reports from patients with gastroesophageal cancer enrolled in The University of Chicago GI oncology tumor bank were used to train and validate a novel composite natural language processing-based pipeline with a supervised machine learning classification step to separate notes into internal (primary review) and external (consultation) reports; a named-entity recognition step to obtain label (accession number), location, date, and sublabels (block identifiers); and a results proofreading step.

RESULTS We analyzed 188 pathology reports, including 82 internal reports and 106 external consult reports, and successfully extracted named entities grouped as sample information (label, date, location). Our approach identified up to 24 additional unique samples in external consult notes that could have been overlooked. Our classification model obtained 100% accuracy on the basis of 10-fold cross-validation. Precision, recall, and F1 for class-specific named-entity recognition models show strong performance.

CONCLUSION Through a combination of natural language processing and machine learning, we devised a re-implementable and automated approach that can accurately extract specimen attributes from semistructured pathology notes to dynamically populate a tumor registry.

JCO Clin Cancer Inform. © 2019 by American Society of Clinical Oncology

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on June 27, 2019 and published at ascopubs.org/journal/cci on July 31, 2019: DOI https://doi.org/10.1200/CCI.19.00008

INTRODUCTION

Tumor tissue repositories contain annotated tissue specimen information that is vital to cancer research. Historically, these represent small grant-funded efforts to curate data manually. However, the big data movement has led to multi-institutional and even international collaborative efforts, which render continuous manual abstraction infeasible if not obsolete.

Despite near-universal electronic medical record use, pathology reports remain as free text containing semistructured elements detailing a specimen's source and gross and microscopic characteristics. Although other groups have developed tools to aid in parsing the cancer type or tumor characteristics from these reports, such as to identify relevant patients for registry inclusion[1-4] or to determine TNM staging,[5-9] none, to our knowledge, have attempted to extract and group semistructured specimen identifiers themselves.

Regardless of the specimen type, these reports customarily include an institutional accession number (label), tissue container identifier for each specimen collected (eg, A) Esophagus, B) Colon), and tissue block number prefaced by its respective container identifier (eg, A1, B1). In large academic centers, many patient samples are collected elsewhere and sent for external consultation. These external reports also contain a collecting institution name (eg, St Anthony's Hospital), and unlike internal primary institution reports, they may contain multiple samples, each with unique sample identifiers from different sites and dates.

Although reports follow a prescribed format, the unstructured text requires tedious data abstraction of discrete elements. All of this information is of great utility, but the heterogeneous nature of the notes makes extraction through rigid dictionary or rule-based approaches infeasible. This motivated us to develop and describe a re-implementable computational pipeline with open-source packages (see Data Supplement

CONTEXT SUMMARY

Key Objective

Repositories with annotated tissue specimen information are vital to cancer research; however, historically, they often contain manually curated data, including specimen accession numbers, accessioning facility, and procedure date. We developed a natural language processing approach with named-entity recognition to extract and group these semistructured specimen identifiers from the report text and evaluated it using pathology notes.

Knowledge Generated

Through supervised classification, reports were reliably separated into internal and external consult notes. A named-entity recognition-based approach applied on external notes was able to identify additional samples that would have likely been overlooked compared with assuming one sample per note.

Relevance

With natural processing language, sample information can be extracted from pathology notes, which could allow automated sample detection for tumor repositories. An integration with approaches that extract tumor characteristics using oncology ontology mapping also would improve sample identification for translational research.

for descriptions of most methodological parts) that natural language processing (NLP) practitioners can re-implement with their own annotated data on the basis of NLP, named-entity recognition (NER), and machine learning approaches. Our motivation is to develop a scalable method that dynamically identifies existing pathology specimens for patients already enrolled in an institutional registry to capture all available tumor specimens from consented patients.

PATIENTS AND METHODS

This computational pipeline was created in support of the data abstraction process for a large retrospective chart review project. Approval for the study was obtained from The University of Chicago's institutional review board. It was considered human-subjects research with minimal risk to participants, and a waiver to obtain consent was obtained. Pathology reports from 2006 to 2016 on patients with gastroesophageal cancer enrolled in The University of Chicago GI oncology tumor bank were used to develop the pipeline. Clinical information was acquired from the clinical research data warehouse.

Our overall approach to obtain knowledge from pathology reports is shown in Figure 1A. If a pathology note has not been classified as internal or external (consult), a name de-identification step is performed, and a machine learning–based binary classification model predicts this class. This classification is of importance for the result quality of the NER models and may be of additional utility for information retrieval purposes as a discrete element in a notes data warehouse but is not of clinical importance. Next, a class-specific NER model is used to extract the named entities of label, received location, received date, and sublabels. A subsequent annotation heuristic then combines the named entities into sample information, calculates the number of accession numbers within a single pathology report (sample count) and the number of tissue specimens in a given report (received number), and performs data normalization/proofreading. By virtue of being a machine learning model and with our normalization procedures, this NER method addresses the challenges with human error and variable nomenclature in the notes.

Preprocessing

The clinical notes were acquired as plain-text files. For the classification step, we removed all names that were identified by PhysioNet de-identification software[10,11] (Data Supplement), which was run with GNU Parallel.[12] In addition, to aid classification, the procedure category (cytology, surgical pathology, dermatology, etc) was added as a text string to the content of the note. For the NER step, the unchanged notes were used.

Classification

The open-source Python package scikit-learn[13] is used to create a machine learning pipeline (Fig 1B). The pathology notes are read from plain-text files and transformed into a high-dimensional bag-of-words representation weighted by the term frequency-inverse document frequency (TF-IDF), a method commonly used in the information retrieval community.[14] For each term (token or phrase; Data Supplement) of a note, TF-IDF calculates the term's importance as a product of two measures: the term frequency and the inverse document frequency. We used TF-IDF as implemented by scikit-learn. Scikit-learn's version of a support vector machine (SVM; sklearn.svm.SVC, which is based on C-SVM[15]), a supervised machine learning model for classification, is trained to predict the class (internal or external consult) of a note on the basis of its numerical TF-IDF vectors. Hyperparameter optimization in scikit-learn was used to account for results from all configurations (Data Supplement). Hyperparameter optimization[16] means that a machine learning method is applied on the same problem with different configuration settings (the hyperparameters) to find the best setting.

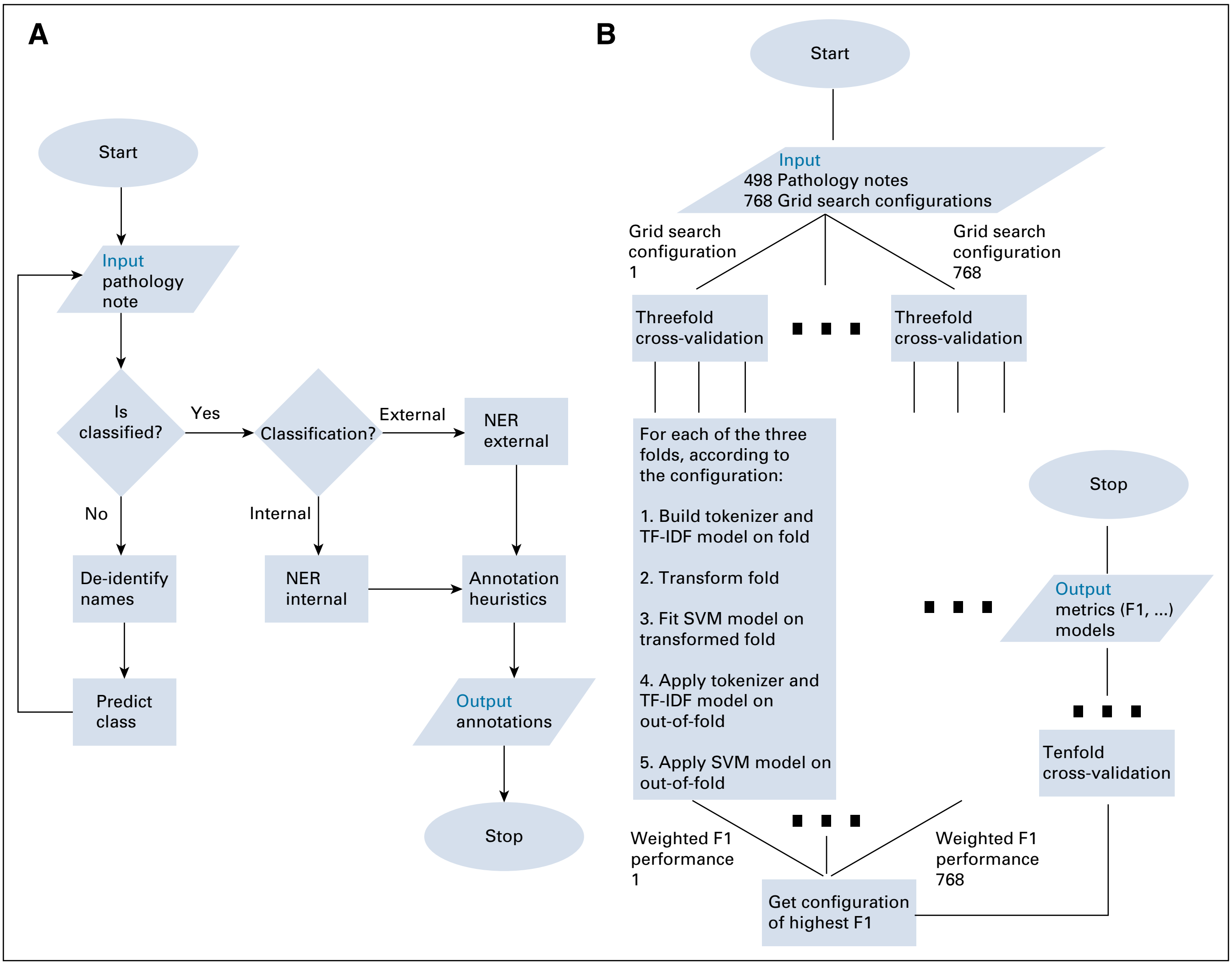


FIG 1. Approach flowcharts. (A) Overview of the composite classification, named-entity recognition (NER), and heuristics algorithm. (B) A grid search over the hyperparameter configurations is performed. The best found configuration is evaluated further with a stratified 10-fold cross-validation test. SVM, support vector machine; TF-IDF, term frequency-inverse document frequency.

NER

We trained Stanford NER[17] models to detect the following named entities (see Fig 2 for annotated examples):

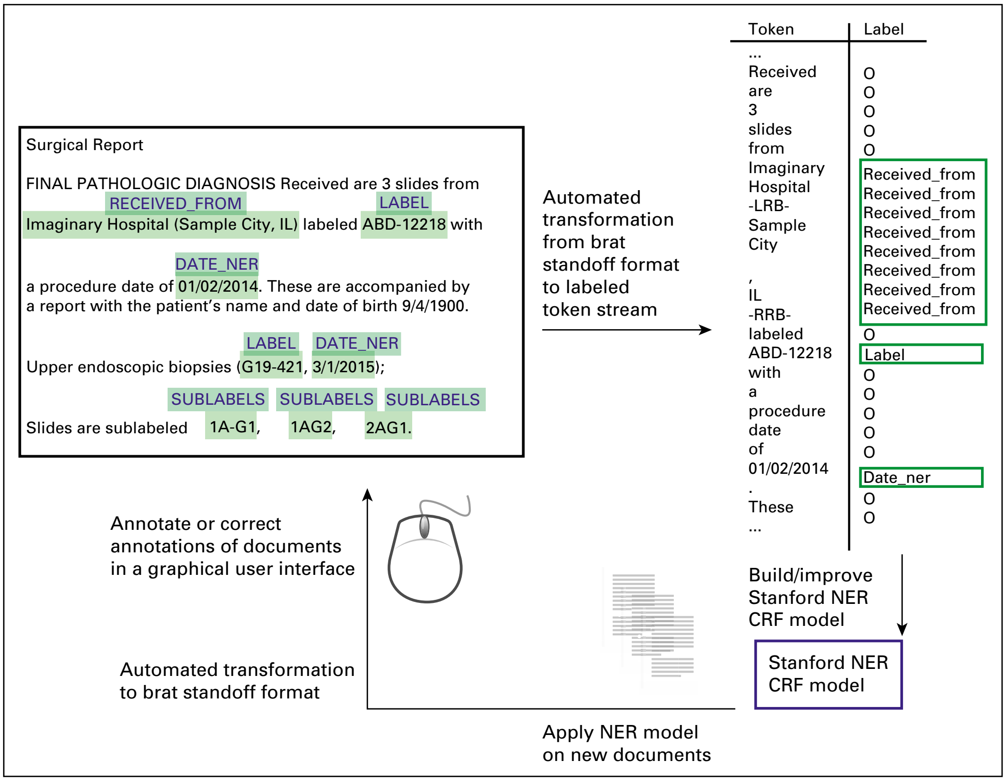
- Labels: specimen accession numbers (report may contain multiple accession numbers for externally collected consult cases)
- Received locations: for external consult cases, this will be the original accessioning facility for each included case
- Dates: procedure date for specimens in the report
- Sublabels: tissue block identifiers

NER Graphical User Interface Loop

Training set reports were annotated with the browser-based brat rapid annotation tool[18] (brat standoff format with only entity annotations). We developed software that integrates the output of the brat rapid annotation tool with Stanford NER, which enables the user to work in an iterative loop similar to the tag-a-little, learn-a-little[19] strategy (Fig 2).

Sample Information

Sample information is formed that consists of label, date, and location on the basis of a heuristic, which is given as pseudocode in the Data Supplement. As a high-level heuristic summary, for each found label, it looks up dates and locations within proximity by following rules that were crafted on the basis of visual inspection of notes to address most of the patterns we found. For example, it considers that there is not another label in between the currently investigated label and a date. The sample information for the synthetic data in Fig 2 would be as follows: Label: ABD-12218, Date: 01/02/2014, Location: Imaginary Hospital (Sample City, IL), and Label: G19-421, Date: 3/1/2015, Location: Imaginary Hospital (Sample City, IL). To obtain the sample count (the count of unique accession



Surgical Report

FINAL PATHOLOGIC DIAGNOSIS Received are 3 slides from Imaginary Hospital (Sample City, IL) labeled ABD-12218 with a procedure date of 01/02/2014. These are accompanied by a report with the patient's name and date of birth 9/4/1900.

Upper endoscopic biopsies (G19-421, 3/1/2015);

Slides are sublabeled 1A-G1, 1AG2, 2AG1.

Token	Label
...	
Received	O
are	O
3	O
slides	O
from	O
Imaginary	Received_from
Hospital	Received_from
-LRB-	Received_from
Sample	Received_from
City	Received_from
,	Received_from
IL	Received_from
-RRB-	Received_from
labeled	O
ABD-12218	Label
with	O
a	O
procedure	O
date	O
of	O
01/02/2014	Date_ner
.	O
These	O
...	

RECEIVED_FROM
LABEL
DATE_NER
LABEL
DATE_NER
SUBLABELS
SUBLABELS
SUBLABELS
Automated transformation from brat standoff format to labeled token stream
Build/improve Stanford NER CRF model
Stanford NER CRF model
Apply NER model on new documents
Automated transformation to brat standoff format
Annotate or correct annotations of documents in a graphical user interface

FIG 2. The iterative loop aids in creating new named-entity recognition (NER) models. The note text contains synthetic data for illustrative purposes. CRF, conditional random field.

numbers [labels]), normalization heuristics remove prefixes [such that labels #ABC-1234 and (1)ABC-1234 would equal ABC-1234] and identify certain variations of a label as the same unique label (eg, MD - 033, MD-0033, and MD-33). The Data Supplement provides the methodological details of these proofreading named-entity normalization heuristics.

Sublabels

Sublabels often appear in a separate section of the note below the occurrences of the label, location, and date named entities (Fig 2). The NER model identifies these; the Data Supplement contains the methodological details of the capturing and normalization of sublabels and the related number of containers (received number) computation.

Classification Analysis

A randomly chosen subset of 498 name de-identified pathology notes from the overall notes set was labeled by a GI oncologist into 403 internal and 95 external notes. The classification hyperparameter optimization maximized the weighted F1 metric on threefold cross-validations, resulting in 768 possible hyperparameter candidate sets and 2,304 model fits. F1 is a widely used quality measure in machine learning[4,19] and can be understood as the harmonic mean of the precision and recall, with 0 being the lowest and worst value and 1.0 being the best value (see Aberdeen et al[19] for definitions). A cross-validation[6,7] performs repeated out-of-sample testing during model building to address overfitting or sample bias. A model that is based on the best configuration is further evaluated with 10-fold cross-validation (Fig 1B). Through the weights of the linear SVM kernel, the top predictive token/ngram features per class are given.

NER Analysis

From the overall notes set, 106 randomly chosen external notes were annotated with label, received location, date, and sublabel annotations, and 82 randomly chosen internal notes were annotated with sublabel annotations by a GI oncologist. To provide the statistical measures of precision, recall, and F1 on an out-of-training set as it is done in NER,[17] using a 1/3 and 2/3 split, we randomly separated the external notes into 70 training and 36 testing notes and the internal notes into 54 training notes and 28 testing notes. To provide a measure for the clinical impact, the sample count was obtained. To examine the difference of building class-specific NER models and validate further the necessity for classification, class-unspecific NER models were built with data from both internal and external notes and evaluated on an out-of-training set, and their F1 metrics were compared. Length metrics for the sublabel named entity from both classes are provided.

RESULTS

Reports from 173 patients with pathology reports from a single institution (either internal or external consultation reports) were analyzed in this study, of which 172 patients had gastroesophageal cancer, and one had a hepatoid small bowel carcinoma. There was a slight predominance of men, as seen in gastroesophageal cancer, and most patients had more advanced stage III or IV disease (Table 1). Table 1 is organized by number of notes in each category and is provided so that other institutions that re-implement our NLP approach can examine the types of notes investigated here when using their own data.

Classification Results

Sixty-eight of 768 hyperparameter configurations obtained a weighted F1 of 1.0 through cross-validation. The average F1 was 0.986, with a sample standard deviation of 0.015. The lowest-performing setting resulted in an F1 of 0.948, which demonstrates that it was sensible to cast this problem as an NLP classification task. This best model parameter setting (Data Supplement) was then used as the basis for a stratified 10-fold cross-validation test[16] on the data set, which resulted in 100% accuracy. Precision, recall, and F1 measures, defined, for instance, in Aberdeen et al[19] or Yala et al,[4] all resulted in the highest value of 1.0.

The highest-weighted features from the linear SVM kernel included “il” in the external class (presumably because local laboratories were mostly located in Illinois) and hospital (because this token would appear in their address or name; Fig 3). The token deid_name_il seems to indicate false positives from the name de-identification. Of note, specific sample descriptions such as pink, tissue, and soft were top tokens for internal notes.

NER Results

The notes were transformed as Stanford NER[17] input with Stanford NLP.[20] As listed in Table 2, the overall F1 measure in row 5 is 0.9014, which demonstrates the predictive power of this approach. This result reflects the findings in Aberdeen et al[19] where an NER model showed a satisfactory quality for many named-entity classes with only several hundred training label instances. The F1 of 0.7778 in row 4 suggests that the most difficult named entities for the model are the sublabels. Our NER method identified 130 samples from 106 reports: 24 more unique accession numbers than notes suggests that our approach may increase sample identification yield from consult reports by up to 23% compared with assuming one sample per report. The model quality for internal notes sublabels is superior to that from the external sublabels.

Combination of External and Internal Training

We built a class-unspecific NER model by combining all 82 internal notes with the set of external training notes and tested it against the same external testing set. The performance for sublabels dropped (0.6752 *v* 0.7778), as listed in row 7 of Table 2. In a similar experiment, we kept only the sublabel annotations in the 106 external notes and added them to the 54 internal training notes. The results are listed in row 8 of Table 2. Similarly as before, the F1 performance dropped (0.8808 *v* 0.9154).

To inspect this performance difference, we calculated named-entity length metrics. In the data, external sublabels had a mean character length of 4.83 with a standard deviation of 4.1. Internal sublabels had a mean character length of 3.1 with a standard deviation of 1.68. These results indicate a difference in sublabel named-entity manifestation in internal versus external notes and further justify the supervised classification of the notes to build class-specific NER models.

DISCUSSION

Electronic records of pathology reports remain inherently unstructured. In consequence, research databases, such as tumor repositories, continue to depend on manual

TABLE 1. Demographics of Included Patients by Number of Notes in Each Category

	Classification		NER	
Demographic	**Internal**	**External**	**Internal**	**External**
No. of notes	403	95	82	106
Median age (range)	64 (16-88)	64 (19-88)	64.5 (20-88)	65 (24-88)
Male sex	218	68	35	80
Stage				
I	58	12	15	8
II	67	3	24	9
III	119	22	23	19
IV	148	57	20	66
Unknown	11	1	0	4
Site				
Esophagogastric	137*	48*	22	63
Gastric	265	46	60	43
Report type				
Cytology	138	14	21	26
Surgical pathology general	154	63	55	65
Surgical pathology report	49	11	0	7
Surgical pathology request	49	7	0	8
Dermatopathology	9	0	4	0
Hematopathology	3	0	2	0
Autopsy	1	0	0	0

Abbreviation: NER, named-entity recognition.

*One patient was inadvertently included in this cohort despite not having gastroesophageal cancer.

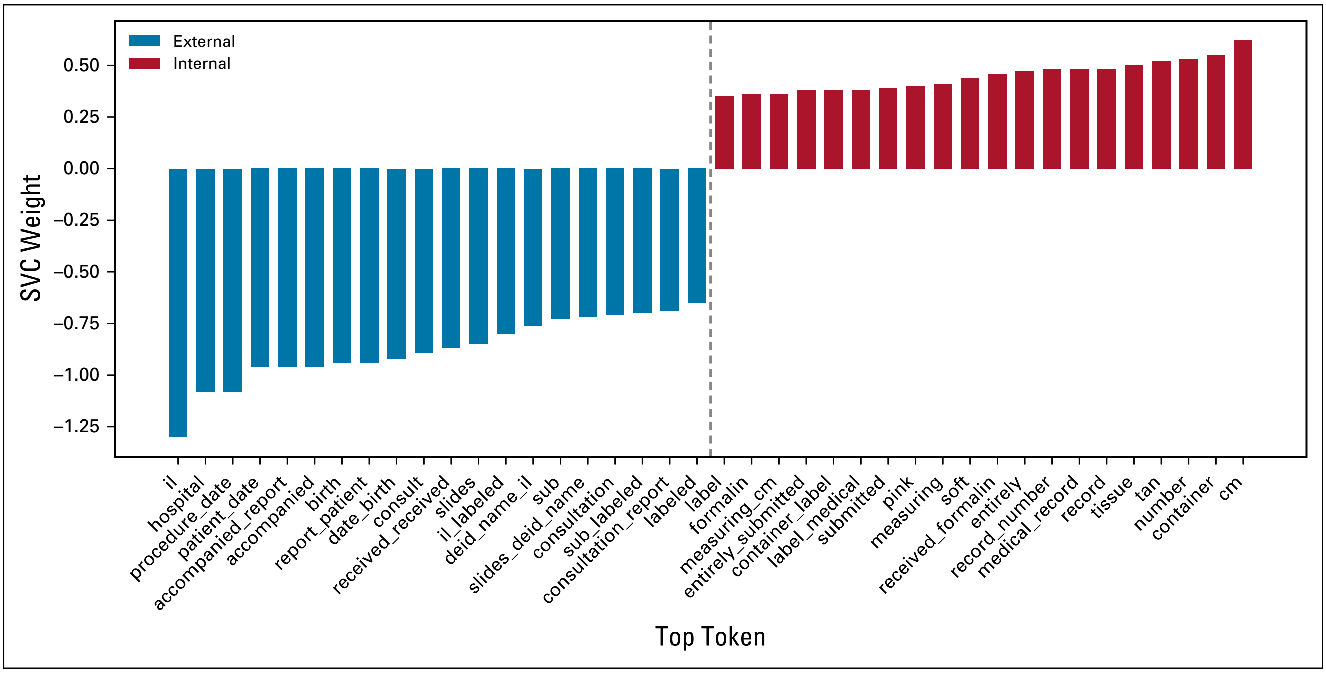


FIG 3. Top predictive model features for the external (left) and internal (right) class. SVC, support vector classifier.

abstraction. As databases grow beyond institutions, this approach becomes infeasible. We developed and applied a machine learning–based NER model to 188 free-text pathology reports and demonstrated that we are able to extract sample information accurately.

Although seemingly a simple task because of documentation heterogeneity, a stepwise approach was required to accomplish this goal. To ensure extraction of multiple specimens from external sample reports, reports first had to be partly de-identified and classified as internal or external (consult). Our model effectively accomplished this classification goal with 100% accuracy in a 10-fold stratified cross-validation. However, ideally the classification model should be externally validated against unrelated notes from another institution. Reports were sourced from a single institution, although they contained a mix of internal reports as well as consultation reports on externally collected samples, which represented only a minority of reports, so overtraining could partly explain our high accuracy, despite our application of cross-validation. The classification performance most likely would drop if the model were to classify notes from another institution without retraining on a part of these. Preferably, the model also should be evaluated on different clinical reports from those used for training and hyperparameter optimization. Scikit-learn was chosen for classification because we did not find an

TABLE 2. NER Model Performances

Setup	Entity	Precision	Recall	F1	TP	FP	FN
External notes	Date	0.9765	0.9651	0.9708	83	2	3
External notes	Label	1.0000	0.8529	0.9206	87	0	15
External notes	Received location	0.9750	0.8864	0.9286	78	2	10
External notes	Sublabel	0.9130	0.6774	0.7778	63	6	30
Totals (rows 1-4) external notes	Totals (rows 1-4)	0.9688	0.8428	0.9014	311	10	58
Internal notes	Sublabel	0.9787	0.8598	0.9154	92	2	15
Trained on external notes training set and all internal notes tested on external testing set	Sublabel	0.8281	0.5699	0.6752	53	11	40
Trained on internal notes training set and all external notes tested on internal testing set	Sublabel	0.9884	0.7944	0.8808	85	1	22

NOTE. The Setup column indicates the training and testing setup for each row. Unless noted otherwise in the Setup column, the training and testing sets are taken from the same given note class (ie, in the first row, only external notes are used, and the model is trained on the external training set and tested on the external testing set).

Abbreviations: FN, false negative; FP, false positive; NER, named-entity recognition; TP, true positive.

accessible combination of stratified cross-validation, hyperparameter optimization, and TF-IDF provided for the classifier component of Stanford CoreNLP. In consideration of the high-dimensional data representation, we chose an SVM with a linear kernel without particular reasons against other linear models. On the basis of the excellent performance, there was no need for more-complex and nonlinear models.

Next, an NER model extracted sample identifiers from the reports. In an ideal world, a rigid rules-based approach could be used, but in reality, the unstructured text phrasing is heterogeneous because of inter-reporter variation and typographic errors. In similar situations, machine learning–based NER has been successful in a variety of domains, such as in detecting people or organizations for the CoNLL NER Reuters news task,[17] detecting protected health information,[19] and identifying significant concepts from radiology reports.[21] We chose Stanford NER because scikit-learn did not offer sequence modeling, which is how NER through machine learning commonly is performed. Our results demonstrate that by applying note classification first and then building class-specific NER models, we can achieve high-quality, albeit not perfect, NER results. Furthermore, we show that the class-unspecific NER models provide a comparatively worse performance. Despite our NER approach, because of repetitions of the same label (accession number) in a note or variable formatting (eg, MD - 033, MD-0033), we required additional heuristics that only counted unique labels and removed such variable formatting (see Data Supplement for methodology) to obtain the sample count. Without these steps, the count of identified label named entities (including the repetitions) resulted in 195 instead of 130, which shows the importance of heuristics to obtain unique labels.

Our approach could be generalized for other clinical applications. Specimen measurements could be captured as an additional NER class and stored in a clinical database. For progress notes, medications, dose, and route of administration with variations of enteral, parenteral, and topical named entities could be detected and grouped. As a limitation, the relation-extraction heuristics would have to be modified because they are based on the encoding of patterns from text examples for specific tasks. In addition, the classifier and NER components would have to be retrained.

Looking forward, this approach not only identifies additional samples that likely were overlooked, but also could be used to flag reports that contain multiple accession numbers for additional review by data abstractors and could be integrated with models that extract tumor characteristics using oncology ontology mapping.[1-4] As a limitation, we considered labels as unique if the normalization heuristics did not identify them as the same labels, and we report them here as such. Other variable formatting of the same labels not foreseen by us ideally would have to be added to the normalization heuristics. For the medication detection example, synonyms likely also would have to be compared. We have demonstrated that by using a stepwise classification and NER approach we are able to extract sample information accurately from both primary and consult pathology reports, which will allow us to automate prospective sample identification for tumor repositories. With the help of this approach, tumor banks can dynamically populate samples from their electronic record systems into their databases, which would vastly improve sample identification and multi-institutional collaboration for translational research.

AFFILIATIONS

[1]The University of Chicago, Chicago, IL
[2]Memorial Sloan Kettering Cancer Center, New York, NY
[3]The University of Chicago Medical Center, Chicago, IL

CORRESPONDING AUTHOR

Tomasz Oliwa, PhD, Center for Research Informatics, The University of Chicago, The Shoreland, 5454 S Shore Dr, Suite 1D, Chicago, IL 60615; e-mail: toliwa@bsd.uchicago.edu.

EQUAL CONTRIBUTION

T.O. and S.B.M. contributed equally to this work.

SUPPORT

Supported by a National Institutes of Health K23 award (CA178203-01A1) and K12 award (5K12CA139160-08), a University of Chicago Comprehensive Cancer Center Award in Precision Oncology—Cancer Center Support grant (P30CA014599), the Anthony F. and Helen J. Castle Foundation, a Live Like Katie Foundation award, the Salvatore Ferrara II Esophageal Cancer Research Fund for Personalized Antibodies for Gastroesophageal Adenocarcinoma, the American Association for Cancer Research, and the Conquer Cancer Foundation of ASCO. The Center for Research Informatics is funded by the Biological Sciences Division, Institute for Translational Medicine/Clinical and Translational Science Awards (National Institutes of Health Grant No. UL1 TR000430), at The University of Chicago.

AUTHOR CONTRIBUTIONS

Conception and design: Tomasz Oliwa, Steven B. Maron, Daniel V.T. Catenacci, Brian Furner, Samuel L. Volchenboum
Financial support: Steven B. Maron, Daniel V.T. Catenacci
Administrative support: Samuel L. Volchenboum
Provision of study material or patients: Steven B. Maron, Daniel V.T. Catenacci
Collection and assembly of data: Tomasz Oliwa, Steven B. Maron, Leah M. Chase, Samantha Lomnicki, Daniel V.T. Catenacci, Brian Furner
Data analysis and interpretation: Tomasz Oliwa, Steven B. Maron, Daniel V.T. Catenacci
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated.

Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Steven B. Maron
Stock and Other Ownership Interests: Calithera Biosciences
Honoraria: RMEI
Consulting or Advisory Role: Health Advances
Research Funding: Genentech (Inst)
Travel, Accommodations, Expenses: Merck

Leah M. Chase
Patents, Royalties, Other Intellectual Property: Co-inventor of CleanCision, a wound retraction and protection system in protecting surgical incisions from intraoperative contamination when used during colorectal surgery

Daniel V.T. Catenacci
Honoraria: Genentech, Roche, Eli Lilly, Amgen, OncoPlex Diagnostics, Foundation Medicine, Taiho Pharmaceutical, Genmab, NantOmics, Guardant Health, Merck, Bristol-Myers Squibb, Gritstone Oncology, Five Prime Therapeutics, Astellas Pharma
Consulting or Advisory Role: Genentech, Roche, Amgen, Merck, Eli Lilly, Taiho Pharmaceutical, Bristol-Myers Squibb, Astellas Pharma
Speakers' Bureau: Guardant Health, Foundation Medicine, Genentech, Eli Lilly, Merck

Samuel L. Volchenboum
Stock and Other Ownership Interests: Litmus Health
Honoraria: Sanford Health
Consulting or Advisory Role: CVS Accordant
Research Funding: AbbVie (Inst)

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT
We thank Julie A. Johnson and Stacie Landron from the Center for Research Informatics (CRI) at The University of Chicago. The computational infrastructure was provided by the applications development team maintained by CRI. Data from this study were provided by the Clinical Research Data Warehouse maintained by CRI.

REFERENCES

1. 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 17:253-264, 2010
2. Hanauer DA, Miela G, Chinnaiyan AM, et al: The registry case finding engine: An automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg 205:690-697, 2007
3. Nguyen AN, Moore J, O'Dwyer J, et al: Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. AMIA Ann Symp Proc 2015:953-962, 2015
4. Yala A, Barzilay R, Salama L, et al: Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 161:203-211, 2017
5. McCowan I, Moore D, Fry MJ: Classification of cancer stage from free-text histology reports. Conf Proc IEEE Eng Med Biol Soc 1:5153-5156, 2006
6. Martinez D, Li Y: Information extraction from pathology reports in a hospital setting. Proc Assoc Comput Machinery Int Conf Inf Knowledge Manage 1877-1882, 2011
7. Napolitano G, Marshall A, Hamilton P, et al: Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. Artif Intell Med 70:77-83, 2016
8. Glaser AP, Jordan BJ, Cohen J, et al: Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform 10.1200/CCI.17.00128
9. Leyh-Bannurah S-R, Tian Z, Karakiewicz PI, et al: Deep learning for natural language processing in urology: State-of-the-art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. JCO Clin Cancer Inform 10.1200/CCI.18.00080
10. Goldberger AL, Amaral LAN, Glass L, et al: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 101:E215-E220, 2000
11. Neamatullah I, Douglass MM, Lehman LW, et al: Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 8:32, 2008
12. Tange O: GNU parallel: The command-line power tool. ;login: The USENIX Magazine:36:42-47, 2011
13. Pedregosa F, Varoquaux G, Gramfort A, et al: Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825-2830, 2011
14. Beel J, Gipp B, Langer S, et al: Research-paper recommender systems: A literature survey. Int J Digit Libr 17:305-338, 2016
15. Chang C-C, Lin C-J: LIBSVM: A library for support vector machines. ACM Transactions Intell Syst Technol 2:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm
16. Buitinck L, Louppe G, Blondel M, et al: API design for machine learning software: Experiences from the scikit-learn project. Eur Conf Machine Learn Principles Pract Knowledge Discovery Databases 108-122, 2013
17. Finkel JR, Grenager T, Manning C: Incorporating non-local information into information extraction systems by Gibbs sampling. Proc 43rd Annu Meeting Assoc Comput Linguistics 363-370, 2005
18. Stenetorp P, Pyysalo S, Topić G, et al: BRAT: A Web-based tool for NLP-assisted text annotation. Proc Demonstrations 13th Conf Eur Chapter Assoc Comput Linguistics 102-107, 2012
19. Aberdeen J, Bayer S, Yeniterzi R, et al: The MITRE Identification Scrubber Toolkit: Design, training, and assessment. Int J Med Inform 79:849-859, 2010
20. Manning CD, Surdeanu M, Bauer J, et al: The Stanford CoreNLP natural language processing toolkit. Proc 52nd Annu Meeting Assoc Comput Linguistics Syst Demonstrations 55-60, 2014
21. Hassanpour S, Langlotz CP: Information extraction from multi-institutional radiology reports. Artif Intell Med 66:29-39, 2016