Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Evaluation of Automated Public De-Identification Tools on a Corpus of Radiology Reports

Jackson M. Steinkamp, MD • Taylor Pomeranz, MD • Jason Adleberg, MD • Charles E. Kahn, Jr, MD, MS • Tessa S. Cook, MD, PhD

From the Department of Radiology, Hospital of the University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104 (J.M.S., T.P., J.A., C.E.K., T.S.C.); and Boston University School of Medicine, Boston, Mass (J.M.S.). Received August 5, 2019; revision requested October 23; revision received May 5, 2020; accepted May 14. **Address correspondence to** J.M.S. (e-mail: *jacksonsteinkamp@gmail.com*).

T.S.C. supported by Radiological Society of North America education grants, an American College of Radiology innovation award, and a Society for Imaging Informatics in Medicine grant.

Conflicts of interest are listed at the end of this article.

See also the commentary by Tenenholtz and Wood in this issue.

Radiology: Artificial Intelligence 2020; 2(6):e190137 • https://doi.org/10.1148/ryai.2020190137 • Content code: AI

Purpose: To evaluate publicly available de-identification tools on a large corpus of narrative-text radiology reports.

Materials and Methods: In this retrospective study, 21 categories of protected health information (PHI) in 2503 radiology reports were annotated from a large multihospital academic health system, collected between January 1, 2012 and January 8, 2019. A subset consisting of 1023 reports served as a test set; the remainder were used as domain-specific training data. The types and frequencies of PHI present within the reports were tallied. Five public de-identification tools were evaluated: MITRE Identification Scrubber Toolkit, U.S. National Library of Medicine–Scrubber, Massachusetts Institute of Technology de-identification software, Emory Health Information DE-identification (HIDE) software, and Neuro named-entity recognition (NeuroNER). The tools were compared using metrics including recall, precision, and F1 score (the harmonic mean of recall and precision) for each category of PHI.

Results: The annotators identified 3528 spans of PHI text within the 2503 reports. Cohen κ for interrater agreement was 0.938. Dates accounted for the majority of PHI found in the dataset of radiology reports (n = 2755 [78%]). The two best-performing tools both used machine learning methods—NeuroNER (precision, 94.5%; recall, 92.6%; microaveraged F1 score [F1], 93.6%) and Emory HIDE (precision, 96.6%; recall, 88.2%; F1, 92.2%)—but none exceeded 50% F1 on the important patient names category.

Conclusion: PHI appeared infrequently within the corpus of reports studied, which created difficulties for training machine learning systems. Out-of-the-box de-identification tools achieved limited performance on the corpus of radiology reports, suggesting the need for further advancements in public datasets and trained models.

Supplemental material is available for this article.

Much of the data in electronic medical records is in the form of clinical notes, which are largely written using unstructured or loosely structured text. This information is immensely valuable for clinical care, quality improvement, and medical research, particularly in the modern era of deep learning technologies. Major advances in automated natural language processing systems, which can process and generate natural language text, have great potential to improve health care quality and research.

The radiology report is often considered the final product of the radiology department, as it contains the reasoning, interpretations, diagnostic conclusions, and recommendations of the radiologist. While a variety of public repositories for medical images exist (1,2), there are few public repositories of radiology reports, and most such repositories contain reports from only one institution (3,4). Development of large-scale report repositories consisting of varied reports from many institutions could enable a variety of new descriptive studies (5,6) and quality improvement tools. However, making these reports available to researchers outside the clinical care stream first requires Health Insurance Portability and Accountability Act (HIPAA)–compliant de-identification (7).

Effective de-identification of protected health information (PHI) within clinical free-text notes remains challenging. A variety of systems designed to de-identify clinical notes has been researched and published, and most of these systems have been trained and evaluated on general purpose unstructured clinical notes (8–13). We were unable to find any system built for or tested primarily on radiology reports. The frequency and distribution of PHI in radiology reports differs from those in other clinical notes, so systems that perform best on general purpose clinical notes may not perform optimally on radiology reports. Studies from other clinical domains have demonstrated that specific domain adaptation is often necessary to achieve acceptable performance (14,15).

Performance standards for an automated de-identification system depend strongly on the intended use of the de-identified documents and the intended viewers. Performance standards also vary with the distribution of PHI. Some categories, such as patient names, may be more

Abbreviations

HIDE = Health Information DE-identification, NeuroNER = Neuro named-entity recognition, NLM = U.S. National Library of Medicine, MIST = MITRE Identification Scrubber Toolkit, MIT deid = Massachusetts Institute of Technology de-identification software, MRN = medical record number, PHI = protected health information

Summary

Out-of-the-box software tools for de-identification of general purpose clinical text did not reach acceptable performance for clinical or research use on a dataset of 2503 annotated radiology reports.

Key Points

- Five out-of-the-box software tools designed for de-identification of general purpose clinical text on a manually annotated corpus of 2503 radiology reports were evaluated.
- Machine learning systems outperformed rule-based systems on the radiology report corpus, with the best-performing system (Neuro named-entity recognition [NeuroNER]) achieving a token-level F1 score of 93.6%, below acceptable levels for clinical use (95% recall) on sensitive categories of protected health information (PHI).
- PHI is relatively rare in the corpus of radiology reports at the authors' institution and consists largely of dates; however, patient names and medical record numbers are still present in rare amounts, making it difficult to create a large, varied, and unbiased sample of positive training examples.

sensitive than others, such as dates of admission. No system can hope to achieve perfect performance. A comparative study established a standard for "acceptable" performance, with recall of 95% and precision of 80% for both patient names and social security numbers, and recall of 85% and precision of 70% for other PHI types (16). For particularly sensitive use cases, an even higher performance may be desired.

In this study, we evaluated and compared existing publicly available de-identification tools on a dataset of radiology reports. We built a large dataset randomly sampled from our entire institutional adult radiology report database, manually annotated these reports to identify the PHI, and evaluated the performance of five publicly available de-identification tools. We compared performance on specific PHI types in order to understand the strengths and weaknesses of the available systems. In addition, we quantified the prevalence of various types of PHI within our institution's reports to demonstrate the need for readily available de-identification tools. Although other de-identification studies have been published (5,16–18), few evaluate the effectiveness of existing software packages available off the shelf to clinician researchers, and none specifically evaluate such systems on radiology reports (a domain with different PHI distributions than other medical texts).

Materials and Methods

Data Collection

This retrospective study used data collected for nonresearch purposes and was approved by our institutional review board. A random sample of 2503 radiology reports was extracted from our institution's database of adult radiology reports collected between January 1, 2012 and January 8, 2019. The reports came from multiple hospitals within a single academic health system and included all patients over the age of 18 years. A subset of 1023 reports was designated as the testing set. Of these reports, 500 were labeled by two annotators to establish a measurement of interannotator reliability and produce high-accuracy labels to ensure accurate performance comparison. The entire text of each report was annotated, including structured template headers and footers. Further description of the dataset is provided in the Results section.

PHI Definition

We used the Safe Harbor method from the HIPAA Privacy Rule to define the standardized types of PHI for our algorithmic comparisons (7) (see Appendix E1 [supplement]). In addition to the original 18 categories, we included three extra PHI categories, including "clinician names," "hospital/institution names," and "vendor and tool names," as this information can be used to identify the location at which a patient received care. The complete set of PHI-type labels is provided in Table 1.

Labeling Procedure

Data annotation was performed by two of the authors (J.M.S. and J.A.) using a custom labeling application. PHI within the documents was annotated at the word-token level, that is, for each word in the document, the annotation specified whether it was part of an element of PHI and, if so, which type of PHI. Tokenization was performed using spaCy (Explosion AI, Berlin, Germany) Python package. Annotations consisted of the start and end positions of a PHI instance within a document, as well as its type (eg, patient name, patient medical record number [MRN]). Conflicts were resolved by discussion and consensus.

Included Software Packages

A review of the literature identified the major publicly available tools for de-identification of clinical text that are available as off-the-shelf software packages (see Appendix E2 [supplement]). Note that this study aimed to compare out-of-the-box performance of tools designed to be used by clinical users without substantial software development experience or machine learning knowledge. Furthermore, we did not evaluate commercial tools. A comparison of these tools is provided in Table 2.

De-Identification Software Comparison

The 1023 reports comprising the test set were used to compare the performance of the software tools. For the machine learning–based algorithms, which learn features from data, we used the remaining 1480 labeled reports as a training dataset. The rule-based algorithms neither learn from nor customize themselves to new data, so the training reports were not used in the evaluation of these tools.

Some of these tools have modifiable parameters that can substantially affect the algorithm performance. For instance, the MITRE Identification Scrubber Toolkit (MIST) (*http://*

mist-deid.sourceforge.net/) allows users to set their own values for various algorithmic parameters (eg, learning rate, gradient descent algorithm, etc). Identifying the optimal parameter settings for a particular task can be difficult, time-consuming, and task dependent. As this study aimed to evaluate readily available tools for nonspecialist users rather than state-of-the-art task-optimized machine learning algorithms, we opted to test the tools using their default parameters rather than perform exhaustive parameter comparisons. For further details on our task definition and parameter selection, see Appendix E3 (supplement).

Table 1: The 21 Categories of Protected Health Information Available to Annotators

Category	Description
1	Names of patients, family members, employers, etc
2	Names of health care workers
3	Names of hospitals, clinics, or other health care organizations
4	Geographical subdivisions smaller than a state, including address, city, county, precinct, and zip code
5	All elements of dates (dates directly related to an individual, including birth date, admission date, discharge date, date of death, and all ages over 89 years, as well as all elements of dates indicative of such age)
6	Phone numbers
7	Fax numbers
8	E-mail addresses
9	Social security numbers
10	Medical record numbers
11	Health plan beneficiary numbers
12	Account numbers
13	Certificate or license number
14	Vehicle identification and serial numbers
15	Device identification and serial numbers
16	Web URLs
17	IP addresses
18	Biometric identifiers
19	Full-face photographic images and any comparable images
20	Any other unique identifying number, characteristic, or code
21	Names of vendors, software, tools, or other institution-specific content

MIST.—MIST (9) comprises a data labeler and a conditional random field-based machine learning system (Carafe; *https://sourceforge.net/projects/carafe/)* (20). As it is a machine learning system, training data must be supplied. However, any number of valid types of PHI tags can be provided.

NLM-Scrubber.—The U.S. National Library of Medicine (NLM)−Scrubber (*https://scrubber.nlm.nih.gov/)* is an end-to-end pipeline with minimal parameter customization that uses rule-based systems to identify four major categories of PHI in clinical reports: names, addresses, dates (including ages), and alphanumeric identifiers (10). For evaluation of this software, we group our 21 PHI categories into these groups.

Emory HIDE.—The Emory Health Information DE-identification (HIDE) platform (*http://www.mathcs.emory.edu/hide/index.html*) is a multicomponent suite of tools designed to de-identify structured and unstructured data, including a machine learning model based on conditional random fields (11). HIDE uses the CRFSuite platform (*http://www.chokkan.org/software/crfsuite/*) and a variety of custom hand-engineered features.

Table 2: Details of Software Packages Evaluated for this Study

Tool	Original Release Year	Major Algorithmic Techniques	Data Used for Creation and Validation
MIST (9)	2010	Conditional random fields	1200 discharge summaries, laboratory reports, letters, and order summaries
NLM-Scrubber (10)	2014	Rules, dictionaries	3093 clinical free-text documents (unspecified)
Emory HIDE (11)	2009	Conditional random fields	100 pathology reports
MIT deid (12)	2008	Rules, dictionaries	2434 nursing notes
NeuroNER (8,13)	2017	Recurrent neural networks, conditional random fields	2939 free-text medical notes (multiple types) from the I2B2 2014 and MIMIC datasets

Note.—HIDE = Health Information DE-identification, I2B2 = Informatics for Integrating Biology and the Bedside 2014 De-identification and Heart Disease Risk Factors Challenge, MIT deid = Massachusetts Institute of Technology de-identification, MIMIC = Medical Information Mart for Intensive Care, MIST = MITRE Identification Scrubber Toolkit, NeuroNER = Neuro named-entity recognition, NLM = U.S. National Library of Medicine.

MIT deid.—The Massachusetts Institute of Technology de-identification (MIT deid) platform (*https://physionet.org/content/deid/1.1/*) uses a host of regular expressions and dictionaries to de-identify PHI from texts (12). It uses no machine learning components.

NeuroNER.—The Neuro named-entity recognition (NeuroNER) software (*http://neuroner.com/*) is a machine learning model that uses recurrent neural networks to identify various PHI forms (8,13). It can be used either as a Python module or as a command-line tool. It also includes files corresponding to pretrained word vectors

and neural network models trained on existing corpora of PHI.

Statistical Analysis and Evaluation Metrics

We provide descriptive statistics of our annotated dataset, including the frequency of PHI by category and the number of documents containing PHI. Interannotator reliability was calculated at the token level using the Cohen κ coefficient and two predicted classes (PHI vs no PHI).

The five PHI de-identification software packages were compared by evaluating token-level recall, precision, and F1 score (the harmonic mean of precision and recall). In PHI de-identification, recall may be considered more important than precision because the consequence of a false positive (PHI is disclosed) is substantially more harmful than that of a false negative (informative information is removed from the report). In this study, we report all three metrics. We also report token-level performance on each subcategory of PHI (eg, patient names, dates) to better characterize model strengths and weaknesses. Last, we report document-level performance metrics for each system, that is, the number of reports that retained at least one element of PHI and the number of reports with at least one non-PHI element falsely redacted.

Table 3: Frequency of Protected Health Information within the 2503 Radiology Reports

Type of PHI	No. of Text Spans	No. of Unique Text Spans
Dates*	2755	1957
Names of health care workers	360	302
Names of hospitals or clinics or care delivery systems	169	70
Names of software, tools, institution-specific vendor content	86	47
Geographical subdivisions smaller than a state†	62	9
Any other unique identifying number, characteristic, or code	49	34
Names of patients or family members	21	15
Medical record numbers	20	14
Phone numbers	6	3

Note.—Frequency of each type of protected health information (PHI) ranked from most to least frequently occurring. In addition, we list the number of unique text spans that fall into a given category. For example, if the PHI phrase "Dr Jones" shows up in three different notes, this counts as only one unique text span. There were no text spans of other types of PHI (fax numbers; mail addresses; social security numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; vehicle identifiers and serial numbers, including license plate; web URLs; IP addresses; biometric identifiers, including finger and voice prints; full-face photographic images and any comparable images) for either table column.
* All elements of dates for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 years and all elements of dates indicative of such age.
† Includes address, city, county, precinct, and/or zip code.

Results

Dataset

The dataset of 2503 reports included more than 200 different imaging examination protocols, with a variety of different system and personal templates, including entirely free-text reports. The dataset consisted of 633 554 tokens, with 254 862 tokens in the test set. Our institution makes routine use of templates and other structured elements. However, it is difficult to know simply by looking at a report how much of it was generated through templated text, so we cannot give precise estimates about the frequency of templated information. Most reports in our dataset include at least some templated information, although the degree of structure varies. Some templates simply have structured findings and impression segments, while others have prespecified organ-level structure or prepopulated default normal finding blocks.

In total, there were 3528 spans of text corresponding to PHI. Dates were by far the most common form of PHI (n = 2755 [78.1%]), followed by names of physicians or other health care workers (n = 360 [10.2%]), names of hospitals or care delivery systems (n = 169 [4.8%]), and names of software tools or vendors (n = 86 [2.4%]). The presence of patient or family member names was comparatively rare in radiology reports compared with other types of PHI (n = 21 [0.6%]), and many types, including e-mail address and IP addresses, did not appear a single time within our sample. The frequency of each type of PHI within our dataset, as well as the number of unique text spans that comprise each category, are provided in Table 3. Of all 2503 reports, 1567 (62.6%) contained at least one instance of PHI, while 936 (37.4%) contained no PHI. However, only 411 (16.4%) included PHI other than dates directly related to a patient.

The Cohen κ for the doubly labeled data, evaluated as a two-class task (PHI vs not PHI) at the token level, was 0.938 (95% confidence interval: 0.925, 0.950). The discrepancies between the two annotators were entirely due to cases in which one annotator failed to notice a span of PHI while the other noticed it; there were no disagreements about whether a span of text constituted PHI.

Performance Comparison

The Figure shows the overall token-level performance (including recall, precision, and F1 score for all PHI types) of each of the evaluated software packages on our test set consisting of 1023 radiology reports.

MIST.—The performance of MIST on the various categories is provided in Tables 4, 5, and 6. It had relatively high

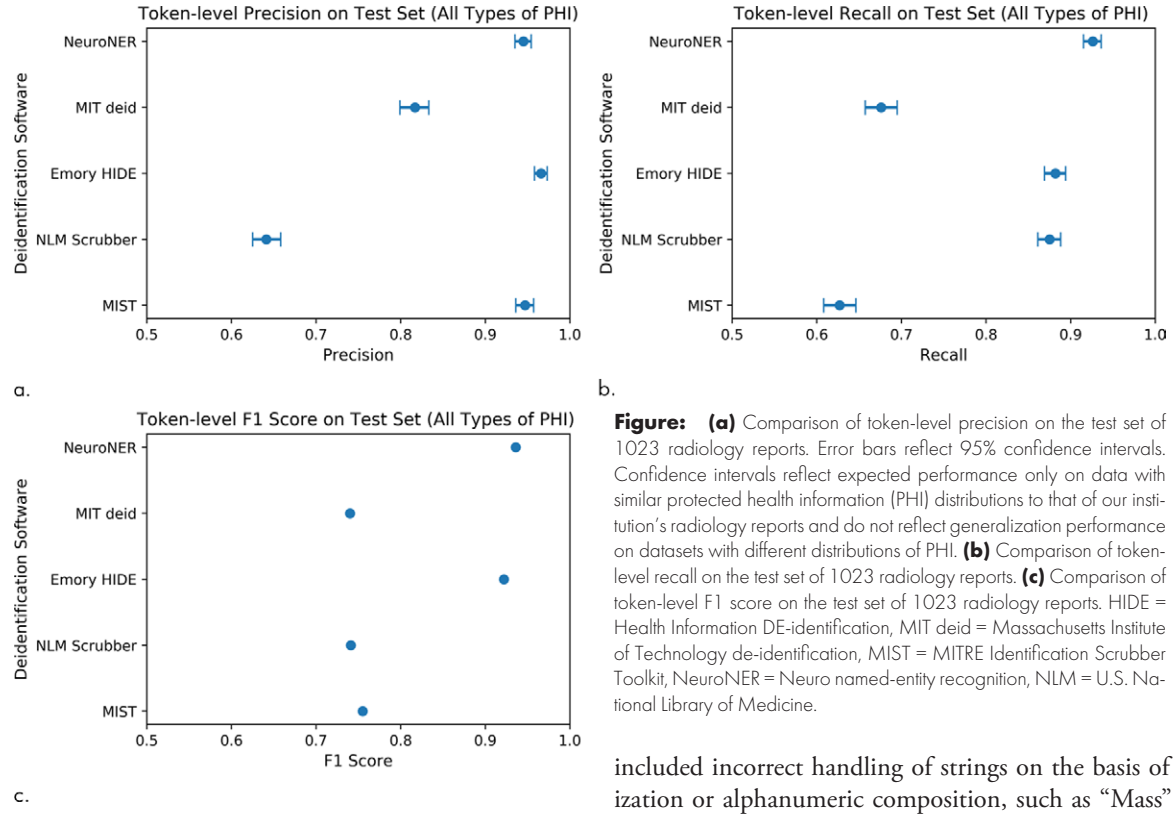

Figure: **(a)** Comparison of token-level precision on the test set of 1023 radiology reports. Error bars reflect 95% confidence intervals. Confidence intervals reflect expected performance only on data with similar protected health information (PHI) distributions to that of our institution's radiology reports and do not reflect generalization performance on datasets with different distributions of PHI. **(b)** Comparison of token-level recall on the test set of 1023 radiology reports. **(c)** Comparison of token-level F1 score on the test set of 1023 radiology reports. HIDE = Health Information DE-identification, MIT deid = Massachusetts Institute of Technology de-identification, MIST = MITRE Identification Scrubber Toolkit, NeuroNER = Neuro named-entity recognition, NLM = U.S. National Library of Medicine.

precision for most categories but poor recall. Of the 1023 documents in the test set, 40 (3.9%) had at least one non-PHI token falsely removed from the document, while 395 (38.6% of all documents [61.1% of documents containing PHI]) retained PHI after application of the tool. Of these, the most commonly missed PHI categories were as follows: 370 documents contained retained dates, 45 contained names of health care providers, seven contained patient names, and seven contained MRNs. On qualitative inspection, the majority of reports with falsely redacted PHI remained largely intact and readable. Even for categories with a large number of varied examples (date, health care provider), the system did not perform at an acceptable level. Many normal date strings and provider names that appeared in stereotyped locations were not caught by the system.

NLM-Scrubber.—The performance of the NLM-Scrubber tool is provided in Tables 4, 5, and 6. It achieved a precision of 98.3%, a recall of 97.5%, and an F1 score of 97.9% on the most common date category but markedly lower scores on the other categories. On the document level, 300 of the 1023 documents (29.3%) had at least one non-PHI token falsely removed from the document, and 119 documents (11.6% of all documents [18.4% of documents containing PHI]) retained PHI after scrubbing. The most common categories for retained PHI were personal names (73 documents), dates (29 documents), addresses (20 documents), and alphanumeric identifiers (four documents). Most of the false negatives consisted of typographical errors or nontraditionally formatted date strings (eg, "to-1–2018"), while the false positives in other categories included incorrect handling of strings on the basis of capitalization or alphanumeric composition, such as "Mass" in "No Mass" redacted as an address, "CT" in "CT scan" redacted as an address, "Osgood–Schlatter" in "Osgood–Schlatter disease" redacted as a personal name, and a comment on "L4-L5" vertebrae redacted as an alphanumeric identifier. These types of false redactions have the potential to substantially compromise the readability of the report.

Emory HIDE.—The performance of the HIDE software is provided in Tables 4, 5, and 6. It performed well on the date, phone number, and identifying code categories, likely owing to the inclusion of a variety of orthographic features (eg, "starts with a digit") into the machine learning classifier. However, on other categories, including names, addresses, and vendor tools, it did not perform as well. As HIDE is a machine learning algorithm, it is possible that with additional training examples the software could perform better on these categories. Of the 1023 documents, 39 (3.8%) had at least one PHI token falsely redacted, while 80 documents (7.8% of all documents [12.4% of documents containing PHI]) retained PHI after the tool was applied. The most common categories for retained PHI were names of health care providers (31 documents), dates (27 documents), location names and addresses (18 documents), vendor tools (12 documents), and patient names (seven documents). A qualitative inspection revealed that the readability for most of the reports containing false positives remained intact.

MIT deid.—The performance of the MIT deid system is given in Tables 4, 5, and 6. It was not able to generalize effectively our radiology report dataset, although it did achieve a performance of 89.0% F1 on the dates category. Of the 1023 documents, 181 (17.7%) had PHI falsely redacted, and 256 documents

Table 4: Token-level Precision for Each of the De-Identification Tools

	MIST		NLM-Scrubber		Emory HIDE		MIT deid		NeuroNER	
Variable	Value (%)	95% CI	Value (%)	95% CI	Value (%)	95% CI	Value (%)	95% CI	Value (%)	95% CI
All PHI	94.7	93.6, 95.7	64.1	62.5, 65.8	96.6	95.8, 97.3	81.7	79.9, 83.3	94.5	93.5, 95.4
Patient names	100	51.0, 100	37.3	34.1, 40.6	0*	NA	37.9	33.3, 42.7	100	75.8, 100
Health care provider names	93.0	89.6, 95.3			97.5	95.0, 98.7			82.0	78.1, 85.3
Vendor tools	86.7	75.8, 93.1			88.6	76.0, 95.0	28.6	8.2, 64.1	82.0	69.2, 90.2
Health care location names	85.2	78.4, 90.1			93.4	87.7, 96.7	51.1	41.0, 61.2	82.0	78.1, 85.3
Addresses/geographic locations	94.9	86.1, 98.3	15.2	11.5, 19.9	97.8	88.4, 99.6			98.1	90.1, 99.7
Dates	97.4	96.2, 98.2	98.3	97.6, 98.9	96.8	95.8, 97.5	96.0	94.8, 96.8	98.4	97.7, 98.9
Phone numbers	100	72.2, 100	19.9	16.4, 23.9	100	72.2, 100	0*	NA	0*	NA
MRNs	0*	NA			0*	NA	0*	NA	100	64.6, 100
Other identifying codes	100	48.9, 100			98.1	(90.2, 99.7)	0*	NA	81.1	70.4, 88.6

Note.—The 95% confidence intervals (CIs) for proportional metrics are calculated using the Wilson score with the Python statsmodels package (*https://www.statsmodels.org/stable/index.html*). The "All PHI" row weights each protected health information (PHI) token equally. Note that token counts may differ between different models due to using prepackaged tokenizers that come as part of the software package, or redefining our PHI categories to accord with the original algorithm specifications (eg, counting or not counting "Dr" as part of a name, depending on the algorithm). HIDE = Health Information DE-identification, MIT deid = Massachusetts Institute of Technology de-identification, MRN = medical record number, MIST = MITRE Identification Scrubber Toolkit, NER = named-entity recognition, NLM = U.S. National Library of Medicine, NA = not applicable.

* Indicates a zero denominator for the proportion (ie, the model never predicted the PHI category within the test set).

(25.0% of all documents [39.6% of documents originally containing PHI]) retained PHI after usage of the tool. The most common categories of retained PHI were dates (166 documents), names (71 documents), locations (63 documents), and vendor tools (21 documents). Most of the false positives consisted of capitalized words at the beginning of sentences, which may have the potential to substantially compromise the readability of the report.

NeuroNER.— The performance of the NeuroNER system is given in Tables 4, 5, and 6. NeuroNER is able to achieve a high performance on the majority of categories. Of the 1023 documents, 34 (3.3%) included at least one token with falsely redacted PHI, while 91 documents (8.9% of all documents [14.1% of documents with PHI]) retained PHI after application of the tool. The most common types of retained PHI were names of health care providers (37 documents), dates (32 documents), addresses or geographic names (20 documents), patient names (seven documents), and MRNs (seven documents). The reports containing false positives remained intact and readable. As with the other machine learning systems, NeuroNER is likely hampered by the small number of positive examples of rarer categories in our training dataset.

Discussion

Our radiology report dataset represents a substantial departure from general purpose clinical text de-identification datasets. For instance, the amount of PHI per document is substantially lower, and the distribution of PHI types heavily skews toward date strings. This likely reflects a concerted effort by our radiology department to keep nonessential PHI out of the reports. However, appearances of nondate PHI are still common (16.4% of reports in our corpus). This imbalance makes it more difficult to assemble training corpora for machine learning systems, which require a wide variety of positive examples of PHI to learn their general features. Even after hand-annotating more than 2500 radiology reports, there were still only 15 unique patient or family member names included in our dataset. Unfortunately, using shortcuts to identify PHI, such as using string matching or regular expressions to prescreen reports for likely PHI instances, introduces substantial bias into the test set, as any PHI instances that do not follow those rules will be missed. It is likely that large initiatives that aggregate data from multiple institutions to build large varied datasets are necessary to create robust and generalizable de-identification systems.

Dates were by far the most common type of PHI included in our radiology report dataset. This is perhaps unsurprising, as radiologists frequently make comparisons to previous studies, which are often referenced multiple times within the same report. In some cases, dates formatted as "to-for-2015" or "20/17" were found, likely owing to dictation software or human input error. While not formatted properly as dates, these text spans certainly represent PHI. Conversely, many text spans formatted identically to a date were actually imaging series number references (eg, "5/7" for image "5 of 7"). Therefore, it is likely impossible

Table 5: Token-level Recall for Each of the De-Identification Tools

	MIST		NLM-Scrubber		Emory HIDE		MIT deid		NeuroNER	
Variable	Value (%)	95% CI	Value (%)	95% CI	Value (%)	95% CI	Value (%)	95% CI	Value (%)	95% CI
All PHI	62.7	60.8, 64.6	87.5	86.1, 88.8	88.2	86.9, 89.4	67.6	65.7, 69.5	92.6	91.5, 93.6
Patient names	10.5	4.2, 24.1	57.8	53.6, 61.9	0	0, 9.2	47.5	42.1, 53.0	31.6	19.1, 47.5
Health care provider names	71.4	66.8, 75.6			77.9	73.5, 81.7			92.6	89.6, 94.8
Vendor tools	89.7	79.2, 95.2			67.2	54.4, 77.9	3.4	0.1, 11.7	70.7	58.0, 80.8
Health care location names	74.7	67.5, 80.8			74.7	67.3, 80.9	22.0	16.9, 28.1	77.9	70.7, 83.7
Addresses/geographic locations	88.9	78.8, 94.5	68.3	56.0, 78.4	74.6	62.2, 83.9			88.1	77.5, 94.1
Dates	61.2	58.9, 63.5	97.5	96.7, 98.2	96.0	94.9, 96.8	83.0	81.1, 84.7	97.5	96.6, 98.1
Phone numbers	62.5	38.6, 81.5	95.6	89.1, 98.3	62.5	38.6, 81.5	0	0, 19.3	0	0, 19.4
MRNs	0.0	0.0, 25.9			0	0.0, 25.9	0	0.0, 25.9	63.6	35.4, 84.8
Other identifying codes	5.9	0, 13.8			84.1	73.1, 91.1	0	0.0, 5.7	88.9	78.8, 94.5

Note.—The 95% confidence intervals (CIs) for proportional metrics are calculated using the Wilson score with the Python statsmodels package (*https://www.statsmodels.org/stable/index.html*). The "All PHI" row weights each protected health information (PHI) token equally. Note that token counts may differ between different models due to using prepackaged tokenizers that come as part of the software package, or redefining our PHI categories to accord with the original algorithm specifications (eg, counting or not counting "Dr" as part of a name, depending on the algorithm). HIDE = Health Information DE-identification, MIT deid = Massachusetts Institute of Technology de-identification, MRN = medical record number, MIST = MITRE Identification Scrubber Toolkit, NeuroNER = Neuro named-entity recognition, NLM = U.S. National Library of Medicine.

to devise a set of rules that capture all dates in a report; for a system to perform well, it must leverage the surrounding context words in the document. Similarly, the large number of unique text spans representing health care worker names makes it difficult to predict them all ahead of time with a list of names or regular expression patterns.

None of the evaluated systems performed at an acceptable level for clinical or research use, particularly on the highly sensitive categories of PHI (patient names, MRNs, phone numbers; desired performance > 95% recall), which were rare in our dataset. The rule-based models, which were designed to identify PHI on the basis of handwritten text patterns detected in different corpora, are prone to false positives, often falsely redacting informative eponyms, such as Osgood–Schlatter disease. The machine learning models, which can be trained on the unique features of new datasets, varied in performance. While both HIDE and MIST used conditional random fields, the HIDE system uses a wider variety of hand-selected word features, including orthographic features, such as "begins with digit," allowing it to learn more complex dependencies between features and perform better. The NeuroNER model also leverages pretrained word vectors constructed from large amounts of general purpose English text, with no hand-selected features, and performs similarly to the HIDE system. Future improvement in radiology report de-identification may require large multi-institutional public datasets. A large report text corpus could enable the training of systems to handle the typical distribution of PHI in radiology reports, such as the rare occurrence of patient names, frequent appearance of dates, and radiology-specific eponyms and capitalization patterns.

Modern machine learning systems achieve state-of-the-art performance in a wide variety of complex language tasks, including named entity recognition tasks such as PHI de-identification (13,21). It is likely that such systems would also perform well on this task with a large enough dataset. However, in our literature search, we were unable to find many neural network–based tools available as off-the-shelf packages designed for use by the average clinician-researcher (with the exception of NeuroNER). Thus, most of these tools are outside the scope of this study. In future work, we plan to evaluate (and build additional) such systems on our dataset. However, it would be wise for those building and evaluating such systems to keep in mind the ultimate goal of more general availability.

One major limitation of our work was the small amount of positive PHI examples for certain important categories (patient name, MRN) found in our dataset. Not only does this lead to wide confidence intervals on the performance estimates for the rarer categories, it also limits the machine learning models' ability to learn and generalize from data. Larger datasets would likely enable more effective comparison and training, but the amount of manual labor required to label such datasets makes the task difficult. This represents a substantial practical concern with respect to off-the-shelf use of these models. Another limitation of our work was that our dataset consists of reports from only one health care system. Our reporting patterns or system-wide templates may differ

from those used at other institutions in ways that substantially alter the performance of de-identification tools. Last, there are commercially available tools that offer similar services and that fall outside the scope of this study.

In summary, we have built a labeled dataset consisting of 2503 radiology reports across all imaging modalities, departments, and anatomic regions and provided descriptive analyses of the PHI patterns. We have identified and evaluated major public off-the-shelf software tools on this dataset to establish performance metrics. In future studies, we will evaluate modern neural network models on the same task and attempt to build systems that achieve an acceptable level of performance. Our ultimate aim is to build large, cross-institutional repositories of de-identified radiology reports to enable new types of research studies and software tools that leverage the promise of unstructured natural language data.

Table 6: Token-level F1 Scores for Each of the De-Identification Tools

Variable	MIST (%)	NLM-Scrubber (%)	Emory HIDE (%)	MIT deid (%)	NeuroNER (%)
All PHI	75.5	74.1	92.2	74.0	93.6
Patient names	19.0	45.4	0	42.1	48.0
Health care provider names	80.8		86.6		87.0
Vendor tools	88.1		76.5	6.2	75.9
Health care location names	79.6		83.0	30.8	86.3
Addresses/geographic locations	91.8	24.9	84.6		92.9
Dates	75.1	97.9	96.4	89.0	97.9
Phone numbers	77.0	33.0	76.9	0	0
MRNs	0		0	0	77.8
Other identifying codes	11.1		90.6	0	84.8

Note.—HIDE = Health Information DE-identification, MIT deid = Massachusetts Institute of Technology de-identification, MRN = medical record number, MIST = MITRE Identification Scrubber Toolkit, NeuroNER = Neuro named-entity recognition, NLM = U.S. National Library of Medicine.

Acknowledgments: Charles Chambers and Darco Lalevic contributed substantially to the data acquisition for this study.

Author contributions: Guarantors of integrity of entire study, J.M.S., T.S.C.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.M.S., T.P., T.S.C.; experimental studies, T.P., T.S.C.; statistical analysis, J.M.S., J.A., T.S.C.; and manuscript editing, all authors.

Disclosures of Conflicts of Interest: J.M.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: cofounder and 25% owner of River Records, a health care start-up company focused broadly on researching, implementing, and disseminating solutions to improve clinician documentation software and electronic medical records and reduce clinician documentation burden; author has not received financial compensation from the company nor was the company involved in this study. Other relationships: disclosed no relevant relationships. **T.P.** disclosed no relevant relationships. **J.A.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received airfare expenses from the RSNA for travel to the RSNA 2019 Annual Meeting. Other relationships: disclosed no relevant relationships. **C.E.K.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: consulting fee and travel expenses from Olympus Corporation of America to present on AI at their Frontiers of Endoscopy meeting; institution receives salary support as Editor of *Radiology: Artificial Intelligence.* **T.S.C.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: board member of SIIM, AUR, PRS, PRRS; received travel reimbursement from SIIM for a board retreat and from PRS for board meetings; royalties from the Osler Institute for cardiac MRI lectures given in 2012.

References

1. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–1057.
2. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE, 2017; 3462–73471.
3. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23(2):304–310.
4. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(1):160035.
5. Jiang X, Sarwate AD, Ohno-Machado L. Privacy technology to support data sharing for comparative effectiveness research: a systematic review. Med Care 2013;51(8 Suppl 3):S58–S65.
6. Aryanto KYE, van Kernebeek G, Berendsen B, Oudkerk M, van Ooijen PMA. Image De-Identification Methods for Clinical Research in the XDS Environment. J Med Syst 2016;40(4):83.
7. Office for Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Information Portability and Accountability Act (HIPAA) Privacy Rule. Office of Civil Rights; 2012. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. Accessed March 8, 2020.
8. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. ArXiv [cs.CL] [preprint]. http://arxiv.org/abs/1705.05487. Posted 2017. Accessed March 8, 2020.
9. Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010;79(12):849–859.
10. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA Annu Symp Proc 2014;2014:767–776.
11. Gardner J, Xiong L. An integrated framework for de-identifying unstructured medical data. Data Knowl Eng 2009;68(12):1441–1451.
12. Neamatullah I, Douglass MM, Lehman L-WH, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8(1):32.
13. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017;24(3):596–606.
14. Lee HJ, Zhang Y, Roberts K, Xu H. Leveraging existing corpora for de-identification of psychiatric notes using domain adaptation. AMIA Annu Symp Proc 2018;2017:1070–1079.
15. Henriksson A, Kvist M, Dalianis H. Detecting Protected Health Information in Heterogeneous Clinical Notes. Stud Health Technol Inform 2017;245:393–397.
16. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol 2012;12(1):109.

17. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res 2019;21(5):e13484.
18. Langarizadeh M, Orooji A, Sheikhtaheri A. Effectiveness of Anonymization Methods in Preserving Patients' Privacy: A Systematic Literature Review. Stud Health Technol Inform 2018;248:80–87.
19. Ramshaw LA, Marcus MP. Text Chunking Using Transformation-Based Learning. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, editors. Natural Language Processing Using Very Large Corpora. Dordrecht: Springer Netherlands; 1999. p. 157–76.
20. Wellner B. Sequence models and ranking methods for discourse parsing [Doctoral thesis]. Waltham, Mass: Brandeis University, 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.7881&rep=rep1&type=pdf. Accessed April 6, 2020.
21. Khin K, Burckhardt P, Padman R. A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation. ArXiv [cs.CL] [preprint]. http://arxiv.org/abs/1810.01570. Posted 2018. Accessed March 8, 2020.