journal homepage: www.ijmijournal.com

Bootstrapping a de-identification system for narrative patient records: Cost-performance tradeoffs

David Hanauer[a,*], John Aberdeen[b], Samuel Bayer[b], Benjamin Wellner[b], Cheryl Clark[b], Kai Zheng[c,d], Lynette Hirschman[b]

[a] Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA
[b] The MITRE Corporation, Bedford, MA, USA
[c] School of Public Health Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA
[d] School of Information, University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form
12 February 2013
Accepted 25 March 2013

Keywords:
Privacy [I01.880.604.473.352.500]
Natural language processing
[L01.224.065.580]
NLP
Electronic health records
[E05.318.308.940.968.625.500]
Medical record systems
Computerized
[E05.318.308.940.968.625]
Medical informatics [L01.313.500]

ABSTRACT

Purpose: We describe an experiment to build a de-identification system for clinical records using the open source MITRE Identification Scrubber Toolkit (MIST). We quantify the human annotation effort needed to produce a system that de-identifies at high accuracy.

Methods: Using two types of clinical records (history and physical notes, and social work notes), we iteratively built statistical de-identification models by annotating 10 notes, training a model, applying the model to another 10 notes, correcting the model's output, and training from the resulting larger set of annotated notes. This was repeated for 20 rounds of 10 notes each, and then an additional 6 rounds of 20 notes each, and a final round of 40 notes. At each stage, we measured precision, recall, and F-score, and compared these to the amount of annotation time needed to complete the round.

Results: After the initial 10-note round (33 min of annotation time) we achieved an F-score of 0.89. After just over 8 h of annotation time (round 21) we achieved an F-score of 0.95. Number of annotation actions needed, as well as time needed, decreased in later rounds as model performance improved. Accuracy on history and physical notes exceeded that of social work notes, suggesting that the wider variety and contexts for protected health information (PHI) in social work notes is more difficult to model.

Conclusions: It is possible, with modest effort, to build a functioning de-identification system *de novo* using the MIST framework. The resulting system achieved performance comparable to other high-performing de-identification systems.

1. Introduction

The passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act and the recent introduction of federal incentive programs and meaningful use criteria have significantly accelerated the widespread adoption of electronic health records (EHRs) across the U.S. [1,2] It is projected that by 2015, nearly all U.S. primary care practices and hospitals will be equipped with certain levels of electronic documentation capabilities [1]. This transition creates an unprecedented opportunity to capture and use patient care data more effectively to achieve broad quality improvement and cost containment goals. It also has the potential to

[*] *Corresponding author at*: University of Michigan Medical School, 5312 CC, SPC 5940, 1500 E Medical Center Dr, Ann Arbor, MI 48109-5940, USA. Tel.: +1 734 615 0599; fax: +1 206 338 4213.
E-mail address: hanauer@med.umich.edu (D. Hanauer).

1386-5056/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijmedinf.2013.03.005

transform the U.S. healthcare system into a self-learning vehicle by supporting secondary use applications such as disease surveillance and clinical, translational, and health services research [1,3,4]. This situation is not unique to the U.S. Indeed, some nations are well ahead of the U.S. in EHR adoption [5,6].

While computable data recorded in discrete, structured fields are highly desirable, a considerable proportion of clinical data will likely continue to exist in a free-text narrative format [7,8]. Maximizing the secondary use of free-text narratives has proven to be extremely challenging. One prominent reason is that narrative documents contain a large quantity of protected health information (PHI), including names, dates, phone numbers, and other identifiers as defined by the Health Insurance Portability and Accountability Act (HIPAA) [9]. The removal of such information from narrative documents often requires costly processes and highly specialized technical expertise.

There are various computational approaches for automatically *scrubbing* narrative data by identifying and erasing patient identifiers or substituting them with pseudo-identifiers. Commonly used methods range from simple string-matching algorithms based on regular expressions (predictable text patterns), [10–12] to more intelligent machine learning (ML) approaches that use adaptive statistical models learned from training data [13–15]. Such statistical models are generally more flexible and more versatile than rule-based or heuristic systems, thus reducing the need for ongoing human-supervised maintenance and tuning. In addition, this approach puts the ability to build these models directly in the hands of the subject matter experts, bypassing the need for linguistic or programming expertise. However, such statistical models intrinsically contain PHI elements originating in training data and thus cannot be readily shared outside of a HIPAA covered entity. Therefore, such systems must be set up locally at each individual adopting institution, which could represent a significant barrier for their widespread adoption.

In light of such constraints, we conducted an empirical experiment with an ML-based de-identification system. The primary objective of this study was to determine the feasibility, in terms of human effort required, of implementing the system de novo at an institution where there was no prior experience in building or using ML-based de-identification methods. The results, obtained by measuring the actual time burden necessary to train the system and the incremental gains in system accuracy while its model was being trained, may provide valuable insights into the practicality of implementing such systems in resource-limited settings. The other objective of the study was to characterize the distribution of PHI among various types of clinical documents. The findings, a delineation of varied frequencies and densities of PHI by document type, may provide a rational basis for more strategic selection of documents to train ML models more efficiently and comprehensively.

2. Methods

2.1. De-identification software

We used the MITRE Identification Scrubber Toolkit (MIST) [15], an open source de-identification tool freely available at SourceForge (http://mist-deid.sourceforge.net/). MIST uses conditional random fields to *learn* patterns in documents in order to identify candidate PHI strings for potential removal or substitution [15]. An advantage of this approach is that once the system is prepared (trained) using adequate training examples, it should work well on any comparable type of clinical document without further re-tuning [16].

MIST provides an intuitive graphical user interface (GUI) that shields end users from the complexity of the underlying ML algorithms (Fig. 1). Through the GUI, a user can annotate PHI elements appearing in training data to help the system refine its de-identification model for better performance. MIST also provides the capability of logging user interactions with the GUI. These log files contain two major types of events: (1) add_annotation, the user action to add a PHI element that MIST may have failed to identify; and (2) remove_annotation, the user action to remove a false-positive PHI element that MIST may have mislabeled. Much of the analysis reported in this paper is based on data drawn from these computer-recorded logs.

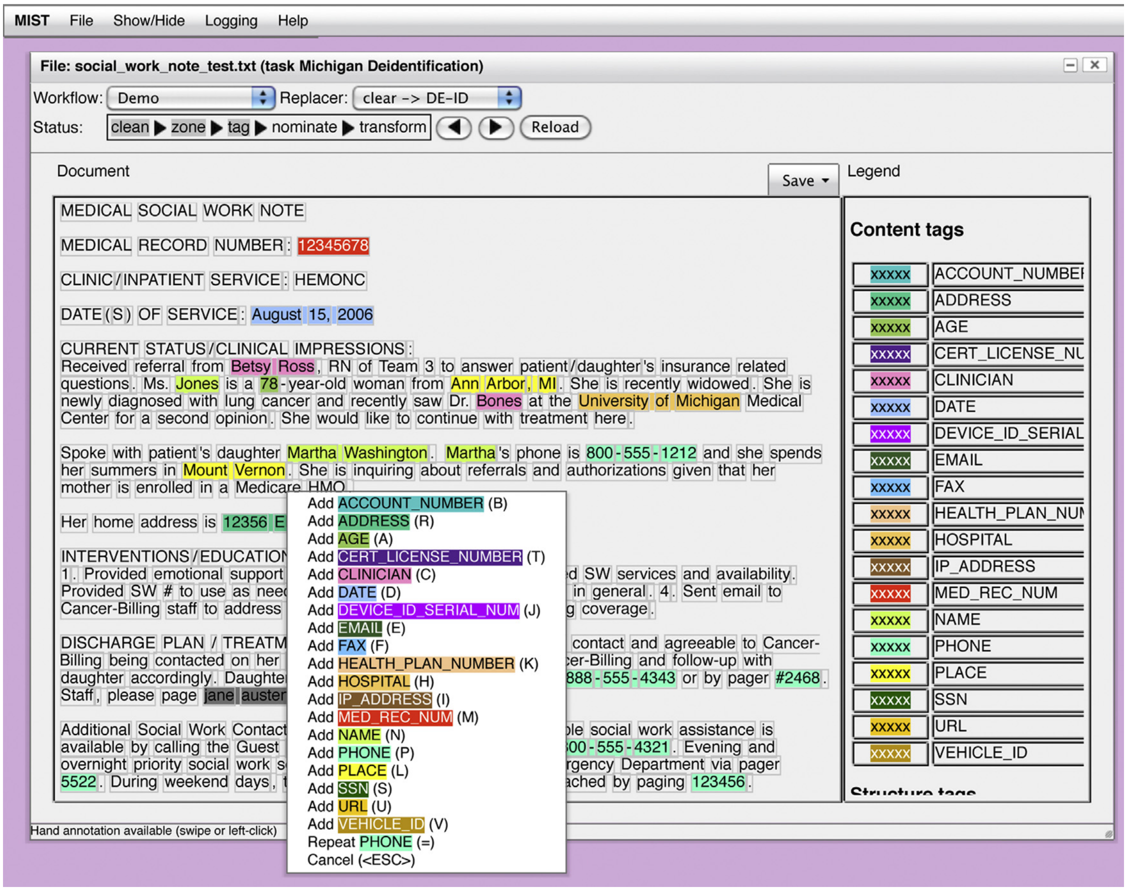
MIST is designed to handle an arbitrary number of PHI elements. In this study, we tested the system using 12 key PHI elements developed from standard HIPAA identifiers including NAME, AGE, ADDRESS, and PHONE (a full list is provided in Table 2). Note that for simplicity, certain HIPAA identifiers such as voice phone and fax were merged as a single PHI entity in this study.

2.2. Clinical setting

The empirical study was conducted at the University of Michigan Health System (UMHS), a large, tertiary care academic medical center that heavily utilizes patient care data for secondary use purposes. UMHS has had a homegrown web-based EHR system, CareWeb, since 1998, that provides a comprehensive repository of clinical data including medications, problem lists, and clinician notes [17]. The clinical documents stored in CareWeb exist in a free-text format, although some of them have been populated via structured or semi-structured templates. At UMHS, over 3 million new documents comprised of over 125 million lines of text are added to the repository each year. About 40% of all documents stored in CareWeb were created via dictation and transcription, with the rest entered via typing through computer keyboard [18].

2.3. Document types

More than 120 different document types might be created during patient care, including emergency department notes, admission notes, inpatient progress notes, discharge notes, and outpatient clinic notes. Additionally, healthcare professionals with a variety of roles, such as physicians, respiratory therapists, nurses, and social workers, collectively contribute to notes composition. Because the purpose of this study was to investigate the performance of ML-based de-identification software, we selected two types of narrative documents that often contain a large quantity and variety of PHI elements: admission history and physical (H&P) and social work (SW) notes. H&P notes generally contain significant PHI since they are the initial “intake” documents created when a patient is



MIST File Show/Hide Logging Help

File: social_work_note_test.txt (task Michigan Deidentification)

Workflow: Demo Replacer: clear -> DE-ID

Status: clean ▶ zone ▶ tag ▶ nominate ▶ transform ◀ ▶ Reload

Document Save ▾ Legend

MEDICAL SOCIAL WORK NOTE

MEDICAL RECORD NUMBER: 12345678

CLINIC/INPATIENT SERVICE: HEMONC

DATE(S) OF SERVICE: August 15, 2006

CURRENT STATUS/CLINICAL IMPRESSIONS:
Received referral from Betsy Ross, RN of Team 3 to answer patient/daughter's insurance related questions. Ms. Jones is a 78-year-old woman from Ann Arbor, MI. She is recently widowed. She is newly diagnosed with lung cancer and recently saw Dr. Bones at the University of Michigan Medical Center for a second opinion. She would like to continue with treatment here.

Spoke with patient's daughter Martha Washington. Martha's phone is 800-555-1212 and she spends her summers in Mount Vernon. She is inquiring about referrals and authorizations given that her mother is enrolled in a Medicare HMO.

Her home address is 12356 E

INTERVENTIONS/EDUCATION
1. Provided emotional support
Provided SW # to use as need
Cancer-Billing staff to address

d SW services and availability.
in general. 4. Sent email to
g coverage.

DISCHARGE PLAN / TREATM
Billing being contacted on her
daughter accordingly. Daughte
Staff, please page jane auster

contact and agreeable to Cancer-
cer-Billing and follow-up with
888-555-4343 or by pager #2468.

Additional Social Work Contact
available by calling the Guest
overnight priority social work s
5522. During weekend days, t

ble social work assistance is
00-555-4321. Evening and
rgency Department via pager
ached by paging 123456.

Hand annotation available (swipe or left-click)

Add ACCOUNT_NUMBER (B)
Add ADDRESS (R)
Add AGE (A)
Add CERT_LICENSE_NUMBER (T)
Add CLINICIAN (C)
Add DATE (D)
Add DEVICE_ID_SERIAL_NUM (J)
Add EMAIL (E)
Add FAX (F)
Add HEALTH_PLAN_NUMBER (K)
Add HOSPITAL (H)
Add IP_ADDRESS (I)
Add MED_REC_NUM (M)
Add NAME (N)
Add PHONE (P)
Add PLACE (L)
Add SSN (S)
Add URL (U)
Add VEHICLE_ID (V)
Repeat PHONE (=)
Cancel (<ESC>)

Content tags

xxxxx	ACCOUNT_NUMBEI
xxxxx	ADDRESS
xxxxx	AGE
xxxxx	CERT_LICENSE_NU
xxxxx	CLINICIAN
xxxxx	DATE
xxxxx	DEVICE_ID_SERIAL
xxxxx	EMAIL
xxxxx	FAX
xxxxx	HEALTH_PLAN_NUI
xxxxx	HOSPITAL
xxxxx	IP_ADDRESS
xxxxx	MED_REC_NUM
xxxxx	NAME
xxxxx	PHONE
xxxxx	PLACE
xxxxx	SSN
xxxxx	URL
xxxxx	VEHICLE_ID

Structure tags

Fig. 1 – MIST web-Based graphical user interface showing a social work note that was tagged by the system using an intermediate model. One of the names was not properly tagged so is being hand-corrected in this screen shot.

admitted to the hospital, so they will usually provide very detailed information about the patient's social history, family history, and past medical history. SW notes also generally have a large quantity of PHI because they contain detailed descriptions about the patient's social contexts, including names of friends and family involved in the patient's life.

We retrieved H&P and SW notes from our clinical data repository belonging to decedents who had a history of cancer. We double-confirmed the vital status of the patients from data contained in CareWeb and in our local Tumor Registry, a separate database maintained by our tumor registrars. Because this study only involved decedents, it was not subject to the institutional review board review. Nonetheless, the research protocol was reviewed and approved by the UMHS Privacy and Compliance Offices, and all data were maintained in a secure and encrypted manner in accordance with HIPAA and other privacy regulations. The final corpus contained 360 documents, of which 130 were H&P notes and 230 were SW notes. The H&P notes collectively had 147,685 words whereas the SW notes had a total of 97,126 words, for a total word count of 244,811.

2.4. *Development of statistical models*

We developed our local de-identification model using the *tag-a-little, learn-a-little* (TALLAL) approach (Fig. 2). This method involves an initial naïve state in which the system possesses no a priori knowledge about the nature of PHI elements that need to be handled in a given empirical environment. Nevertheless, the system, through its learning mechanisms, can quickly improve its performance by developing and evolving its statistical models with training examples hand-annotated by human experts. We report on our experiments to measure how many cycles of TALLAL are needed for the system to reach satisfactory performance.

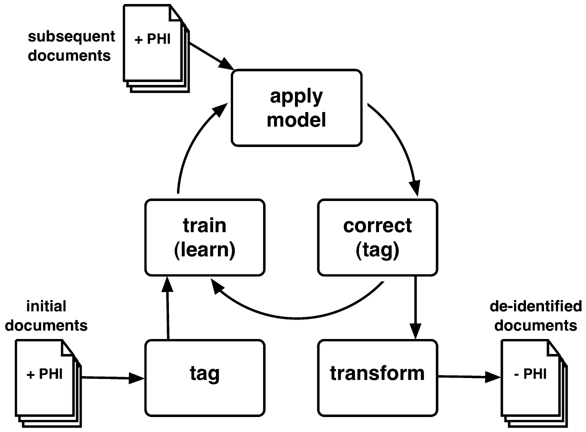


Fig. 2 – Tag-a-little, learn-a-little (TALLAL) training loop.

A typical TALLAL cycle performed in this study proceeds as follows. First, using the MIST interface, an annotator processes a set of documents to highlight the location of every PHI element appearing (*tag-a-little*), and assigns the appropriate label to each PHI element (e.g., name, phone number, date). The results are then used by MIST to develop an initial de-identification model (*learn-a-little*). Next, the provisional model developed in the previous step is applied to a new set of training documents to automatically annotate the documents. The human reviews the annotations made by the system, and manually corrects any errors, using the MIST interface. This approach bootstraps the process of hand-annotating the documents and relegates human annotators to a primarily reviewing role for making corrections only when the system errs. This TALLAL process is performed iteratively until sufficient accuracy is achieved wherein human corrections are no longer, or rarely, needed.

All annotations were performed by a single experienced clinician who was familiar with the content and structure of the medical documents at UMHS. This individual had over 10 years of clinical experience, as well as experience with annotating documents. At the end of each TALLAL iteration, three measures were computed to assess the accuracy of the intermediate model, namely precision, recall, and *F*-score (**precision** is the fraction of annotated instances that are correctly annotated; **recall** is the fraction of relevant instances in the document that are annotated, and F-measure is the harmonic mean of precision and recall). The annotator also tracked clock time taken to annotate documents or correct MIST suggestions as an estimate of human effort needed to achieve high de-identification accuracy levels. The computational time to build the models based on each successive round of annotations was not recorded, as this did not require human effort.

Once all of the documents had been hand annotated, we conducted a series of experiments in which we varied the composition of the document types in both the training and test sets, using the manually added annotations to build the model for the training data and to measure performance in the test data. This was done to understand how the performance of MIST reflects variation in size and composition of the training data. These variations included training and testing on only one document type, changing the number of documents in the training set, as well as using different ratios of note types for model building.

2.5. *Characterization of PHI distribution*

In our experiments we used only H&P and SW notes, but many others note types are available, and there is currently no consensus as to the most valuable notes, in terms of PHI content, from which a de-identification model could be developed. We therefore sought to characterize the PHI content across a variety of note types used in our health system. We extended our final model used in the prior experiments by including an additional 240 documents selected at random from a large pool of decedent patients and constructed a larger model built from a total of 600 trained documents. We did not record the time/effort for this additional tagging. These additional documents included 25 out of a possible 122 distinct document types. The 3 most common ones were (1) Letter/Note-Revisit, (2) Phone Note, and (3) Discharge Summary. Using this larger, more representative model we processed 123,747 documents and characterized the distribution of the tagged PHI elements across all document types. In this paper, we report on the 30 most frequently occurring note types that covered 115,782 documents (93.6% of all documents) – see Fig. 4.

3. Results

3.1. *Performance gains contrasting human effort required*

The initial tag-a-little, learn-a-little results are shown in Table 1. With just a single 10-note round of annotation the trained model achieved an *F*-score of 0.89. Precision, recall, and *F*-scores varied in the early rounds of annotation and training, as novel constructions had great influence over the resulting models. As training rounds progressed, the effect of novel constructions lessened and, by the later rounds, the precision, recall, and *F*-scores stabilized in the 0.95–0.96 range. The final model developed using 80% of the data for training (288 documents) and the remaining 20% for testing (72 documents) achieved an *F*-score of 0.96. The distribution of PHI elements in this model is shown in Table 2.

Table 1 also shows the annotation time for each round of tagging. After just over 8 h of total annotation (round 21) an *F*-score above 0.95 was reached. Annotation time for the entire experiment was just under 12 h. While there was much variability in time required from round to round, the general trend was that the average amount of time required to correctly label the PHI in each document decreased as the rounds progressed, as fewer corrections had to be made as the accuracy of the model improved. The average time required per document for the first five rounds was 143.2 s whereas the average time required per document for the last five rounds was 96.4 s, an approximate 33% decrease in time.

The user interaction logs of this experiment reveal clear patterns in the use of add_annotation and remove_annotation, as shown in Fig. 3. For the first round, no model has been built, and all annotations present in the notes were supplied by the human annotator, hence the large number of add_annotation actions. For Round 2 an initial model has been applied to the data before the human annotator begins, and there are far fewer instances of add_annotation, as the model has borne some of the workload. There is variability in the use of add_annotation in subsequent rounds (just as there is variability in the instances of PHI in different notes), but the overall trend is downward, as the trained models become more accurate.

In the first round remove_annotation is used a small number of times, as the annotator corrected himself in applying PHI tags. Use of remove_annotation increases in subsequent rounds, and the early models create some incorrect instances of PHI markup that the human annotator must remove. After round 18 the use of remove_annotation lessens as the models improve. In the final rounds the annotator is back to using remove_annotation largely to correct himself rather than the model in applying PHI tags.

Table 1 – Total time required and gains in perforomance after each round of *tag-a-little, learn-a-little*. Intermediate de-identification models were developed using 80% of the cumulative hand-corrected notes; performance was evaluated using the remaining 20% of the cumulative hand-corrected notes. The rightmost column lists how many PHI elements were in the documents for each round.

Round	Cumulative notes	Annotation time		Precision	Recall	*F*-score	Average PHI elements per document
		Per note (min)	Cumulative (h)				
1	10	3:18	0:33	0.902	0.873	0.888	24.8
2	20	1:36	0:49	0.926	0.846	0.884	20.7
3	30	2:00	1:09	0.986	0.783	0.873	21.6
4	40	2:36	1:35	0.947	0.774	0.852	28.1
5	50	2:26	1:59	0.929	0.829	0.876	26.5
6	60	1:45	2:17	0.936	0.826	0.878	19.1
7	70	2:15	2:39	0.936	0.836	0.883	24.2
8	80	1:57	2:59	0.926	0.932	0.929	26.2
9	90	2:27	3:23	0.938	0.895	0.916	33.2
10	100	2:01	3:44	0.966	0.920	0.942	27.9
11	110	2:24	4:08	0.947	0.919	0.932	33.9
12	120	2:26	4:32	0.967	0.916	0.941	29.5
13	130	2:33	4:57	0.953	0.881	0.916	30.3
14	140	2:06	5:18	0.964	0.901	0.931	23.2
15	150	2:42	5:45	0.972	0.925	0.948	31.7
16	160	1:39	6:02	0.971	0.916	0.943	21.0
17	170	2:15	6:24	0.977	0.929	0.952	25.5
18	180	2:07	6:46	0.959	0.915	0.937	25.0
19	190	2:12	7:08	0.955	0.920	0.937	25.6
20	200	2:32	7:33	0.960	0.917	0.938	30.0
21	220	1:50	8:10	0.969	0.938	0.953	25.6
22	240	1:32	8:40	0.964	0.941	0.952	22.3
23	260	1:37	9:13	0.967	0.951	0.959	24.7
24	280	1:44	9:47	0.969	0.946	0.957	21.8
25	300	1:21	10:14	0.961	0.949	0.955	22.1
26	320	1:36	10:46	0.967	0.957	0.962	23.7
27	360	1:44	11:56	0.973	0.955	0.964	22.5

3.2. *Experiments with a mix of note types*

The results of the six experiments (denoted A-F) with the mix of note types are presented in Table 3. The first experiments (A and B) compared 100 History and Physical (H&P) and Social Work (SW) notes trained and tested using equal numbers of documents. The overall model accuracy was higher for 100 H&P documents (*F*-score: 0.957) than it was for 100 SW documents (*F*-score: 0.915), although accuracy of AGE, NAME, and PHONE was higher for 100 SW than for 100 H&P. Adding 100 additional SW documents (C) improved *F*-scores slightly overall and for most PHI types, although the scores for DATE and HEALTH PLAN NUM remained unchanged. Overall *F*-score for 200 SW (C) was still lower than the overall *F*-score for 100 H&P (A). We next attempted to boost the accuracy of de-identification of SW notes by adding the 100 H&P notes to the training set (D), creating a corpus of 300 documents. This further raised the overall *F*-score to 0.922, with small boosts to

Table 2 – Total PHI elements and average number of PHI elements per note obtained in training and test notes for the final model developed, shown in round 27 of Table 1.

PHI type	Training set (N = 288)		Test set (N = 72)	
	Total PHI elements	Average per note	Total PHI elements	Average per note
Address	5	0.02	2	0.03
Age	490	1.70	132	1.83
Clinician	1003	3.48	265	3.68
Date	3125	10.85	753	10.46
Email	1	<0.01	0	0
Health Plan Number	5	0.02	4	0.06
Hospital	315	1.09	66	0.92
Medical Record Number	5	0.02	0	0
Name	1235	4.29	284	3.94
Phone	714	2.48	166	2.31
Place	327	1.14	76	1.06
URL	5	0.02	0	0
Total from above	7230	25.10	1748	24.28

Table 3 – Results of the six experiments in which varying numbers of H&P and SW notes were used to build models that were then tested against either 30 H&P or 30 SW notes. The *F*-score for each PHI type is shown for each experiment.

Experiment	A	B	C	D	E	F
Training Corpus	100 H&P	100 SW	200 SW	100 H&P, 200 SW	100 H&P, 100 SW	100 H&P, 200 SW
Test Corpus	30 H&P	30 SW	30 SW	30 SW	30 H&P	30 H&P
PHI type						
Address	†,*	0.000	0.000	0.000	*	*
Age	0.914	0.940	0.965	0.988	0.916	0.937
Clinician	0.951	0.951	0.957	0.963	0.943	0.948
Date	0.980	0.991	0.992	0.996	0.978	0.978
Email	†,*	†	†	†	†,*	†,*
Health Plan Number	†,*	0.400	0.400	0.400	*	*
Hospital	0.824	0.692	0.741	0.741	0.885	0.885
Medical Record Number	†,*	*	*	*	*	*
Name	0.759	0.845	0.885	0.897	0.841	0.862
Phone	0.845	0.965	0.975	0.979	0.935	0.935
Place	0.845	0.860	0.880	0.901	0.873	0.838
URL	†,*	†	†	†	†,*	†,*
All PHI types (average)	0.957	0.905	0.915	0.922	0.960	0.961

H&P = History & physical note; SW = Social work note; Health Plan Number = Health plan number (i.e., health insurance plan number).

† No training items available for the PHI element in this experiment.

* No testing items available for the PHI element in this experiment.

the *F*-scores of most of the PHI types. Still, *F*-scores for the SW test corpus never reached the level of *F*-scores for the H&P corpus alone. Finally, we evaluated balanced (E) and unbalanced (F) models against the H&P test corpus, with slight increases in scores with each increment in number of training notes, regardless of note type.

Some of the variability of the *F*-score for each PHI element across experiments can likely be explained by the variability of PHI in the test and training corpora, as shown in Table 4. For example, there are no instances of EMAIL or URL in any of the training corpora, and thus no hope of training a model to find such PHI. Similarly, there are very few instances of ADDRESS, Health Plan Number, and Medical Record Number in the training corpora. The H&P test corpus contains no instances of ADDRESS, EMAIL, Health Plan Number, Medical Record Number, or URL.

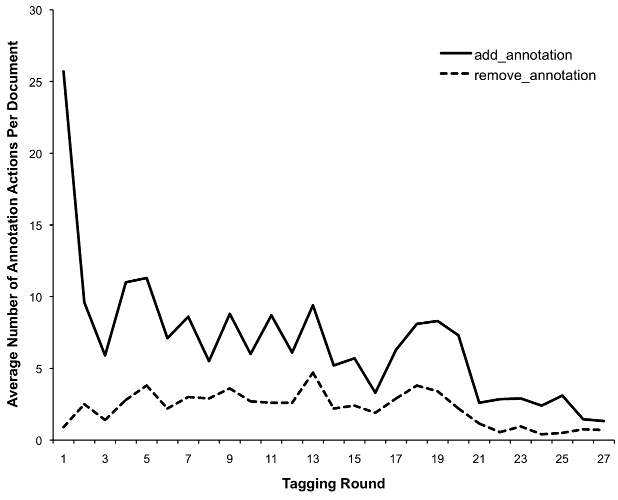


Fig. 3 – Use of add_annotation and remove_annotation. The add_annotation action was required when the system missed a PHI element, whereas the remove_annotation action was required when the system inaccurately labeled a non-PHI element as PHI. As the TALLAL cycles progressed, the need to manually add or remove annotations decreased as the models became more accurate.

3.3. *PHI distribution in different types of clinical documents*

Fig. 4 exhibits a heat map delineating a summary of the distribution of 12 types of PHI across the 30 most frequently occurring document types, drawn from a pool of about 116,000 documents. Document types (rows) are sorted according to overall density of PHI, from top to bottom; those at the top would likely be the best choices for training a de-identification model. Based on this heat map, the top three document types for model training would be “Discharge Summary,” “Outpatient Consult,” and “Admission History & Physical.” “Social Work Notes,” which we had selected for our initial experiments, ranked tenth on the list. URL was located primarily in “Physical Therapy and Occupational Therapy” notes, which was a result of common templates used that included URLs of reference materials. The rarity of EMAIL was noted across nearly all document types except for one called an “Email note”. However, this note type was used infrequently and is therefore not displayed in the heat map.

4. Discussion

The results of our initial *tag-a-little*, *learn-a-little* (TALLAL) bootstrapping experiment show that it is feasible from a human labor perspective to generate de novo a de-identification system with reasonable performance. Bootstrapping has been used in similar tasks involving named entity tagging [19] but to our knowledge the technique has not previously been applied to a de-identification system. In slightly under 3 h of annotation, and using only 80 documents, we achieved an overall *F*-score above 0.90 with the system beginning to plateau around 0.96 after 12 h of annotation and 360

Table 4 – Average number of PHI elements per record contained in training and test corpora for the experiments reports in Table 3.

PHI type	Training corpora					Test corpora	
	100 H&P	100 SW	200 SW	200 Hybrid	300 Hybrid	30 H&P	30 SW
Address	0	0.03	0.03	0.02	0.02	0	0.07
Age	2.47	1.17	1.26	1.82	1.66	2.67	1.43
Clinician	3.88	3.26	3.29	3.57	3.49	4.40	3.00
Date	15.96	7.47	7.31	11.72	10.19	20.47	6.90
Email	0	0	0	0	0	0	0.03
Health Plan Number	0	0.01	0.03	0.01	0.02	0	0.13
Hospital	1.43	0.81	0.93	1.12	1.09	1.03	0.73
Medical Record Number	0	0.01	0.01	0.01	<0.01	0	0
Name	2.57	5.35	5.29	3.96	4.38	2.60	4.23
Phone	0.37	3.50	3.61	1.94	2.53	0.47	3.60
Place	0.75	1.17	1.27	0.96	1.10	0.80	1.67
URL	0	0	0	0	0	0	0.17

H&P = History & physical note; SW = Social work note; Health Plan Number = Health plan number (i.e., health insurance plan number).

documents. This level of performance is comparable to the systems included in the 2006 i2b2 De-identification Challenge [20].

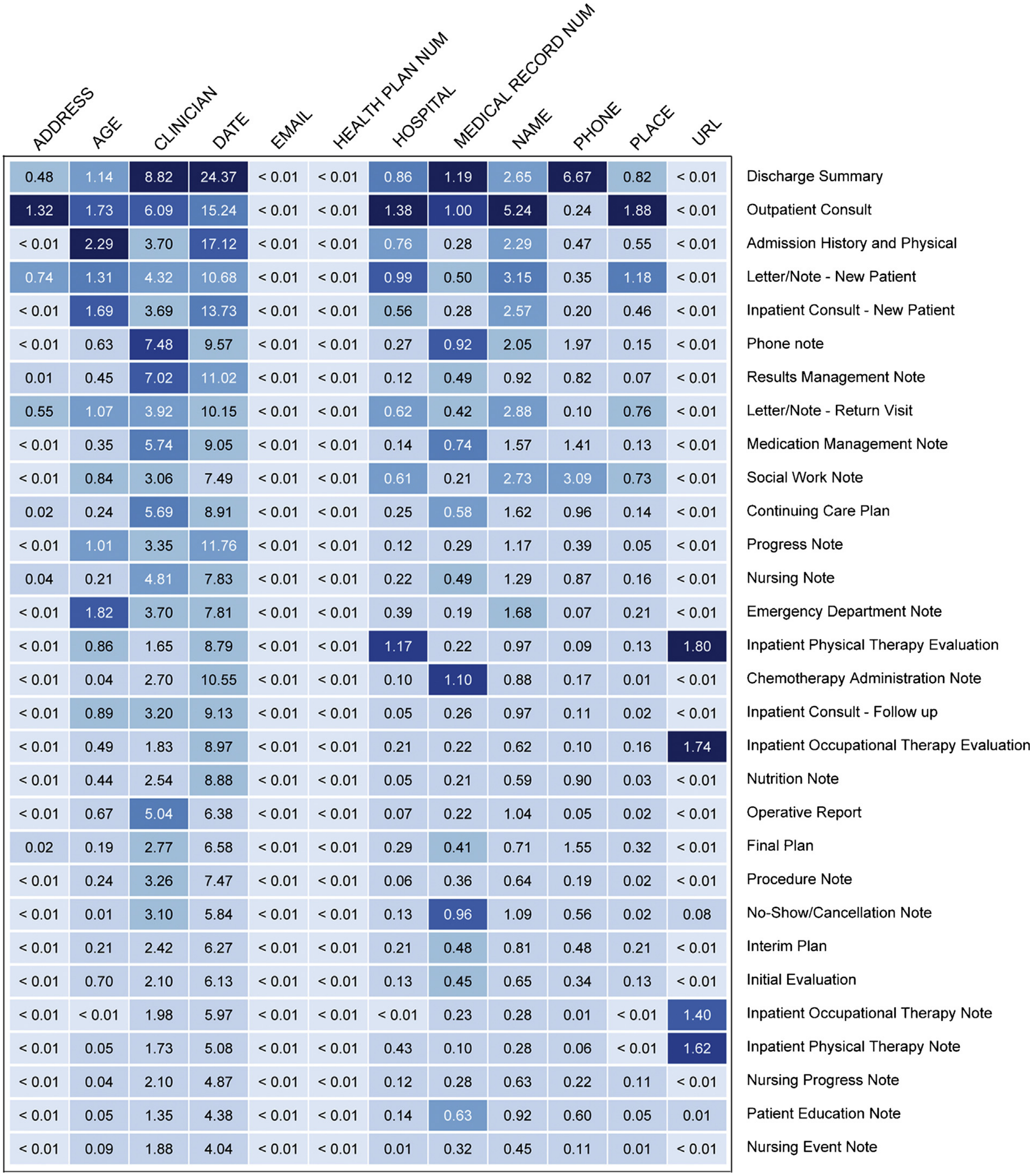
While more training data generally continued to improve performance, the incremental gains diminished as the process progressed. Through the TALLAL approach, it is possible to develop labor vs. performance metrics that could be useful when allocating resources for such endeavors. Additionally, it may be possible to quantify ideal stopping points for the training process once a certain threshold of performance has been reached or when it can be shown that the system is near the maximum performance achievable [21].

As is evident in the experiments with different combinations of History and Physical (H&P) and Social Work (SW) notes, the performance of the system was highly dependent on the richness of document types selected for training data, and the match between the training data and test data [15,16]. When building a model it is ideal to select documents with a high density and distribution of PHI, as this will likely lead to a more robust model. In selecting the note types for our experiment, we did choose documents with a reasonable PHI content, although other document types may have been comparable if not better. In considering the selection of document types, it may also be worthwhile to include those that are significantly syntactically and semantically different from one another to ensure a broad range of training examples for the system. Recent work on clustering document types based on such measures demonstrated that “Admission History/Physical” notes were quite distinct from “Social Services Note [Inpatient]” and therefore our choice of H&P and SW notes may have been reasonable based on this measure [22].

Rare PHI types such as EMAIL and URL are potentially the most stable in terms of intrinsic structure. Therefore, adding a small number of regular expressions to detect these elements could increase system performance despite the lack of adequate training examples. By contrast, some PHI elements such as NAME and PLACE were comparatively abundant but the models failed to reach high levels of performance for these. This is likely due to the greater variation of the internal structure of these elements, and the external context where these PHI elements appear. Performance could potentially be boosted by use of a dictionary containing census data or the names of health care institutions including hospitals, clinics, and nursing homes. Use of dictionaries is common in de-identification systems, although a hybrid approach using dictionaries and machine learning is not [23,24]. One might even consider including a local dictionary of names continually updated from the electronic health record as an additional data resource. MIST does provide facilities for including external lists (e.g., clinician names and patient names), but we did not include it in our experiments.

The distribution of PHI types in our documents differed from those reported in other studies. For example, the top five most frequent types in our corpus included DATE (43%), NAME (patient or family name, 17%), CLINICIAN (14%), PHONE (including FAX, 10%), and AGE (7%). This compares to a study by Dalianis and Velupillai [25] in which the top five PHI types were HEALTH CARE UNIT (28%), CLINICIAN (27%), DATE (23%), NAME (4%), and LOCATION (3%), and a study by Neamatullah et al. [11] reported the most frequent PHI types to be DATE (33%), CLINICIAN (33%), LOCATION (21%), NAME (10%), and PHONE (2%). In all cases DATEs, NAMEs, and CLINICIANs were among the 5 most frequent types. The remaining differences may simply be a reflection of the document types selected for inclusion. Our study oversampled social work notes for the perceived benefits of the PHI content which likely explains, for example, why PHONE was much more common – social workers tended to leave in the notes phone numbers at which they could be reached. Additionally, comparisons are not perfect due to a lack of consensus between studies of what should be annotated and how elements should be categorized [25].

In our experiment we were able to annotate 360 documents in just under 12 h, a rate of about 30 documents, or 20,500 words per hour. Additionally, our documents contained a relatively high number of PHI elements, and the rate of annotation was calculated at 752 PHI elements per hour. By contrast, a study in which house officers manually labeled PHI in nursing notes reported that the annotators were able to process about 18,000 words, and 90 PHI elements, per hour [11,26]. The vastly different rate of PHI per hour may simply be a reflection of the density of PHI in the documents, since our documents had



ADDRESS	AGE	CLINICIAN	DATE	EMAIL	HEALTH PLAN NUM	HOSPITAL	MEDICAL RECORD NUM	NAME	PHONE	PLACE	URL	
0.48	1.14	8.82	24.37	< 0.01	< 0.01	0.86	1.19	2.65	6.67	0.82	< 0.01	Discharge Summary
1.32	1.73	6.09	15.24	< 0.01	< 0.01	1.38	1.00	5.24	0.24	1.88	< 0.01	Outpatient Consult
< 0.01	2.29	3.70	17.12	< 0.01	< 0.01	0.76	0.28	2.29	0.47	0.55	< 0.01	Admission History and Physical
0.74	1.31	4.32	10.68	< 0.01	< 0.01	0.99	0.50	3.15	0.35	1.18	< 0.01	Letter/Note - New Patient
< 0.01	1.69	3.69	13.73	< 0.01	< 0.01	0.56	0.28	2.57	0.20	0.46	< 0.01	Inpatient Consult - New Patient
< 0.01	0.63	7.48	9.57	< 0.01	< 0.01	0.27	0.92	2.05	1.97	0.15	< 0.01	Phone note
0.01	0.45	7.02	11.02	< 0.01	< 0.01	0.12	0.49	0.92	0.82	0.07	< 0.01	Results Management Note
0.55	1.07	3.92	10.15	< 0.01	< 0.01	0.62	0.42	2.88	0.10	0.76	< 0.01	Letter/Note - Return Visit
< 0.01	0.35	5.74	9.05	< 0.01	< 0.01	0.14	0.74	1.57	1.41	0.13	< 0.01	Medication Management Note
< 0.01	0.84	3.06	7.49	< 0.01	< 0.01	0.61	0.21	2.73	3.09	0.73	< 0.01	Social Work Note
0.02	0.24	5.69	8.91	< 0.01	< 0.01	0.25	0.58	1.62	0.96	0.14	< 0.01	Continuing Care Plan
< 0.01	1.01	3.35	11.76	< 0.01	< 0.01	0.12	0.29	1.17	0.39	0.05	< 0.01	Progress Note
0.04	0.21	4.81	7.83	< 0.01	< 0.01	0.22	0.49	1.29	0.87	0.16	< 0.01	Nursing Note
< 0.01	1.82	3.70	7.81	< 0.01	< 0.01	0.39	0.19	1.68	0.07	0.21	< 0.01	Emergency Department Note
< 0.01	0.86	1.65	8.79	< 0.01	< 0.01	1.17	0.22	0.97	0.09	0.13	1.80	Inpatient Physical Therapy Evaluation
< 0.01	0.04	2.70	10.55	< 0.01	< 0.01	0.10	1.10	0.88	0.17	0.01	< 0.01	Chemotherapy Administration Note
< 0.01	0.89	3.20	9.13	< 0.01	< 0.01	0.05	0.26	0.97	0.11	0.02	< 0.01	Inpatient Consult - Follow up
< 0.01	0.49	1.83	8.97	< 0.01	< 0.01	0.21	0.22	0.62	0.10	0.16	1.74	Inpatient Occupational Therapy Evaluation
< 0.01	0.44	2.54	8.88	< 0.01	< 0.01	0.05	0.21	0.59	0.90	0.03	< 0.01	Nutrition Note
< 0.01	0.67	5.04	6.38	< 0.01	< 0.01	0.07	0.22	1.04	0.05	0.02	< 0.01	Operative Report
0.02	0.19	2.77	6.58	< 0.01	< 0.01	0.29	0.41	0.71	1.55	0.32	< 0.01	Final Plan
< 0.01	0.24	3.26	7.47	< 0.01	< 0.01	0.06	0.36	0.64	0.19	0.02	< 0.01	Procedure Note
< 0.01	0.01	3.10	5.84	< 0.01	< 0.01	0.13	0.96	1.09	0.56	0.02	0.08	No-Show/Cancellation Note
< 0.01	0.21	2.42	6.27	< 0.01	< 0.01	0.21	0.48	0.81	0.48	0.21	< 0.01	Interim Plan
< 0.01	0.70	2.10	6.13	< 0.01	< 0.01	0.13	0.45	0.65	0.34	0.13	< 0.01	Initial Evaluation
< 0.01	< 0.01	1.98	5.97	< 0.01	< 0.01	< 0.01	0.23	0.28	0.01	< 0.01	1.40	Inpatient Occupational Therapy Note
< 0.01	0.05	1.73	5.08	< 0.01	< 0.01	0.43	0.10	0.28	0.06	< 0.01	1.62	Inpatient Physical Therapy Note
< 0.01	0.04	2.10	4.87	< 0.01	< 0.01	0.12	0.28	0.63	0.22	0.11	< 0.01	Nursing Progress Note
< 0.01	0.05	1.35	4.38	< 0.01	< 0.01	0.14	0.63	0.92	0.60	0.05	0.01	Patient Education Note
< 0.01	0.09	1.88	4.04	< 0.01	< 0.01	0.01	0.32	0.45	0.11	0.01	< 0.01	Nursing Event Note

Fig. 4 – Heat map visualization of density of occurrence across 12 PHI elements and the 30 most frequently occurring document types. Darker colors represent higher densities. Color density was calculated uniquely for each PHI element (columns) to highlight which document types (rows) have the highest concentrations of each PHI element. Precision, recall and F-measure are not reported as we have no gold standard for this larger data set.

seven times the density of PHI. Another study that assessed the time costs for manual de-identification reported an average rate of 41 documents, 13,100 words and 326 PHI elements per hour [27]. In this case the density of PHI was closer to our corpus, for which we had only 1.5 times the density of PHI. In comparing various studies it is difficult to draw conclusions about efficiencies, but the results do suggest that the bootstrapping method, and/or the intuitive user interface of MIST, may have provided time efficiency benefits. Even with the benefits of the MIST system providing most of the annotations automatically, time was still required to carefully read each note to look for mistakes (which as the experiment progressed became increasingly rare). As such, a minimum floor for time required likely exists for each annotator based on their speed of reading through clinical documents and the care taken to ensure that errors are avoided.

Our study does have several limitations. The models we developed and tested were built with documents from a single institution, from decedents with cancer, and from just two document types. The results of the PHI heat map we generated depend on the applicability of our model to other document types that were not included in the training set and it is possible that certain PHI elements may have been mislabeled as a result. We know from previous work that models trained on document types with high PHI density are more portable to other document types than models trained on documents with low PHI density [15]. For example, applying a model trained on discharge summaries (high PHI density) to order summaries (low PHI density) yields higher accuracy than the reverse (0.924F vs. 0.662F) [15]. We also know that hybrid models trained from several document types perform well on a variety of test document types [15].

An additional limitation is that the document types used in this study were locally developed at UMHS and may not directly overlap with those from other institutions, although many would likely be similar even if the precise names are different from what has been reported elsewhere [22,28]. Adoption of standardized document types could aid in addressing this issue [29]. Furthermore, in this experiment we did not compare the TALLAL approach directly to pure manual annotation in terms of the final accuracy of the model or the time required for annotation. Future work should involve such a comparison.

The tagging in this study was performed by a single annotator, albeit one with significant clinical experience. Adding additional annotators can lead to greater accuracy in identifying and redacting PHI, but at significant cost, and with potential trade-offs [30]. Additional verification of annotations at the end of each round could improve the accuracy of the de-identification model, as could checking for inter-annotator-agreement (also known as inter-rater agreement) if more than one annotator was involved. Another study involving the manual review of clinical documents reported an intraclass-correlation coefficient of 0.99 when including all words in the corpus, suggesting that most of the time the annotations were correct [27]. Other studies, however, have reported lower levels of agreement [31]. Future work should study the cost/accuracy trade-offs required for practical applications. For example, de-identification of documents for use within an institution may be performed at lower cost than de-identification of documents to be released externally, where ensuring completeness of PHI removal is paramount.

The logging mechanism of MIST does not fully capture the context in which actions occurred. For example, the remove_annotation actions recorded during each round could have been due to the need to remove items incorrectly tagged by the system or to remove items incorrectly tagged by the annotator during that round. Thus, it is possible that our measure of remove_annotation over-estimated the number of times the system incorrectly labeled text. However, our experience in using the system was that the vast majority of the times remove annotation was used, it was to remove annotations incorrectly made by the system.

Finally, all de-identification processes, whether automated or manual, leave behind a certain number of residual identifiers [11,27,32]. Replacing PHI elements with synthetic surrogates can mitigate the effect of residual identifiers [33]. Manual review for residual identifiers might also be warranted in some cases, especially when the risks are high and the total number of documents is reasonable for a human to review. In practice, a successful de-identification strategy for secondary use should include both technology (e.g., an automated de-identification system) and policy (e.g., a data use agreement governing usage of the data and prohibiting attempts to re-identify the data).

Summary points

What was already known on the topic:

- Removal of protected health information constitutes a major stumbling block for sharing information from medical records for research purposes.
- High-performing machine learning approaches to de-identification can be applied to a variety of medical record types, but require annotated training corpora.

What this study added to our knowledge:

- Using an iterative tag-a-little, learn-a-little approach, a de-identification system can be built for a particular document type with less than one day's worth of annotation effort.
- The resulting de-identification system achieves accuracy levels that are comparable with published state-of-the-art systems.

5. Conclusion

We have shown that it is possible, with about 8 h of annotation time, to build a functioning de-identification system de novo using the MIST framework that achieves an *F*-score of 0.95. The performance achieved was comparable to other systems and could likely be further improved with additional approaches and efforts, such as the use of dictionaries. Nevertheless, a system that produces a set of documents with the vast majority of identifiers removed could be highly useful and beneficial to patient privacy for internal research and quality improvement initiatives.

Author contributions

DH, JA, SB, BW, CC and LH were responsible for the design and analysis of the experiments. JA, SB, BW, CC and LH were responsible for the development of MIST. DH performed the annotations, built the models, ran the experiments, and provided feedback to improve MIST. All authors contributed to the writing of the paper.

Conflicts of interest

None.

Acknowledgements

This work was supported in part by Grant UL1RR024986 received from the National Center for Research Resources – a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research. The MITRE Corporation is a not-for-profit organization chartered to work in the public interest and has made its de-identification tool, MIST, freely available as an open source application.[1]

REFERENCES

[1] D. Blumenthal, Stimulating the adoption of health information technology, N. Engl. J. Med. 360 (April (15)) (2009) 1477–1479.

[2] Blumenthal D. EHR Adoption Set to Soar. 2010 [cited 17 February 2010]; Available from: http://healthit.hhs.gov/portal/server.pt?open=512&mode=2&objID=3357

[3] D.W. Bates, A. Bitton, The future of health information technology in the patient-centered medical home, Health Aff. (Millwood) 29 (April (4)) (2010) 614–621.

[4] C.P. Friedman, A.K. Wong, D. Blumenthal, Achieving a nationwide learning health system, Sci. Transl. Med. 2 (November (57)) (2010) 57cm29.

[5] B.H. Gray, T. Bowden, I. Johansen, S. Koch, Issues in international health policy: electronic health records: an international perspective on "meaningful use", Issue Brief (Commonwealth Fund) (2011) 1–18.

[6] A.K. Jha, D. Doolan, D. Grandt, T. Scott, D.W. Bates, The use of health information technology in seven nations, Int. J. Med. Inform. 77 (12) (2008) 848–854.

[7] S.T. Rosenbloom, J.C. Denny, H. Xu, N. Lorenzi, W.W. Stead, K.B. Johnson, Data from clinical notes: a perspective on the tension between structure and flexible documentation, J. Am.Med.Inform.Assoc. 18 (March–April (2)) (2011) 181–186.

[8] K. Zheng, Q. Mei, Hanauer DA, Collaborative search in electronic health records, J. Am.Med.Inform.Assoc. 18 (May (3)) (2011) 282–291.

[9] Standards for privacy of individually identifiable health information. 65th ed. Federal Register: US Department of Health and Human Services; 2000. p. 82462-510.

[10] K. Tu, J. Klein-Geltink, T. Mitiku, C. Mihai, J. Martin, De-identification of primary care electronic medical records free-text data in Ontario, Canada, BMC Med. Inform. Decis. Mak. 10 (1) (2010) 35.

[11] I. Neamatullah, M.M. Douglass, L.W. Lehman, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, G.D. Clifford, Automated de-identification of free-text medical records, BMC Med. Inform. Decis. Mak. 8 (2008) 32.

[12] F.J. Friedlin, C.J. McDonald, A software tool for removing patient identifying information from clinical documents, J. Am. Med. Inform. Assoc. 15 (September–October (5)) (2008) 601–610.

[13] G. Szarvas, R. Farkas, R. Busa-Fekete, State-of-the-art anonymization of medical records using an iterative machine learning framework, J. Am. Med. Inform. Assoc. 14 (September–October (5)) (2007) 574–580.

[14] Ö. Uzuner, T.C. Sibanda, Y. Luo, P. Szolovits, A de-identifier for medical discharge summaries, Artif. Intell. Med. 42 (January (1)) (2008) 13–35.

[15] J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, L. Hirschman, The MITRE Identification Scrubber Toolkit: design, training, and assessment, Int. J. Med. Inform. 79 (December (12)) (2010) 849–859.

[16] R. Yeniterzi, J. Aberdeen, S. Bayer, B. Wellner, L. Hirschman, B. Malin, Effects of personal identifier resynthesis on clinical text de-identification, J. Am. Med. Inform. Assoc. 17 (March–April (2)) (2010) 159–168.

[17] W. Bria, Web-enabled clinical information systems: a new era begins, in: M. Brady, M. Hassett (Eds.), Clinical Informatics, Healthcare Information and Management Systems Society, Chicago, IL, 2000, pp. 103–111.

[18] K. Zheng, Q. Mei, L. Yang, F.J. Manion, U.J. Balis, D.A. Hanauer, Voice-dictated versus typed-in clinician notes: linguistic properties and the potential implications on natural language processing, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2011, pp. 1630–1638.

[19] R. Ghani, R. Jones, A comparison of efficacy and assumptions of bootstrapping algorithms for training information extraction systems, in: Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data at the Linguistic Resources and Evaluation Conference (LREC 2002), 2002.

[20] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (September–October (5)) (2007) 550–563.

[21] F. Laws, H. Schätze, Stopping criteria for active learning of named entity recognition, in: Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 465–472.

[22] O. Patterson, J.F. Hurdle, Document clustering of clinical narratives: a systematic study of clinical sublanguages, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2011, pp. 1099–1107.

[23] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 10 (2010) 70.

[24] B. Wellner, M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, A. Yeh, J. Hitzeman, L. Hirschman, Rapidly retargetable approaches to de-identification in medical records, J. Am. Med. Inform. Assoc. 14 (January (5)) (2007) 564–573.

[25] H. Dalianis, S. Velupillai, De-identifying Swedish clinical text – refinement of a gold standard and experiments with Conditional random fields, J. Biomed. Semantics 1 (April (1)) (2010) 6, doi:10.1186/2041-1480-1-6.

[26] M.M. Douglass, G.D. Clifford, A. Reisner, G.B. Moody, R.G. Mark, Computer-assisted deidentification of free text in the MIMIC II database, Comput. Cardiol. 31 (2004) 341–344.

[27] D.A. Dorr, W.F. Phillips, S. Phansalkar, S.A. Sims, J.F. Hurdle, Assessing the difficulty and time cost of de-identification in clinical narratives, Methods Inf. Med. 45 (2006) 246–252.

[28] C. Arnott Smith, Effect of XML markup on retrieval of clinical documents, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2003, pp. 614–618.

[29] R.H. Dolin, L. Alschuler, C. Beebe, P.V. Biron, S.L. Boyer, D. Essin, E. Kimber, T. Lincoln, J.E. Mattison, The HL7 clinical document architecture, Journal of the American Medical Informatics Association: JAMIA. 8 (November–December (6)) (2001) 552–569.

[30] S. Shen, B.R. South, F.J. Friedlin, S. Meystre, Coverage of manual de-identification on VA clinical documents, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2011.

[31] S. Velupillai, H. Dalianis, M. Hassel, G.H. Nilsson, Developing a standard for de-identifying electronic patient records

[1] http://mist-deid.sourceforge.net/.

written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial, Int. J. Med. Inform. 78 (December (12)) (2009) e19–e26.

[32] J. Mayer, S. Shen, B.R. South, S. Meystre, F.J. Friedlin, W.R. Ray, M. Samore, Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2009.

[33] D. Carrell, B. Malin, J. Aberdeen, S. Bayer, C. Clark, B. Wellner, L. Hirschman, Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text, J. Am. Med. Inform. Assoc. 20 (March–April (2)) (2013) 342–348.