Effects of personal identifier resynthesis on clinical text de-identification

Reyyan Yeniterzi,[1] John Aberdeen,[2] Samuel Bayer,[2] Ben Wellner,[2,3] Lynette Hirschman,[2] Bradley Malin[4]

▶ Supplementary appendices are published online only at http://jamia.bmj.com/content/vol17/issue2.

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey
[2]The MITRE Corporation, Bedford, Massachusetts, USA
[3]Department of Computer Science, Brandeis University, Waltham, Massachusetts, USA
[4]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee, USA

Correspondence to
Dr Bradley Malin, 2525 West End Avenue, Suite 600, Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN 37203, USA; b.malin@vanderbilt.edu

Received 31 May 2009
Accepted 14 December 2009

ABSTRACT

Objective De-identified medical records are critical to biomedical research. Text de-identification software exists, including "resynthesis" components that replace real identifiers with synthetic identifiers. The goal of this research is to evaluate the effectiveness and examine possible bias introduced by resynthesis on de-identification software.

Design We evaluated the open-source MITRE Identification Scrubber Toolkit, which includes a resynthesis capability, with clinical text from Vanderbilt University Medical Center patient records. We investigated four record classes from over 500 patients' files, including laboratory reports, medication orders, discharge summaries and clinical notes. We trained and tested the de-identification tool on real and resynthesized records.

Measurements We measured performance in terms of precision, recall, F-measure and accuracy for the detection of protected health identifiers as designated by the HIPAA Safe Harbor Rule.

Results The de-identification tool was trained and tested on a collection of real and resynthesized Vanderbilt records. Results for training and testing on the real records were 0.990 accuracy and 0.960 F-measure. The results improved when trained and tested on resynthesized records with 0.998 accuracy and 0.980 F-measure but deteriorated moderately when trained on real records and tested on resynthesized records with 0.989 accuracy 0.862 F-measure. Moreover, the results declined significantly when trained on resynthesized records and tested on real records with 0.942 accuracy and 0.728 F-measure.

Conclusion The de-identification tool achieves high accuracy when training and test sets are homogeneous (ie, both real or resynthesized records). The resynthesis component regularizes the data to make them less "realistic," resulting in loss of performance particularly when training on resynthesized data and testing on real data.

INTRODUCTION

Medical privacy regulations, such as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA), prohibit the disclosure of identifiable medical records without the explicit consent of the patients whose records are involved.[1] The Privacy Rule; however, permits the dissemination of medical records if the data are de-identified or stripped of patient-identifiable features, termed protected health information (PHI). There are several standards by which the Privacy Rule permits de-identified data sharing. One model in particular, the Safe Harbor standard, allows distribution of records once 18 enumerated PHI identifiers have been suppressed from a patient's record, such as names, dates, and geographic information. Thus, over the past decade, medical informaticists and computer scientists have developed various automated de-identification tools to find and suppress HIPAA-designated identifiers from natural language texts. These tools utilize various dictionaries and rules,[2–7] artificial intelligence,[8] and machine-learning[9–15] techniques. For medical record de-identification tools to be adopted widely, they must be ready for application "off the shelf". Presently, this is difficult to achieve because most de-identification tools need to be trained on, or tailored to, the electronic medical records for a given institution[11] and the particular medical record class to which it will be applied. Moreover, most healthcare institutions have limited personnel and experience with software development, such that training or adapting a system to each new institutions's record format and class presents a formidable obstacle to deployment.[16]

One of the principal challenges that designers of de-identification tools face is a limited availability of real medical records for the development and evaluation of their software. While developers that reside within a healthcare setting often have access to large quantities of medical records for their investigations, the proprietary nature of such information significantly hinders the sharing of rich collections to the research community at large. This is problematic because, with restrictive access to a common set of medical records, it is difficult for researchers at disparate institutions to empirically and objectively compare de-identification systems. Non-healthcare focused institutions are at an even greater disadvantage due to their inability to access medical information in a timely manner, if at all. It is clear that the biomedical community, and de-identification researchers in particular, would benefit greatly if a corpus of electronic medical records could be made publicly, or semi-publicly, available.

Without access to identified electronic medical records, de-identification researchers rely on "resynthesized" versions, such as the corpus of hospital discharge records made available in the i2b2 de-identification evaluation held at the 2006 AMIA Annual Fall Symposium.[17] In such corpora, real patient identifiers are replaced with "fake" identifiers to simulate natural language. Though such a practice increases the availability of data for evaluation purposes, the development of de-identification models through synthetic information raises the question of how accurate the resulting text de-identification systems are when applied to real medical records. Moreover, if the resultant systems are biased, it is necessary to determine the cause and make corrections as necessary.

This paper reports on a systematic investigation into the effect of using resynthesized medical records on the performance of an automated de-identification tool. This investigation partnered researchers from the Vanderbilt University Medical Center (VUMC) with the MITRE Corporation. The VUMC supplied a large set of diverse medical records, and MITRE supplied a state-of-the-art, open-source text de-identification system based on machine learning,[14] known as the MITRE Identification Scrubber Toolkit (MIST). We report on the effect of various training and testing environments on the accuracy of de-identification in real-world scenarios. In general, our findings suggest that training and testing de-identification tools with resynthesized medical records provide a slightly inflated accuracy in comparison to tools that use only real medical records. We further find that training tools with resynthesized data and testing with real data cause a significant decrease in de-identification accuracy. We hypothesize that these results can be explained in terms of two effects: the tendency of the resynthesis component to lose some of the variability in the original records; and the fact that any mismatch between training and test will degrade results. The remainder of this paper addresses these issues in more depth, explores causes of these discrepancies, and postulates how they can be resolved.

BACKGROUND

Resynthesis in text de-identification

As a first step towards impartial comparison of clinical text de-identification tools, Informatics for Integrating Biology and the Bedside (i2b2), a National Center for Biomedical Computing based at Partners Healthcare System, held a challenge evaluation of text de-identification tools in 2006.[17] In preparation for the challenge, a corpus of approximately 650 hospital discharge summaries from the Partners Healthcare System were shared to interested researchers for tuning and training text de-identification systems. Real patient identifiers in the medical records were replaced with synthetic information with a form similar to real identifiers. For instance, the personal name “John Doe” might be replaced by “William Withersby.” Challenge participants developed de-identification tools using the synthetic corpus, and a different, smaller set of discharge summaries also containing synthetic identifiers was used to evaluate the systems. In all, there were over 16 de-identification systems submitted from seven teams. The results of the challenge were revealed at a workshop associated with the 2006 AMIA Annual Fall Symposium. The evaluation results indicated that de-identification tools based on machine learning approaches[11 14] were superior to those based on rules, regular expressions, and dictionaries alone.

MITRE De-identification system

The MITRE team developed a de-identification tool for the i2b2 challenge using a machine learning classifier based on conditional random fields (CRF), as implemented in the open source Carafe toolkit.[18] The de-identification task was approached as a sequence-labeling problem, akin to the well understood named entity identification task.[19] The de-identification tool was developed by iteratively training the system and evaluating its performance, customizing the feature set, and experimenting with a parameter to skew performance towards recall at the expense of precision. The hypothesis was that for de-identification, high recall, or over-redaction, was preferable to high precision where some identifiers might be “exposed”. The MITRE system achieved some of the best performance scores at the challenge.[17]

MITRE submitted three system runs, two of which represented a precursor of the system presented in this paper. The most direct precursor achieved a token F-measure of 0.997 and phrase F-measure of 0.972. (A“phrase” corresponds to a PHI instance and can consist of one or more “tokens”.) The highest performance run involved the same system but with task-specific post-processing for regular expressions. This system achieved a token F-measure of 0.997, and phrase F-measure of 0.974. The i2b2 results are most comparable to the results for the second experiment in this paper (ie, R⇒R) with respect to Discharge Summaries. However, it should be recognized that the work reported here included no task-specific post-processing, in order to focus on the machine-learning aspects of the de-identification tool.

METHODS

Materials

For this project, we selected patient files from StarChart, the VUMC electronic medical record system.[20] Each patient file is organized as a longitudinal series of records that contain structured (eg, treatment and diagnosis codes) as well as unstructured data (eg, narrative progress and procedure notes), thus making it a desirable source for biomedical research projects. The system contains information dating back to 1984 and receives feeds from a diverse set of sources, including lab results, radiology reports, and external transcription companies. StarChart additionally provides a set of templates that enable clinicians to enter free text notes. We focused on patients with robust medical histories, such that each file contained at least one ECG and more than 10 documented records. The average size of a file was 14 565 lines with 334 records.

Rather than hand-annotate each term in every record, we applied a specialized version of DE-ID, a commercially available rules-based de-identification software tool,[5] to indicate the position and syntactic type (eg, name vs. date) of each patient identifier. DE-ID replaces identifiers with generic placeholders for the syntactic type. Our decision was based on several factors. First, DE-ID, in conjunction with several pre- and post-processing modules developed at the VUMC, is the current approach by which the VUMC de-identifies all of its medical files (over 1.5 million) for local investigator-initiated research projects. Second, in a pilot project conducted at the VUMC, it was shown that the specialized version of the DE-ID software exhibited a recall of over 99.9% for HIPAA Safe Harbor identifiers.[21] We note that the intent of this study was to replicate the DE-ID process through a machine-learning approach and not to determine if machine learning was better than DE-ID at achieving de-identification.

Resynthesis system

In preparation for the current research, MITRE prepared a resynthesis engine to create synthesized, but realistic, English replacements for PHI. The engine operates on medical records in which the PHI has been identified; these can be the original records or ones that have been scrubbed in the style of the DE-ID system. The information available to the resynthesis engine depends on how the documents are prepared; for example, for DE-ID scrubbed records, DE-ID preserves information about the number of tokens (words) in names, but not in places. Figure 1 provides an example of fictional pre- and post-scrubbed data.

The resynthesis engine first gathers a set of features from the input PHI. For instance, in the case of names, it attempts to

Figure 1 Left: DE-ID annotation. Right: resynthesized version of a partial Vanderbilt University Medical Center (VUMC) medical record.

De-identified VUMC Record	Resynthesized Record
PHYSICIAN: **NAME[WWW VVV], M.D.	PHYSICIAN: Dudley, Carmen, M.D.
PATIENT: **NAME[AAA, BBB M].	PATIENT: Ahmad, Jane Q.
MRN: **ID-NUM	MRN: ID43729
ADMITTED: **DATE[Jan 17 2003]	ADMITTED: Aug 21 2003
DISCHARGED: **DATE[Jan 20 2003]	DISCHARGED: Aug 24 2003
**NAME[BBB AAA] is a **AGE[over 89]-year-old woman with a history of a left renal mass who presented for laparoscopic partial nephrectomy… She was instructed to follow up with Dr. **NAME[UUU] in one week. She was given prescription for Percocet for pain control…	Jane Ahmad is a 95-year-old woman with a history of a left renal mass who presented for laparoscopic partial nephrectomy... She was instructed to follow up with Dr. Williams in one week. She was given prescription for Percocet for pain control…

reproduce the number of tokens, the capitalization pattern, and whether the phrase corresponds to a name with the last name first. For dates, it attempts to preserve the offset from the earliest date in the document, as well as the specific details of how the date was formatted. Any features that cannot be determined from the input are assigned randomized values based on a weighted, handcrafted estimation of the frequency of the possible values.

Once the feature values are determined, the engine generates replacement "fillers." The tokens for the fillers are drawn from a variety of sources, including weighted lists of first and last names provided by the US Census, and lists of cities, states, streets, medical facilities, and zip codes derived from various on-line resources.[22] The replacement fillers are assembled based on the features that the engine has already gathered; so, for instance, if the engine determines that a name consists of a last name followed by a first name and an initial (as it might determine from a pattern like **NAME[AAA, BBB M] or a PHI instance such as "Philips, Bruce R"), it will generate a new filler, such as "Ahmad, Jane Q" as shown in figure 1. If name coreference information is available, as is typically the case for DE-ID output, name fillers are selected in a consistent way to preserve the coreference. Similarly, date offsets are preserved, so that the pair of dates Jan. 17 and Jan. 20 is shifted, but the 3-day difference is preserved. The engine also caches replaced name tokens on a record-by-record basis; so "AAA" will correspond to "Ahmad" throughout the document.

Experimental environment

To test the effect of resynthesis on the performance of MIST, we designed an experimental system as presented in figure 2.

Generation of ground truth data

First, the VUMC-specialized DE-ID tool was applied to the original medical records (OMRs) to provide redacted medical records, as shown on the left side of figure 1. The OMRs and redacted records were then "relinked" to provide a gold standard set of annotated OMRs, where annotations capture the category and location of each redacted PHI in the original record. This created a pool of OMRs for training and testing of the de-identification models. The tag categories for PHI used in this study consisted of (1) AGE, (2) DATE, (3) DEVICE ID, (4) EMAIL ADDRESS, (5) ID-NUMBER, (6) INITIAL, (7) INSTITUTION, (8) NAME, (9) PHONE NUMBER, (10) PATH NUMBER (which corresponds to the tracking number associated with a pathology sample and is derivative of DE-ID's initial design for handling pathology reports), (11) PLACE (an umbrella term for geographical terms, such as cities and towns), (12) ROOM, (13)

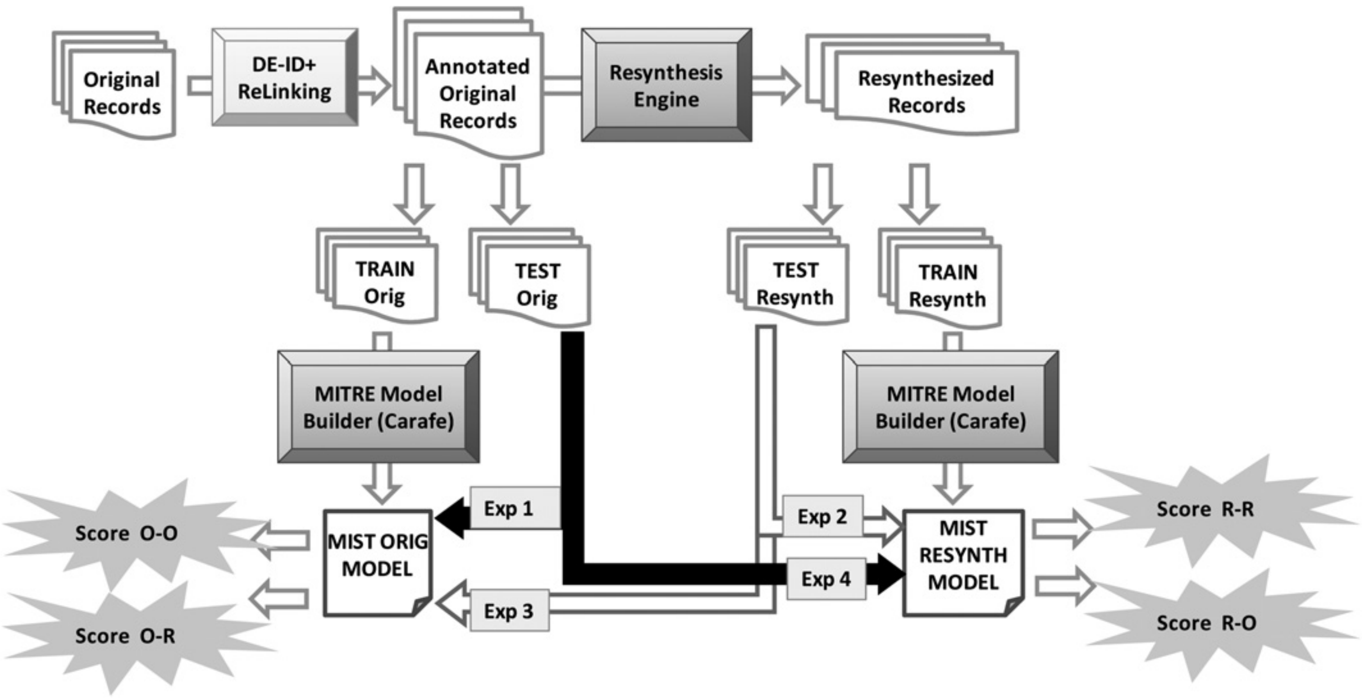

Figure 2 Training and testing framework for text de-identification with resynthesis.

STREET ADDRESS, (14) WEB LOCATION (eg, URLs and IP addresses), and (15) ZIP CODE.

We evaluated four classes of medical records, as defined by the VUMC: (1) Discharge Summaries (DS), (2) Letters, (3) Laboratory Reports (LAB), and (4) Orders. We also replicated the experiments on a hybrid corpus that consists of records from each class.

Production of training and test data

Next, we prepared training and test data for our resynthesis experiments. To create a pool of resynthesized medical records (RMRs), we began with the output of the DE-ID tool described above and ran these records through MITRE's resynthesis engine. The annotated OMRs and RMRs were then split into two groups and supplied to the MIST de-identification tool:

1. records to train de-identification models; and
2. records to test the resulting models.

The models built from the OMRs and RMRs are termed MIST-ORIG and MIST-RESYNTH, respectively (see figure 2).

Given the annotated medical records and the two de-identification models, we performed four sets of experiments. The first two experiments used training and test corpora drawn from the matched data sources (eg, training and testing with OMRs or training and testing with RMRs). The second two experiments used unmatched training and test corpora (eg, training drawn from OMRs, testing from the RMRs, or vice versa). The experiments were motivated as follows:

- Experiment 1 (O⇒O). The de-identification model was trained and tested with OMRs. This experiment evaluates the performance of the de-identification model with respect to the original identifiers. The context replicates the ideal training and evaluation environment; that is, when the tool developer has access to the actual medical records on which the tool will be applied.
- Experiment 2 (R⇒R). The de-identification model was trained and tested with RMRs. This evaluates the performance of the de-identification model with respect to synthesized data. The context simulates the scenario when the tool developer does not have access to real medical records. This experiment is also comparable to the i2b2 de-identification challenge competition.
- Experiment 3 (O⇒R). The de-identification model was trained with OMRs and tested with RMRs. This is an integrity check that investigates how well a model trained on real data can be applied to synthetic data. It determines how well the model created from the real data covers the RMRs. This environment corresponds to the situation in which developers use local data to train their system and then evaluate it in a public forum.
- Experiment 4 (R⇒O). The de-identification model was trained with RMRs and tested on OMRs. This replicates the scenario in which developers use public data to train their system, and the resulting models are subsequently applied as an "out of the box" solution to private records.

Training and test methods

For all experiments described in this paper, we applied the trainable de-identification component described in Wellner *et al*,[16] based on the Carafe conditional random field toolkit. We used the exact same set of features, learning parameters and decoding bias (which slightly prefers higher recall) that corresponds to the best results achieved on the AMIA i2b2 challenge data.[17] While it might be possible to improve performance by customizing the feature set for the data used in this paper, controlling for the set of features and learning parameters is motivated by the fact that: (1) it allows for easier interpretation and reproduction of the results; (2) this set of features was found to perform at a very high level on a similar task[14]; and (3) the use of an existing, fixed feature set applied to new data types, and possibly new PHI categories, represents a typical real-world use case for this type of system.

Metrics

We report the results of the evaluation using the following traditional information retrieval performance measures: recall, precision, F-measure, and accuracy. These measures are calculated at the token level, where each token in a phrase containing PHI must be identified as belonging to the specific PHI category to be considered correct.

Precision, recall, F-measure, and accuracy

Precision, recall, and F-measure are defined based on the number of PHI phrases or tokens identified. Accuracy is defined as the correct labeling for each word in the record, where true negatives are included in the score.

$$\text{Precision (p)} = \frac{\text{number of correctly identified PHI}}{\text{number of PHI identified by the system}}$$

$$\text{Recall (r)} = \frac{\text{number of correctly identified PHI}}{\text{number of PHI in the gold standard}}$$

$$\text{F-measure (f)} = \frac{2pr}{p+r}$$

$$\text{Accuracy} = \frac{\text{number of correctly labeled words}}{\text{all words in records}}$$

Error versus PHI exposure

We performed a detailed analysis to examine the errors associated with NAME identifiers at the PHI phrase level with respect to the O⇒O and R⇒R experiments. The aim of this analysis was twofold: first, to understand the kinds of errors that are committed by the de-identification tool; second, to characterize the rate at which de-identification errors actually expose PHI and, thus, create potential privacy violations.

Four cases of errors were considered:

- Missed: the system failed to identify any part of the PHI;
- Tag-Clashes: the system correctly identified a span of text, but assigned an incorrect identifier category;
- Span-Clashes: the system assigned the correct identifier category, but the extent was incorrect;
- Both-Clashes: neither the identifier category nor the extent was correct.

Examples for these cases are depicted in table 1. Note that Tag-Clashes never result in PHI exposure. In contrast, for Span-Clashes and Both-Clashes, PHI is exposed only if the marked span is too short because PHI is not exposed if the marked span is too long.

RESULTS

The number of records selected for the training and test corpora for each medical record class are summarized in table 2. We

Table 1 Examples of error types

Error type	Example
Correctly matched	… by <NAME>Sarah T. Smith </NAME>, head of…
Missed	… by Sarah T. Smith, head of…
Tag-clash	… by <PLACE>Sarah T. Smith </PLACE>, head of…
Span-clash	… by <NAME>Sarah T. Smith, </NAME> head of…
Both-clash	… by <PLACE>Sarah T.</PLACE> Smith, head of…

selected two times more records for the LAB and the Order corpora than in the DS and Letter corpora to compensate for the relatively small amount of text that resides in records of the latter classes. The training set of the hybrid corpus contains the same number of records from each class, whereas the test set contains all of the records from each of the four classes' test sets.

De-identification model training and testing

The number of instances for each PHI category (in terms of phrase) in each medical record class is summarized in table 3. Note that some classes lack instances of a certain PHI category. For example, the LAB training corpus contains no instances of AGE. As a result, a model trained from this corpus has no way of identifying ages in any test corpus. In addition, different tag categories appear with greatly different frequencies. Only six of the 15 PHI categories found in any of the record classes appear in at least three of the four classes: NAME, DATE, INITIAL, INSTITUTION, ID-NUMBER, and PLACE.

The results for the experiments are summarized in table 4, which is grouped by train-test pair. The rows of the table report the results for (1) each of the four record classes, (2) an aggregated result that was obtained by combining the results from each of the four record classes, and (3) a hybrid model (as described earlier). Since we are comparing results across different test sets, calculating statistical significance is problematic. To simplify matters, we assessed the reliability of the results using a bootstrap resampling approach. Details of this process are and its results can be found in online appendix A.

Here, we review the results and interpret certain findings. First, we begin with the O⇒O experiments, which evaluated the performance of the de-identification model trained on, and tested with, the OMRs. For this experiment, the F-measure ranged from 0.93 to 1.00. Second, in the R⇒R experiments, the de-identification model was trained and tested on RMRs. The F-measure was similar to the previous set of experiments, with a range between 0.96 and 0.99. We also observed that the F-measure tended to increase in comparison to its counterpart in the previous experiment. Third, the O⇒R experiment explores the relationship between original and resynthesized data by training the model with OMRs and testing with RMRs. Notice that the F-measure significantly drops in comparison to the first two experiments. Specifically, the range of the F-measure was from 0.78 to 0.89. Additionally, the precision tends to be higher than the recall. This reverses the pattern seen in the O⇒O and R⇒R results, where recall was higher than precision. Finally, in the R⇒O experiment, the de-identification model was trained

Table 2 Class and medical record sizes in Vanderbilt University Medical Center medical-record corpora

	Evaluation		Record information (mean)		
Record class	Train	Test	No. of lines	Line length	No. of identifiers
DS	200	50	95	36	23
LETTER	200	50	60	31	16
LAB	400	100	24	19	9
ORDER	400	100	17	18	5
HYBRID	200	300	44	30	12

Table 3 Distribution of protected health information phrases in medical-record corpora

Protected health information category	Record class									
	DS		LAB		LETTER		ORDER		HYBRID	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Age	173	37	0	0	65	7	0	0	58	44
Date	1390	349	401	100	810	208	1170	285	754	942
Device-ID	0	0	0	0	1	0	0	0	0	0
Email	0	0	0	0	0	0	0	0	0	0
ID-num	970	223	1470	379	497	142	427	110	612	854
Initials	3	1	4	0	22	7	0	0	4	8
Institution	218	54	0	0	46	13	6	1	66	68
Name	1878	445	1273	321	1620	413	682	163	1142	1342
Path-NUMBER	0	0	400	100	0	0	0	0	50	100
Phone	1	0	2	1	0	0	0	0	0	1
Place	57	11	11	5	137	34	4	2	43	52
Room	1	0	0	0	3	2	0	0	0	2
Street	6	0	0	0	21	9	0	0	7	9
Web-loc	0	0	0	0	3	0	0	0	0	0
Zip-code	6	0	0	0	102	30	0	0	23	30
Total	4703	1120	3561	906	3327	865	2289	561	2759	3452

Table 4 Performance measures for train-test experiments and label-blind protected health information exposure

Record class	Recall	Precision	F-measure	Accuracy	Protected health information exposure (1-label-blind recall)
O⇒O Experiment					
DS	0.986	0.946	0.966	0.993	0.014
Lab	0.966	0.905	0.935	0.983	0.034
Letter	0.956	0.931	0.944	0.986	0.040
Order	0.999	0.993	0.996	0.999	0.001
Aggregate	0.978	0.943	0.960	0.990	0.022
Hybrid	0.962	0.925	0.943	0.986	0.035
R⇒R Experiment					
DS	0.986	0.972	0.979	0.998	0.010
Lab	0.995	0.991	0.993	0.999	0.005
Letter	0.965	0.962	0.963	0.996	0.032
Order	0.990	0.989	0.989	0.999	0.010
Aggregate	0.983	0.977	0.980	0.998	0.014
Hybrid	0.970	0.960	0.965	0.997	0.022
O⇒R Experiment					
DS	0.871	0.919	0.894	0.990	0.101
Lab	0.731	0.843	0.783	0.987	0.268
Letter	0.832	0.910	0.869	0.987	0.155
Order	0.788	0.984	0.875	0.992	0.212
Aggregate	0.816	0.913	0.862	0.989	0.171
Hybrid	0.842	0.911	0.875	0.990	0.147
R⇒O Experiment					
DS	0.674	0.887	0.766	0.961	0.324
Lab	0.348	0.723	0.470	0.899	0.652
Letter	0.769	0.852	0.808	0.955	0.224
Order	0.766	0.834	0.799	0.926	0.234
Aggregate	0.642	0.841	0.728	0.942	0.355
Hybrid	0.404	0.789	0.535	0.914	0.592

with RMRs and tested with OMRs. Notice the F-measure dropped even further than in the O⇒R experiment, ranging from 0.47 to 0.81. Although both recall and precision drop compared to the O⇒O case, the drop in F-measure is dominated by the much larger drop in recall, Also, we found that the best F-measure in this experiment was achieved by the Letter corpus, whereas in the previous experiment, this honor belonged to the DS corpus.

Effect of training set size

Figure 3 provides an indication of how training set size influences the recall, precision, and F-measure of the various experiments. These results correspond to models trained for discharge summaries only. When training and testing on like data (O⇒O and R⇒R experiments), high F-measures in the mid-90s are obtained with a small quantity of 25 training discharge summaries, and scores climb with additional training data and level off with minimal improvement after about 100 notes. This suggests that getting an O⇒O system up and running in a new setting can be accomplished with a very small seed effort devoted to creating a training set.

F-measures for the cross-experiments (O⇒R and R⇒O) are lower. For the O⇒R experiment, F-measures start in the mid-80s, when training with 25 notes, and follow a similar curve as training notes are added. After 100 notes, the scores level off near an F-measure of 0.9. The scores for the R⇒O experiment are the lowest of all the experiments and show the most variability with training data. With 25 training discharge summaries, the F-measure is below 0.5 but eventually climbs to the mid-70s before leveling off, again at 100 notes. The regularity introduced by training on resynthesized discharge summaries, and the correspondingly poorer fit of the resynthesized model to the greater variability of original data, is a likely reason that the scores are lowest in this condition. The relatively larger increase in scores as more training discharge summaries are used (ie, F-measures climb from the mid-40s to the mid-70s) may be an indication that using more resynthesized data introduces more diversity to the model, resulting in a better, although still weak, fit to the original data.

Analysis of NAME errors and error metrics

Table 5 summarizes the performance of the system on the NAME category, with particular attention to what percentage of each error type reveals PHI. We observe that for most types of error, such as Tag-clash and Both-clash, no PHI is leaked. For the Span-clash case, fewer than 20% of these errors result in PHI disclosure, because in the vast majority of cases the name content was contained in another neighboring tag, with no PHI exposure. Adding the NAME errors that do not expose PHI to the correctly matched group leads to a 3–5% increase in the total protected score; the last column of table 4 summarizes PHI exposure by experiment and document type.

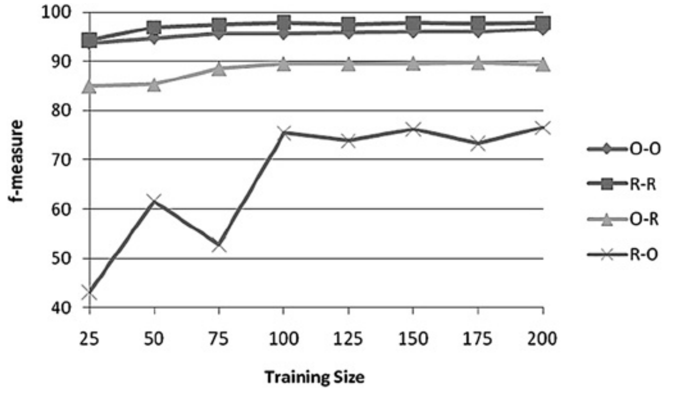

Figure 3 Impact of number of training documents on F-measure for discharge records.

In addition to providing a better sense of the system's PHI exposure rate, the error analysis supplies useful information for improving the system. We observed that many of the error types are instances in which the system produced tags that are too large, for example, encompassing extra content or including additional PHI and thus not resulting in any exposure. In future research, we plan to explore this phenomenon by experimenting with different tokenization methods that may result in finer-grained tagging.

Different types of errors have different effects and potential consequences. Failure to mask any word within an occurrence of PHI risks exposure of identity. Thus, a true "miss" is a very serious error. This should be distinguished from a Tag-clash, where the word is tagged with an incorrect category of PHI. Such an error may reduce the readability of the resulting de-identified record, but it will not expose PHI. To calculate the frequency of the system's exposure of PHI, we collapsed all PHI categories into a single "all-PHI" group and evaluated the system's ability to distinguish tokens in PHI from tokens that contained no PHI and therefore did not need to be masked. The higher the recall of the all-PHI measure, the lower the exposure of PHI. Of course, a system could trivially achieve perfect recall, and thus zero risk of PHI exposure, by masking all tokens in each record. Therefore, it is critical to measure precision as well as recall to gain insight into the system performance. In addition, if the goal is to preserve information, then tag accuracy is also important because an incorrect tagging would make reading the resulting record more difficult. It remains an interesting open issue to understand the extent to which such tag clashes or spurious tags cause loss of important biomedical information.

Table 5 Error rates and exposure analysis for the NAME protected health information (PHI) category

		Record class				
		DS	Lab	Letter	Order	Hybrid
O⇒O Experiment						
Total no of cases		438	3864	4587	5161	2205
Correct match (%)		93.4	92.2	92.1	91.9	91.7
False negatives (%)		6.6	7.8	7.9	8.1	8.3
False-negative type	Missed (%)	3.2	3.5	3.6	3.8	3.9
	Span-clash (%)	3.4	4.2	4.2	4.2	4.3
	Tag-clash (%)	0	0.1	0.1	0.1	0.1
	Both-clash (%)	0	0.1	0.1	0.1	0.1
% of errors leading to PHI exposure	Missed (%)	93	73	72	69	81
	Span-clash (%)	13	14	15	14	18
	Tag-clash (%)	0	0	0	0	0
	Both-clash (%)	0	0	0	0	0
R⇒R Experiment						
Total no of cases		883	4185	4998	5124	3535
Correct matches (%)		93.2	92.6	91.7	92.1	91.7
False negatives (%)		6.8	7.4	8.3	8.0	8.4
False-negative type	Missed (%)	2.7	3.3	3.9	3.7	3.7
	Span-clash (%)	3.9	4.0	4.3	4.2	4.5
	Tag-clash (%)	0.2	0.1	0.1	0.1	0.1
	Both-clash (%)	0	0.1	0.1	0.1	0.1
% of errors leading to PHI exposure	Missed (%)	88	72	69	69	73
	Span-clash (%)	9	14	14	14	15
	Tag-clash (%)	0	0	0	0	0
	Both-clash (%)	0	0	0	0	0

Token level PHI exposure for the different experiments and document types is summarized in the last column of table 4. For the O⇒O experiment, the rate of PHI exposure ranges from 0.1% for Order to 4% for Letter. For the R⇒R experiment, the rates are somewhat lower, ranging from 0.5% for Lab to 3.2% for Letter. On the whole, the system trained and tested on OMRs has about 50% higher PHI token exposure than the results from training and testing on RMRs (2,2% vs 1.4% for the aggregate O⇒O vs R⇒R results).

Examining the resynthesized model versus real medical records penalty

The above experiments simulate two use cases. In the first case, users train models with annotated OMRs from their local organization. In the second case, the users apply a model constructed from external data, at least for an initial round of annotation. In the latter scenario, the training data would most likely be drawn from a community-wide corpus of RMRs. Unfortunately, we are unable to simulate this scenario precisely because, at the moment, we only have access to medical data from a single institution. However, since the RMRs are drawn from the same source that yields the OMRs, the R⇒O experiment indicates the most likely best-case performance of the system in such a scenario.

The above results raise the question "Why is the F-measure for the R⇒R (ie, the evaluation on resynthesized data) substantially greater than that of the R⇒O (ie, the out-of-the-box application) setting?" One hypothesis for the disparity is that the test corpus in the R⇒O experiment harbors a substantial share of vocabulary that fails to appear in the training corpus, or "out-of-vocabulary" terms. To investigate this hypothesis, we considered the out-of-vocabulary rates, as well as the F-measures, for each of the experiments and document classes. Figure 4 visualizes the relationship between the F-measures by grouping the tests by experiment and document class, respectively. We note that in the homogeneous train-test cases (R⇒R and O⇒O) there are fewer out-of-vocabulary words than in the mixed train-test cases (O⇒R and R⇒O). Yet, on the whole, there is only a weak correlation between the out-of-vocabulary rate and the F-measure in both graphs. Further details regarding this analysis can be found in online appendix B.

We further explored the performance across the different experimental conditions by comparing the F-measure in each experiment by PHI category, as summarized in table 6. We focus on the dominant identifier types that account for over 90% of the PHI occurrences, namely AGE, DATE, ID-NUMBER, INSTITUTION and NAME.

1. Robust identifiers. AGE is an identifier that yields highly similar results in all experimental settings.

Table 6 F-measure by identifier category and number of training instances

Protected health information category	No. of training instances	F-measure			
		O⇒O	R⇒R	O⇒R	R⇒O
Age	238	0.989	0.989	0.989	0.989
Date	16876	0.978	0.991	0.871	0.781
Id-number	3737	0.877	0.999	0.568	0.216
Initials	87	0.829	0.667	0.000	0.000
Institution	547	0.964	0.950	0.298	0.933
Name	13288	0.965	0.975	0.950	0.802
Path-num	1104	1.000	1.000	0.830	0.054
Place	217	0.812	0.752	0.505	0.703
Street-address	88	0.735	0.839	0.533	0.513
Zip-code	168	0.969	0.983	0.966	0.718
Total	36395	0.960	0.980	0.862	0.729

2. Non-robust identifiers. This category represents identifiers that exhibit a considerable decrease in F-measure. Identifiers in this category include ID-NUM and INSTITUTION. In these cases, the resynthesized ID-NUM or INSTITUTION may bear little resemblance to the IDs in the OMRs, because the replacement engine was unable to appropriately replicate the patterns.
3. Moderately robust identifiers. These are identifiers that exhibit a moderate decrease in performance and dominate the inventory of PHI. For example, the NAME and DATE identifier categories yielded a 15 and 19 point decrease in F-measure, respectively, and cover 82% of all tags in the training set. These dominate the inventory of PHI, and as a result, the same pattern appears in the overall scores: R⇒R scores are highest, slightly outperforming the O⇒O experiment. Next there is a larger drop-off to the scores of the O⇒R experiment, and then there is the largest drop-off to the scores of the lowest-performing experiment, R⇒O. As discussed earlier, these results stem from the fact that resynthesis regularizes data, which loses some of the variability present in OMRs. The more homogenous nature of resynthesized data accounts for R⇒R scores being higher than O⇒O scores. While the scores for the cross-experiments are lower, training on the more regularized data and testing on the original data (R⇒O) results in the lowest scores.

DISCUSSION

This study illustrates how identifier resynthesis influences medical text de-identification under various real-world conditions. There are several take-home messages that can be extracted from our investigations, as well as limitations that should be highlighted.

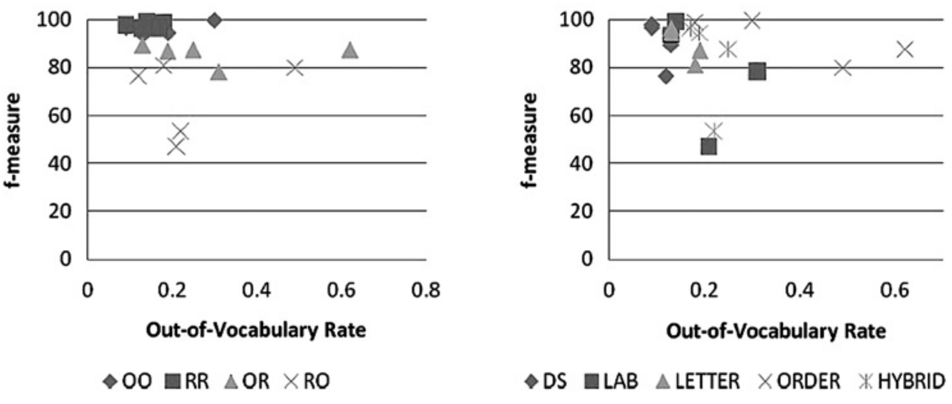

Figure 4 Relative change in F-measure as a function of out-of-vocabulary rate. The scores are grouped by (left) experiment type and (right) document type.

Principal findings

First, it is important to note that while training and testing on synthetic records is more accurate than OMRs, the results do not imply that the evaluation of medical record de-identification tools on resynthesized data is a hazardous endeavor. Rather, we found that there is a high similarity in the F-measure scores generated by the results in the O⇒O and R⇒R experiments. For example, in figure 5, we plot the normalized difference between the F-measures. Formally, this is calculated as (x−y)/x, where x and y are the F-measures from the O⇒O and R⇒R experiments, respectively. Notice that the difference is never greater than 6% (for PHI tokens of the LAB class) and it is as low as 0.7% for PHI tokens of the ORDER class.

This analysis further suggests that the relationship between the synthetic and original results is strongest in information-rich documents, as characterized by the correlation between relative difference in F-measure and the average line length per record. This indicates that the results achieved by training and testing on synthetic medical records may hold true when training and testing on the OMR models. The results of the O⇒R and R⇒O experiments, however, suggest that if a model is trained on resynthesized data, the accuracy of the model may not necessarily translate to real systems, and vice versa. The R⇒O experiment, in particular, is a strong indicator of this finding. Here, the model was trained on synthetic data, which exhibited a high F-measure when tested on synthetic data, but a much lower F-measure when tested on the OMRs.

We further explored whether the large difference between the results of O⇒O and R⇒O experiments occurred by chance by performing a statistical significance test. We applied the same significance testing methodology that was used in the 2006 AMIA Challenge.[11] With the given two models, we randomly shuffled the records in the test set 99 times to create pseudo-models. For each of these models, we checked whether the difference between the F-measures is greater than the difference between the F-measures when actual models are used. This experiment indicated that the difference between the models trained with OMRs and RMRs is significant at level 0.1. This result is a clear example of the effects of resynthesis on medical records.

Second, many de-identification errors do not actually expose PHI, but rather map PHI to the wrong identifier category. While this is still an issue with identifier classification accuracy, it is not a detriment to privacy protection in medical record sharing. It is doubtful that institutional review boards, overseeing the data sharing process, would punish data managers for misclassification errors.

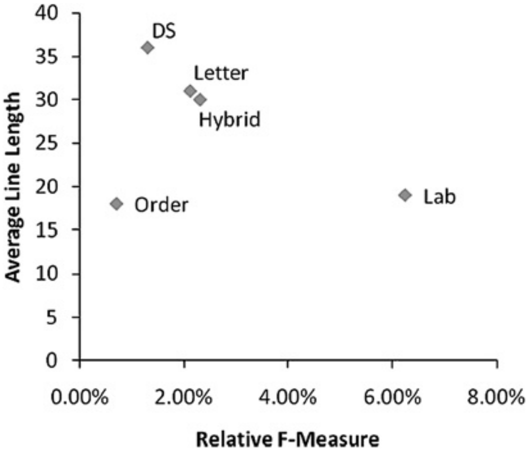

Figure 5 Normalized difference in the F-measures of the results in O⇒O and R⇒R experiments for protected health information tokens.

Third, it appears that seeing the PHI tokens in more contexts is better for a de-identification tool than seeing them in fewer contexts. This is not a radical or particularly surprising result for the training strategy used in our experiments, but it does point in some useful directions. For instance, the corpus-source conditions are not perfectly reproducible in the real-world conditions that R⇒O simulates. The builders of the de-identification model do not have access to the identifiers from the target corpus and, in fact, cannot. However, there are strategies for augmenting a previously constructed model with additional synthesized identifiers at decode time, which may very well reduce the drop-off by several points of F-measure.

Technical limitations

Manual hand annotation of medical records is a labor-intensive process that requires a significant amount of time and money. Instead, we used DE-ID, an existing de-identification system. Though DE-ID is not 100% accurate, it was specifically tailored for VUMC records to maximize performance. Nonetheless, even with the specific implementation of DE-ID, the annotated records may contain some errors that add noise to the models.

We also explored the effects of data preparation on the performance for different identifier categories by comparing the F-measure in each experiment by PHI category across the four experiment types, as summarized in table 6. There are two ways in which the data preparation for the experiments has interacted with the reported results. First, the preparation of the training data was influenced by the use of DE-ID. As described earlier, the OMRs were processed using DE-ID, and the resulting annotated records were the basis for the training and test annotations, as well as input for the resynthesized data. DE-ID removes PHI using specific strategies that retain varying amounts of information about the internal structure of the identifiers, depending on the identifier in question.

- DATE: DE-ID performs a systematic date shift within each document, and in the process it also normalizes the appearance of the date expressions; for example, a date such as “May 6, 1995” might be remapped to “**DATE[Feb 2 1995]” (omitting the comma separating day and year).
- NAME: DE-ID preserves the internal structure of the name as well as identifiers within the document. If there is a name “John Z. F. Smith” and a later mention of “Smith,” DE-ID would produce in the first instance: **NAME[GGG H. I. JJJ] for the first, and NAME[JJJ] for the second mention.
- ID: DE-ID does not preserve the internal structure on ID-NUM. All identifiers are simply mapped to **ID-NUM.

The resynthesis module then operated on the DE-ID annotated text, making use of any structural information available. Specifically, when operating on DE-ID output, the resynthesis module uses the internal information preserved for NAME and DATE, but it has no information on the internal structure of ID-NUM.

To investigate how this may have influenced our results, we performed a small study to compare the distribution of DE-ID NAME patterns, compared to those produced for the resynthesized data. In a sample of 1000 documents, DE-ID tagged over 60 distinct types of **NAME structures, but when the DE-ID tagged records were used as the input to the resynthesis module, only eight structures were produced, as shown in table 7. However, it should be noted that these structures account for over 90% of the structures in the original data.

Given this apparent regularization of the data during resynthesis, it is not surprising that we see a modest increase in the

Table 7 Frequency distribution of name patterns generated by DE-ID and the MITRE Identification Scrubber Toolkit resynthesizer

Pattern	No. of tokens	Frequency	
		DE-ID	Resynthesized
**NAME[AAA]	1	1113	1241
**NAME[AAA AAA]	2	693	798
**NAME[AAA, AAA]	2	629	635
**NAME[AAA, AAA A.]	3	606	627
**NAME[AAA, AAA AAA]	3	310	351
**NAME[AAA A. AAA]	3	306	329
**NAME[AAA AAA AAA]	3	198	301
**NAME[AAA A. A. AAA]	4	0	3
**NAME[A]	1	114	0
Over 60 other patterns	Variable	316	0
Total		3969	4285

R⇒R F-measure for NAME (0.975) compared to O⇒O (0.965). Training on OMRs and testing on RMRs (ie, O⇒R) produces a more noticeable drop to 0.950, and finally training on RMRs and testing on OMRs (ie, R⇒O) produces a large drop to 0.802. The large drop is somewhat surprising, given the estimated 90% overlap in distribution between the internal structure of names in the original (DE-ID tagged) data and the resynthesized data. Some of this may be due to apparent errors in the DE-ID tagging, which produces unlikely name patterns such as "**NAME[AAA, AAA AAA, AAA AAA]," perhaps resulting from the merger of several names. A system trained on more regular patterns might find several identifiers in such a string, leading to a false negative and multiple false positives, which might exaggerate the effect of the different distributions for training and test in the R⇒O case.

We have observed similar discrepancies in performance for DATE. The DATE expressions undergo two stages of regularization: one from the DE-ID offset and regularization, and a second via the resynthesis module. Finally, for ID-NUM, because all internal identifying structure in the ID is lost at the DE-ID stage, the resynthesized training data are almost a complete mismatch for identifying IDs in the OMRs (R⇒O of 0.22%). If the resynthesis had operated directly on the OMRs, it would have preserved the alphanumeric patterns of the various identifiers, and we would expect performance to improve considerably.

Future directions

There are numerous ways in which this research is expected to move the field forward. Here, we list only a few of the more significant. First, this work illustrates (eg, the O⇒O experiments) that machine learning-based de-identification tools require a relatively small number of documents to achieve results that are, although not equivalent, at least competitive with a tailor-made system (ie, DE-ID as tailored to Vanderbilt). Second, healthcare organizations with limited time will search for ready-made solutions, but many industrial software shops and natural language-processing experts reside outside of the healthcare realm with limited access to medical records. Thus, it is anticipated that de-identification models and software may be trained on resynthesized data, and the end users of such tools will apply it directly to their medical record systems. In this case, our work (eg, experiments R⇒O) provides a clear example of why such a practice is undesirable. Third, this research illustrates that the results of systems trained and applied to RMRs (eg, the R⇒R experiments) may hold when trained and evaluated on OMRs (eg, the O⇒O experiments). This is important because it offers justification for organizations to invest time and effort in annotating a relatively small number of documents (eg, we used 200 in our experiments).

Beyond moving the field forward, our investigations suggest several routes for new informatics research. First, it is important to recognize that our work does not imply that rules-based systems are worthless and should be discarded. The out-of-vocabulary results suggest that it is possible that a robust solution will require the combination of rules and machine-learning-based systems. We believe that additional research is necessary to determine the optimal combination of such techniques. Second, our study is only as robust as the MIST system and the applied resynthesis process, but there is no standard resynthesis framework. We believe that this work provides justification for further work in the field of resynthesis theory and tools for evaluation. Third, it should be noted that this study fixed the set of features and parameters for the de-identification system. An interesting avenue for future work is to examine how different feature sets, as well as machine learning frameworks, affect the results of training on one type of corpus (eg, resynthesized) and testing on another (eg, original data). For example, in some cases, the learner appears to be over-fitting; a smaller feature set may help the learner generalize better to "different" data. Along these lines, methods developed to improve domain and genre adaptation for statistical phrase identification systems would be interesting to pursue.[23]

CONCLUSIONS

Limited access to real medical records forces de-identification system developers to work with resynthesized records. Thus, resynthesized medical records are crucial for developing accurate de-identification systems that can be used on a wide scale. In this paper, we investigated the effects of resynthesis on the accuracy of medical text de-identification by experimenting on the original medical records from VUMC with the MIST de-identification system. We evaluated various real-world scenarios to train, compare, and apply de-identification tools in the context of synthetic information. The results indicated that when the training and testing records are produced from the same class of medical record (eg, both synthetic), the system returns a high accuracy. However, this finding did not hold when different types of records were used for the training and testing (eg, training on synthetic and testing on real). Our investigations also indicated how the resynthesis biases evaluations and which factors should be considered when designing medical record resynthesis tools. In future work, we intend to apply this knowledge to disseminate robust medical record resynthesis software to enable research with de-identified medical records and objective comparison of de-identification tools.

Acknowledgments The authors would like to thank J Paulett and K Benitez at the Vanderbilt Health Information Privacy Laboratory, for helpful discussions and technical assistance. The authors also recognize the assistance of D Giuse, David Staggs, and S Wang at the Vanderbilt University Medical Center, for their assistance in organizing and accessing the electronic medical records studied in this project. Finally, the authors need to thank M Basford, D Masys, and J Pulley, for their timely feedback on the Vanderbilt Synthetic Derivative, and Y Saygin and S Yeniterzi of Sabanci University, for helpful discussions during this research project.

Contributors This work was conducted while RY was a visiting graduate student at Vanderbilt University.

Funding The Vanderbilt participation was sponsored in part by a seed grant from the Vanderbilt University International Office.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **US Dept. of Health and Human Services.** Standards for privacy of individually identifiable health information, final rule. *Fed Regist* 2002;**67**:53181—273. 45 CFR, Parts 160—4.
2. **Beckwith B,** Mahaadevan R, Balis U, *et al*. Development and evaluation of an open source software tool for de-identification of pathology reports. *BMC Med Inform Decis Mak* 2006;**6**:12.
3. **Berman J.** Concept-match medical data scrubbing: how pathology text can be used in research. *Arch Pathol Lab Med* 2003;**127**:680—6.
4. **Friedlin J,** McDonald C. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* 2008;**15**:601—10.
5. **Gupta D,** Saul M, Gilberson J. Evaluation of a de-identification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol* 2004;**121**:176—86.
6. **Morrison F,** Li L, Lai A, *et al*. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *J Am Med Inform Assoc* 2009;**16**:37—9.
7. **Ruch P,** Baud R, Rassinoux AM, *et al*. Medical document anonymization with a semantic lexicon. *Proc AMIA Annu Symp* 2000:729—33.
8. **Sweeney L.** Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Symp* 1996:333—7.
9. **Bickel D,** Schwartz R, Weischedel R. An algorithm that learns what's in a name. *Mach Learn* 1999;**34**:211—31.
10. **Neamatullah I,** Douglass M, Lehman L, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008;**8**:32.
11. **Szarvas G,** Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using a iterative machine learning framework. *J Am Med Inform Assoc* 2007;**14**:574—80.
12. **Taira R,** Bui A, Kangarloo H. Identification of patient name references within medical documents using semantic selectional constraints. *Proc AMIA Annu Symp* 2002:757—61.
13. **Uzuner O,** Sibanda T, Luo Y, *et al*. A de-identifier for medical discharge summaries. *Artif Intell Med* 2008;**42**:13—35.
14. **Wellner B,** Huyck M, Mardis S, *et al*. Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc* 2007;**14**:564—73.
15. **Zhang L,** Pan Y, Zhang T. Focused named entity recognition using machine learning. *Proc 27th Annual International ACM SIGIR Conference* 2004:281—8.
16. **Dorr D,** Phillips W, Phansalkar S, *et al*. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods Inf Med* 2006;**45**:246—52.
17. **Uzuner O,** Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**:550—63.
18. *Carafe: conditional random fields, etc*. http://sourceforge.net/projects/carafe/
19. **Ananiadou S,** Friedman C, Tsujii J. Introduction: named entity recognition in biomedicine. *J Biomed Inform* 2004;**37**:393—5.
20. **Jirjis J,** Weiss J, Giuse D, *et al*. A framework for clinical communication supporting healthcare delivery. *Proc AMIA Annu Symp* 2005:375—9.
21. **Roden D,** Pulley J, Basford M, *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;**84**:362—9.
22. **US Census Bureau.** Frequently occurring first names and surnames from the 1990 census. http://www.census.gov/genealogy/names/
23. **Daumé H III.** Frustratingly easy domain adaptation. *Proceedings of the Conference of the Association for Computational Linguistics* 2007.