Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text

David Carrell,[1] Bradley Malin,[2,3] John Aberdeen,[4] Samuel Bayer,[4] Cheryl Clark,[4] Ben Wellner,[4] Lynette Hirschman[4]

► Additional appendices are published online only. To view these files please visit the journal online. (http://dx.doi.org/10.1136/amiajnl-2012-001034).

[1]Group Health Research Institute, Seattle, Washington, USA
[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
[3]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA
[4]The MITRE Corporation, Bedford, Massachusetts, USA

Correspondence to
Dr David Carrell, Group Health Research Institute, 1730 Minor Ave, Suite 1600, Seattle WA 98101, USA; carrell.d@ghc.org

Received 16 April 2012
Accepted 12 June 2012
Published Online First 6 July 2012

ABSTRACT

Objective Secondary use of clinical text is impeded by a lack of highly effective, low-cost de-identification methods. Both, manual and automated methods for removing protected health information, are known to leave behind residual identifiers. The authors propose a novel approach for addressing the residual identifier problem based on the theory of Hiding In Plain Sight (HIPS).

Materials and Methods HIPS relies on obfuscation to conceal residual identifiers. According to this theory, replacing the detected identifiers with realistic but synthetic surrogates should collectively render the few 'leaked' identifiers difficult to distinguish from the synthetic surrogates. The authors conducted a pilot study to test this theory on clinical narrative, de-identified by an automated system. Test corpora included 31 oncology and 50 family practice progress notes read by two trained chart abstractors and an informaticist.

Results Experimental results suggest approximately 90% of residual identifiers can be effectively concealed by the HIPS approach in text containing average and high densities of personal identifying information.

Discussion This pilot test suggests HIPS is feasible, but requires further evaluation. The results need to be replicated on larger corpora of diverse origin under a range of detection scenarios. Error analyses also suggest areas where surrogate generation techniques can be refined to improve efficacy.

Conclusions If these results generalize to existing high-performing de-identification systems with recall rates of 94—98%, *HIPS could increase the effective de-identification rates of these systems to levels above 99% without further advancements in system recall.* Additional and more rigorous assessment of the HIPS approach is warranted.

INTRODUCTION

A significant quantity of information about a patient's health, diagnosis, and treatment status is documented only in clinical narratives, such as progress notes, hospital discharge summaries, and radiology reports.[1] The unstructured data in these reports are rich in description and invaluable to a growing variety of activities, ranging from public health[2–4] to biomedical research investigations.[5–7] At the same time, there is a need to make such information available on a broader scale to further the development of next-generation natural language processing and biomedical data mining systems.[8 9] However, the presence of identifying information in such documents constrains their use and dissemination. This limitation is rooted in the threat to privacy, as laid out in the Privacy Rule of the US Health Insurance Portability and Accountability Act (HIPAA), which regulates the dissemination of protected health information (PHI).[10]

The Privacy Rule permits dissemination of patient data, including clinical text, when it is 'de-identified', and provides several methods for rendering health records into such a state. One such method is Safe Harbor, which corresponds to the removal of 18 types of potential identifiers.[10] Various automated de-identification systems have been developed to satisfy this Rule in the context of unstructured narratives which are available as commercial,[11] as well as free open-source systems.[12 13]

These systems decompose de-identification into two steps: first, they use rule-based and/or statistical algorithms to *identify* PHI; and second, they *replace* identified PHI with surrogates, which may be symbols, 'scrambled' strings of characters similar to those in the original PHI, or fictitious surrogates.[14 15] Sweeny replaced proper names with fabricated names that resembled, but did not exactly match, any known name (eg, the fabricated female first name 'Kathel').[14] The best automated and manual de-identification systems have produced comparable results, achieving recall rates (defined as the number of identifiers detected divided by the total number of identifiers in the reference standard) in the 94—98% range.[12 16–22] Performance varies by PHI type and document type. For example, Gardner and colleagues reported recall rates of 100%, 98.8%, and 96.3% for dates, medical record numbers, and ages, respectively, in pathology reports[23]; and Friedlin and McDonald reported recall rates of 98.4% and 95.7%, overall, for all identifier types in pathology reports and narrative reports, respectively.[24]

To date, automated approaches have not been widely adopted, and many organizations tend to rely upon manual de-identification strategies instead. In a survey of Institutional Review Boards (IRBs) conducted by us, only two out of eight respondents said that their IRB relied solely on automated de-identification methods (online appendix A, question 3); the reasons given for using manual de-identification included effectiveness, ease of explanation, and ease of implementation (online appendix A, question 6). However, manual

approaches are costly, do not scale to large corpora, and may in fact be less accurate than automated systems.[20–22] Consequently, clinical text tends to be made available in limited quantities, for specific projects, and at considerable expense. This results in missed opportunities for multisite collaborations, such as large-scale population-based studies, and impedes development of robust clinical natural language-processing methods, which require training and testing corpora from multiple institutions.[1]

Overcoming barriers to more widespread secondary use of clinical text requires scalable, cost-effective solutions to remove PHI. In particular, it requires addressing the 'residual identifier problem': current de-identification processes—whether manual or automated—will fail to redact some (small) proportion of identifiers, leaving 'residual' (unredacted) identifiers in place, thereby placing patient privacy at risk. Data use agreements, which provide safeguards against re-identification or other misuse of shared text, do not supplant the need for effective de-identification; even limited data use agreements require de-identification of direct patient identifiers. Given the inherent richness of many clinical notes, we believe it is unlikely that any method of removing personal identifiers, automated or manual, can consistently achieve 100% recall. Assuming some degree of residual identifiers will always be present, we believe scalable solutions will achieve greater efficacy by rendering residual identifiers less risky, rather than trying to eliminate them entirely.

De-identification systems that replace detected identifiers with arbitrary symbols assure that 100% of residual identifiers will be clearly evident—simply because they are the only unaltered identifiers present. It has been proposed that this shortcoming may be addressed via the theory of Hiding In Plain Sight (HIPS),[25] which is based on the principle of obfuscation.[26] Obfuscation entails hiding the true meaning of something by making it ambiguous, confusing, or otherwise difficult to interpret. According to the theory of HIPS, replacing detected identifiers with realistic-looking but fictitious synthetic surrogates (eg, names and phone numbers that appear to be real, but which are not) may effectively obfuscate any residual or 'leaked' identifiers that remain in the text; 'leaked' PHI that is indistinguishable from synthetic surrogates presents a minimal threat to privacy. The novelty of the HIPS approach is that it attempts to improve de-identification, not by improving a system's recall rate, but rather by accepting imperfect recall as a given, and using obfuscation to address the residual identifier problem. This shortcoming of traditional redaction approaches and the potential advantages of HIPS are illustrated in figure 1. We hypothesize that if the vast majority of PHI identified by existing high-performing systems is replaced with realistic surrogates, those surrogates will, collectively, protect the relatively small number of residual identifiers left behind from being recognized by recipients of the data.

The goal of this paper is to report on a pilot study to assess the capacity of a specific implementation of HIPS to conceal from human readers any residual identifiers in actual clinical narratives de-identified by an automated system. In preparation for this study, we presented the HIPS concept to IRB managers to obtain their views on its viability as a de-identification strategy. Managers responding to a brief survey (online supplement, appendix A) indicated that HIPS may be an acceptable approach for automating de-identification of a clinical text if proven effective by rigorous evaluation.

METHODS

Materials

Clinical document corpora

Two sets of clinical documents were used in these experiments; one containing a high density of identifiers, and the other an average density. The high-density corpus consisted of progress notes from 131 first-time oncology department encounters for randomly selected Group Health patients with pathologically confirmed primary breast cancers in 2008. Such notes typically contain numerous identifiers, as the oncologist documents the new patient's personal and health history. We also included the corresponding pathology report. The average-density corpus consisted of 150 randomly selected progress notes from Group Health family practice encounters in 2011. To increase the proportion of notes of average length, we oversampled notes with 400–500 words. The resulting 175-note corpus represented 175 unique patients.

The reference standard for all documents used in our experiments was created manually. Any of 18 types of HIPAA identifiers,[10] as well as practitioner and institution names, were manually annotated following written annotation guidelines (see online appendix B). Practitioner and institution names were included because some require their removal for external data sharing. Documents were reviewed in batches of at least five documents. Multiple abstractors reviewed each batch successively and independently. Each abstractor reviewed a batch one time only. A batch of documents was considered de-identified when two successive reviewers found no new instances of identifiers in any of its documents. Calculating inter-annotator agreement was not possible because each abstractor in the series (except the last two) reviewed different versions of the documents.

Original PHI	**Redacted <PHI> & Leaked PHI	Surrogate PHI & Hiding PHI
Smith, 61 yo ...	**pt_name<A>, **age<60s> yo ...	Jones, 64 yo ...
daughter, Lynn, to ...	daughter, Lynn, to ...	daughter, Lynn, for ...
oncologist Dr. White ...	oncologist Dr. **MD_name<C> ...	oncologist Dr. Howe ...
5/13/10 to consider ...	**date<5/28/10> to consider ...	5/28/10 to consider ...
SWOG protocol 1811, ...	SWOG protocol **other_id, ...	SWOG protocol 1798, ...
was randomized 5/10 ...	was randomized 5/10 ...	was randomized 5/10 ...
to call Mr. Smith on ...	to call Mr. **pt_name<A> on ...	to call Mr. Jones on ...
PLAN:Dr White and I ...	PLAN:Dr White and I ...	PLAN:Dr White and I ...

Figure 1 Illustration of original PHI, leaked PHI, and hiding PHI in clinical text.

Automated natural language de-identification

A priori, we surmised that a reasonable test of the obfuscation efficacy of the HIPS approach would require using a real-world automated de-identification system that replaced at least half the identifiers with surrogates (ie, recall ≥0.5), in a corpus large enough to contain at least 10 instances of residual identifiers of each type being investigated (eg, patient names, dates, phone numbers). This would, we believe, ensure ample opportunity for HIPS surrogates to have an obfuscation effect, and would enable calculation of detection rates with reasonable precision. However, given the low frequency of naturally occurring identifiers in actual clinical text, and given that high-performing de-identification systems have residual PHI exposure rates in the 2–6% range (ie, recall of 94–98%),[17 27] it was not feasible for us to conduct these pilot experiments at reasonable cost under completely natural conditions. A completely natural experiment would require test corpora containing 170–>3000 documents, respectively, for detection testing of patient names and phone numbers (assuming use of a de-identification system with 95% recall, and identifier frequencies comparable with those in the present test corpora). An additional 300–400 documents would be needed to train the de-identification models, according to our prior work.[12 18]

To enable pilot testing of the HIPS approach at reasonable expense, we opted to create test corpora with artificially inflated quantities of residual identifiers. We did this by undertraining an otherwise high-performing automated de-identification system. Specifically, we used training sets that were too small (<100 documents) and of mismatched document type, thereby yielding residual rates several times higher than those expected with a properly trained system. We implemented this approach using the MITRE Identification Scrubber Toolkit (MIST) version 1.2.[12] MIST incorporates an identifier prediction engine based on a machine-learning framework that learns from manually annotated training documents (a training set) to detect identifiers in new documents (a test set). An earlier version of MIST was the highest-performing automated system in the Informatics for Integrating Biology and the Bedside (i2b2) de-identification challenge, achieving recall (defined above) and precision (defined as the number of correctly predicted identifiers divided by the number of predictions) of 0.96–0.98 and 0.98–0.99, respectively.[17 27] The version of MIST used in these experiments (version 1.2) differs from the earlier version by omitting some features that were derived from regular expressions specifically tailored to the i2b2 de-identification dataset.

We experimented training MIST with corpora of various sizes (60–100 documents) and document composition—oncology notes, family practice notes, pathology reports, and mixtures thereof. Despite several attempts, we were unable to produce de-identification models that achieved desired levels of recall and residual counts for all types of identifiers. For example, models that met our minimum criteria for patient names, either underperformed or overperformed for other identifier types. Since patient names present the greatest risk of disclosure,[20] we therefore selected de-identification models that allowed us to evaluate HIPS' ability to conceal (1) the most commonly occurring PHI types in the high-PHI-density corpus, and (2) patient names and dates, which were among the most sensitive PHI types found in our average-density clinical corpus.

For the high-density identifier experiment, we trained MIST on a randomly selected set of 100 oncology notes. The remaining 31 notes were used as the test set. The types and quantities of identifiers are summarized in table 1.

For the average-density identifier experiment, we trained MIST on a randomly selected set of 90 documents which included 30 oncology notes, 30 family practice notes, and 30 pathology reports. We randomly selected 50 of the remaining family practice notes to use as the test set. As shown in table 2, this corpus contained 343 HIPAA identifiers and 116 other identifiers, but only patient names and dates met the criteria for reasonable HIPS evaluation.

Generating HIPS surrogates

HIPS surrogates were generated in both corpora using MIST's built-in 'clear-to-clear' redaction option without modification (details are in online supplement appendix C). Briefly, clear-to-clear redaction replaces tagged identifiers with system-generated fictitious content that resembles real identifiers. For example, a patient's name, such as 'Ms. Holli Larsen' may be replaced with the name 'Ms. Roxanne Sutherland', and an encounter date of '29 July 2010' may be replaced with a surrogate of '8 August 2010' based on a randomly generated date offset applied consistently throughout the document. Proper name formatting is preserved as much as possible, including use of honorifics (eg,

Table 1 Automated de-identification system performance by identifier type in the high-density identifier corpus consisting of 31 oncology progress notes (15 512 words)

Identifier type A	Instances in the corpus B	Instances replaced by the system C	System recall* D	Residual identifier instances in corpus E	Reasonable opportunity to test HIPS?† F
HIPAA PHI					
Pat. name	35	29	83%	6	Yes
Age	86	79	92%	7	Yes
Phone #	2	0	0%	2	No
Address	6	4	67%	2	No
Date	180	163	91%	17	Yes
MRN	3	0	NA	3	No
Acct. #	1	0	NA	1	No
Other ID #s	10	1	NA	9	No
Subtotal	323	276	85%	47	
OTHER PHI					
MD name	82	73	89%	9	Yes
Org. name	27	7	26%	20	No
Subtotal	109	80	73%	29	

*A suboptimal training set was used to degrade system recall, thereby increasing residual PHI for experimental purposes.
†We defined a reasonable opportunity to test the HIPS approach as recall (col. D) ≥∼0.5 and N residual PHI instances in the corpus (col. E) ≥∼10.

Table 2 Automated de-identification system performance by identifier type in the average-density identifier corpus consisting of 50 family practice progress notes (22 525 words)

PHI type A	N PHI instances in corpus B	N PHI instances replaced by system C	System PHI recall* D	N residual PHI instance in corpus E	Reasonable opportunity to test HIPS?† F
HIPAA PHI					
Pat. name	59	27	46%	32	Yes
Age	50	50	100%	0	No
Phone #	3	3	100%	0	No
Address	3	0	0%	3	No
Date	228	194	85%	34	Yes
MRN	0	0	NA	0	No
Acct. #	0	0	NA	0	No
Other ID #s	0	0	NA	0	No
ALL HIPAA	343	274	80%	69	
OTHER PHI					
MD name	53	4	8%	49	No
Org. name	63	2	3%	61	No
ALL OTHER	116	6	5%	110	

*A suboptimal training set was used to degrade system recall, thereby increasing residual PHI for experimental purposes.
†Criteria for inclusion in the detection experiment were system recall (col. D) $\geq \sim 0.5$ and N residual instances (col. E) $\geq \sim 10$.

Ms., Dr.), middle initials, and capitalization. Name surrogates are drawn from the public US Bureau of the Census name files,[28] matching on first-name gender for gender-specific names, and gender-ambiguous names otherwise. Location information is replaced using a strategy that identifies the components of location (street address, city, state, zip) and generates consistent replacements. A built-in option accepts user-supplied lists of institution names (such as healthcare facilities), from which MIST randomly selects its surrogates. Accordingly, we supplied a list of the top 200 local institution names referenced in Group Health external claims. Date surrogate generation applies random offsets to all dates, and accommodates many date formats. Surrogates for numeric and alphanumeric identifiers, such as accession, medical record, and phone numbers are created by randomly replacing some parts of the identifier (eg, the last four digits of a phone number) preserving the original format.

Experimental design

Human detection experiments

Two human detection experiments were conducted with the approval of the Group Health IRB. One experiment involved a corpus of high-density identifiers, and the other a corpus of average-density identifiers. Paper copies of the sets of documents were reviewed independently by two reviewers. Reviewers read each note twice. In the first reading, they marked all identifiers regardless of whether they thought they were real or surrogates. In the second reading, they considered each marked instance and selected those they predicted were actual residual identifiers, and recorded the reasoning supporting their predictions. One of the authors (DC) then compared each prediction with the reference standard and tallied correct and incorrect predictions by reviewer and by identifier type.

Three reviewers participated in the experiments. Reviewers #1 and #2 were trained chart abstractors with 4 and 3 years experience, respectively. Reviewer #3 was an informaticist familiar with natural language processing and automated de-identification methods. Reviewers #1 and #2 participated in the high-density experiment; reviewers #1 and #3 participated in the average-density experiment.

We hypothesized that the reviewers would be unable to distinguish residual (leaked) identifiers from HIPS surrogates in the experimental corpora. Specifically, we hypothesized the reviewers' recall for detecting residual PHI would approach zero, and that the precision of their guesses (individually and combined) would be less than the precision expected by chance.

Evaluation

The numbers of correct and incorrect predictions by each reviewer were used to calculate recall (equation 1) and precision (equation 2), by abstractor and by identifier type:

$$Recall = (correctly\ predicted\ residual\ identifiers)/(all\ residual\ identifiers) \quad (1)$$

$$Precision = (\text{correctly predicted residual identifiers})/(all\ predicted\ identifiers) \quad (2)$$

We also calculated the joint performance of both abstractors combined, using the union of their independent predictions (unduplicated). To serve as a baseline against which to evaluate the accuracy of the reviewers' predictions, we calculated the rate of precision expected by chance (equation 3):

$$Precision\ Expected\ by\ Chance = (all\ residual\ identifiers)/(all\ identifiers) \quad (3)$$

If the precision of a reviewer's predictions exceeded the precision expected by chance, we considered this evidence that the reviewer was successful in detecting residual identifiers.

RESULTS

High-density experiment

Results of the high-density detection experiment are shown in table 3; there were a total of 47 instances of residual PHI (Table 3, column C). Together, the two reviewers predicted there were 69 instances of residual identifiers in the corpus. Six of the 69 predictions were correct, for a recall rate of 0.13 (table 3 column O). Eighty-seven per cent of residual identifiers were not detected by either reviewer. The overall precision was 0.09, which was less than the precision expected by chance of 0.15 (table 3 column P).

Table 3 Results of the high-density identifier (oncology notes) human detection experiment by identifier type and reviewer

Test corpus				Reviewer #1 (abstractor)				Reviewer #2 (abstractor)				Both reviewers combined*			
Identifier type A	PHI instances B	Residual PHI C	Expected precision† D	Predictions E	Correct F	Recall G	Precis. H	Predictions I	Correct J	Recall K	Precis. L	Predictions M	Correct N	Recall O	Precis. P
HIPAA															
Pat. name	35	6	0.17	0	0	0.00	—	12	4	0.67	0.33	12	4	0.67	0.33
Age	86	7	0.08	5	0	0.00	0.00	12	0	0.00	0.00	17	0	0.00	0.00
Phone #	2	2	1.00	0	0	0.00	—	1	1	0.50	1.00	1	1	0.50	1.00
Address	6	2	0.33	1	0	0.00	0.00	0	0	0.00	—	1	0	0.00	0.00
Date	180	17	0.09	1	0	0.00	0.00	35	1	0.06	0.03	36	1	0.06	0.03
MRN	3	3	1.00	0	0	0.00	—	0	0	0.00	—	0	0	0.00	—
Acct. #	1	1	1.00	0	0	0.00	—	0	0	0.00	—	0	0	0.00	—
Other ID #s	10	9	0.90	0	0	0.00	—	2	0	0.00	0.00	2	0	0.00	0.00
ALL	323	47	0.15	7	0	0.00	0.00	62	6	0.13	0.10	69	6	0.13	0.09
OTHER															
Prac name	82	9	0.11	5	4	0.44	0.80	8	4	0.44	0.50	12	7	0.78	0.58
Org. name	27	20	0.74	8	6	0	0.75	3	1	0	0.33	10	7	0.35	0.70
ALL	109	29	0.27	13	10	0	0.77	11	5	0.17	0.45	22	14	0.48	0.64

*Based on unduplicated count of N predictions and N correct across the two reviewers.
†Defined as the number of residual PHI instances (col. C) divided by the total number of PHI instances (col. B).

The efficacy of HIPS varied according to PHI type. While reviewers were unable to detect residual dates or ages with precision greater than expected by chance, they correctly guessed 4 of 12 patient names, with a precision of 0.33 compared with the expected precision of 0.17. All four detected names (two first names and two first and last names) were family members of the same patient, mentioned in the Social History section of a progress note that also indicated the family had emigrated from Africa. Contextual information in the document suggested the patient and family members had African names. However, the patient's first and last names had been replaced with European surrogates. The reviewer correctly reasoned that the inconsistency between European and African names was caused by incomplete de-identification. The same reviewer also correctly predicted one of two residual phone numbers.

Reviewers predicted other (non-HIPAA) identifiers with mixed success. They were unable to predict organization names with precision (0.70) greater than that expected by chance (0.74). The corresponding recall rate of 0.35 indicates nearly two-thirds of residual organization names were not detected. However, the two reviewers correctly identified seven of nine residual practitioner names (out of 12 predictions) achieving a precision of 0.58, which greatly exceeded the expected precision of 0.11. Reasons given for all correctly identified practitioner names were the same: the names matched those of actual practitioners with whom the reviewers were familiar, based on extensive chart abstraction experience. The same reason was given, however, for one erroneous prediction.

There were substantial differences between Reviewer #1 and Reviewer #2 in the number of predictions issued. Reviewer #1 made 7 and 13 predictions among HIPAA and other identifiers, respectively, compared with 69 and 22 predictions made by Reviewer #2. Recall and precision rates were similar across the two reviewers, though our sample size was too small to perform a statistical test for differences between reviewers.

Average-density experiment

Results of the average-density detection experiment are shown in table 4. Together, the two reviewers predicted that 14 identifiers were residual, with six correct predictions out of 66 possible, yielding a recall rate of 0.09 (table 4 column O). Ninety-one per cent of residual identifiers were not detected. The overall precision of reviewers' predictions was 0.43. This was higher than the overall precision of 0.23 expected by chance (table 4 column D), and failed to confirm our hypothesis, an issue we address in the Discussion section.

Results varied by reviewer. Reviewer #3 issued more predictions and with greater precision than Reviewer #1. Reviewer #3 correctly predicted patient names with a precision of 0.80, which was well above the rate expected by chance (0.54), indicating imperfect but real disclosure. Reviewer #1 correctly predicted one of two patient names with a precision of 0.50, which was approximately equal to the expected precision. Two of the detected names appeared as salutations in letters to

Table 4 Results of the average-density identifier (family practice notes) human detection experiment by identifier type and reviewer

Test corpus				Reviewer #1 (abstractor)				Reviewer #3 (informaticist)				Both Reviewers Combined*			
PHI type A	N PHI instances B	N residual PHI C	Expected precision† D	Predictions E	Correct F	Recall G	Precision H	Predictions I	Correct J	Recall K	Precision L	N predictions M	N correct N	Recall O	Precis. P
HIPAA															
Pat. name	59	32	0.54	2	1	0.03	0.50	5	4	0.13	0.80	6	4	0.13	0.67
Date	228	34	0.15	0	0	0.00	0.00	8	2	0.06	0.25	8	2	0.06	0.25
ALL	287	66	0.23	2	1	0.02	0.50	13	6	0.09	0.46	14	6	0.09	0.43

*Unduplicated count of N predictions and N correct across the two reviewers.
†Defined as the number of residual PHI instances (col. C) divided by the total number of PHI instances (col. B).

patients that had been copied into the progress notes. MIST replaced patient names elsewhere in the notes, but not in the letter text. Reviewer #3 correctly predicted that both salutations contained residual names, and Reviewer #1 correctly predicted one of them. The third name correctly predicted was the first name of a person mentioned in an emergency room transcript that had been pasted into a progress note. The progress note containing the fourth correctly predicted name was for a patient with East Indian first and last names. The note mentioned that the patient was East Indian. MIST replaced the patient's last name with a European-sounding surrogate, but did not replace the first name, which the reviewer correctly reasoned was residual.

Two dates were correctly predicted by Reviewer #3, achieving a precision of 0.25, which was above the 0.15 rate expected by chance. One of the dates was an historical reference to a clinical event in 'Sep 2010'. The reviewer assumed MIST could not produce a date in this format. This assumption was false, but the prediction was correct in this instance. Elsewhere, MIST generated date surrogates using this same format, four instances of which Reviewer #3 incorrectly predicted were residual. In two other incorrect date predictions, irregular formatting was reported as the (faulty) reasoning. The other correct date prediction involved a reference to a future event on '6 March 2011', which was not redacted in a progress note for a February encounter. Elsewhere in this progress note, MIST used a randomly generated offset of +91 days to replace two mentions of the encounter date with the surrogate '15 May 2011'. The May surrogates rendered illogical the reference to March as a future date, which Reviewer #3 correctly reasoned was residual.

DISCUSSION

Summary of findings

While this pilot study is too small to provide definitive evidence, its findings suggest that our hypothesis about the efficacy of the HIPS approach merits more thorough evaluation. In clinical text de-identified using a simple implementation of HIPS, 87–91% of residual identifiers were successfully concealed from human readers charged with finding them. If this efficacy rate is generalizable to high-performing de-identification systems achieving recall rates of 94–98%, and in larger corpora of more diverse clinical and institutional origin, *HIPS could increase the effective de-identification rates of such systems to levels above 99%*. For example, if a de-identification system replaced 95% of identifiers with surrogates and left 5% exposed, and if HIPS surrogates concealed 90% of the residual identifiers, then the effective de-identification rate would be $0.95+(0.05\times(0.90))$ or 99.5%. Further, had our reviewers been attackers intent on re-identifying patients, their attack would have been hobbled by the fact that 90% of the clues they might have used to deduce patient identity were misleading. HIPS may thus enhance privacy by raising the cost of re-identification relative to traditional redaction approaches.

These experiments suggest that the efficacy of HIPS surrogates depends, in part, on contextual information, which includes content found elsewhere in a document and a reader's prior knowledge of the institutional setting from which a document comes. Contextual clues may help distinguish synthetic names from 'leaked' names, as illustrated in our experiments by person names with discordant national origins. Similarly, a reader's familiarity with a corpus' institutional setting may make it more challenging to devise effective surrogates for 'leaked' practitioner names (also illustrated in our experiments). It is thus important to consider both, the audience and the purpose, for which a corpus is being de-identified. Different surrogate generation strategies may be required for different audiences (eg, practitioners, informaticists, internal vs external readers) and purposes (eg, concealing whether patients were treated at a particular institution vs concealing patients' individual health conditions). HIPS may be most robust in external settings where detailed knowledge of the healthcare system producing the documents is limited.

Limitations

We wish to highlight several limitations of this pilot study, which can serve as guideposts for extension. First, the experimental setting was highly specialized and underpowered. The evaluation, for instance, was limited to small samples of one type of text (progress notes) from two medical specialties. Additional, larger-scale investigations are needed, applying high-performing de-identification methods to multiple note types, to determine whether residual identifiers can be effectively concealed by the HIPS approach. Such residual identifiers are likely not to be random, and may have characteristics that make concealing them more challenging. Moreover, our reviewers were local personnel and did not address all types of potential users. We believe such deficiencies can be addressed by replicating our study on naturally occurring clinical text from multiple institutions in a larger-scale experiment with other types of readers.

Second, our experiments relied on the same de-identification system to create test corpora. Other systems, manual and automated, may produce different results.

Third, we did not subject HIPS to machine-based detection attacks. It is possible that statistical approaches could be used to distinguish surrogates from residual identifiers at rates different from those detected by humans.

Next steps

While our experimental results are encouraging, they also suggest areas where further improvements are needed. One potential improvement would entail accounting for the national origins of person names when generating surrogates. Another strategy that may enhance obfuscation efficacy is to introduce natural-appearing spelling or formatting errors. This is based on the observation that the reviewers leveraged formatting clues to help them detect some residual dates.

We also acknowledge that these experiments did not provide an adequate evaluation of HIPS' efficacy for concealing phone numbers, addresses, medical record numbers, account numbers, and other ID numbers of which there were 22 in the high-density experiment (column F in tables 1 and 2). However, had traditional redaction methods been applied, all 22 of the residual identifiers would have been disclosed. Rigorous evaluation of HIPS' efficacy for these identifier types remains to be done. However, because automated systems can generally detect such identifiers with high recall, and because surrogate generation for alphanumeric strings is generally straightforward, we expect HIPS to perform well in this area.

An issue we did not address experimentally, but which should be explored in future research, is the importance of tailoring surrogate generation techniques to specific de-identification objectives. Devising surrogate practitioner names is illustrative. Choosing surrogates from the universe of local practitioner names may be effective if the de-identification goal is to conceal the fact that a particular doctor is associated with a particular

episode of care. On the other hand, if the de-identification goals were to provide anonymity for the institution providing the corpus, surrogate practitioner names would need to be drawn from a more generic universe. The extent to which hybrid surrogate generation strategies may achieve multiple de-identification objectives should be investigated.

CONCLUSION

This paper proposed a HIPS approach to obfuscate residual identifiers in de-identified natural language clinical documents. Our study, with real readers of real clinical text, suggests that HIPS has the potential to significantly reduce detection of residual identifiers. If the results of our study hold when used in conjunction with high-performing de-identification systems, HIPS may yield a 10-fold reduction in the risk of disclosing residual identifiers in otherwise de-identified clinical text compared with traditional redaction approaches. Such a reduction would yield effective de-identification rates well above 99%, far surpassing efficacy levels attained by manual methods and the best automated systems evaluated in competition. In addition, our research suggests several intuitive improvements to surrogate identifier generation that may improve obfuscation of person names.

Contributors LH, JA, SB, CC, and BW conceived of the concept of using realistic surrogates to improve de-identification efficacy and developed the MIST system. SB developed the software that implemented the surrogate generation system. BM provided expertise in privacy and security issues. All authors contributed to the design of the human detection experiments. DC implemented the experiments. All authors interpreted the results of the experiments and contributed to writing the manuscript.

Funding This manuscript was made possible, in part, by funding from the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology, U01HG006385 from the National Institutes of Health, the Electronic Medical Records and Genomics (eMERGE) consortium, and CCF-0424422 from the National Science Foundation. The contents of the manuscript/abstract are solely the responsibility of the authors.

Competing interests None.

Ethics approval This study was conducted with the approval of the Group Health Human Subjects Review Committee, who considered the two chart abstractors and one informaticist who participated in the human detection experiments to be human subjects.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Chapman WW,** Nadkarni PM, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;**18**:540—3.
2. **Elkin PL,** Froehling DA, Wahner-Roedler DL, *et al*. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Ann Intern Med* 2012;**156**:11—18.
3. **Matheny ME,** Fitzhenry F, Speroff T, *et al*. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. *Int J Med Inform* 2012;**81**:143—56.
4. **South BR,** Chapman WW, Delisle S, *et al*. Optimizing A syndromic surveillance text classifier for influenza-like illness: does document source matter? *AMIA Annu Symp Proc* 2008:692—6.
5. **Jiang M,** Chen Y, Liu M, *et al*. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc* 2011;**18**:601—6.
6. **Peissig P,** Rasmussen L, Berg R, *et al*. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc* 2012;**19**:225—34.
7. **Wilke RA,** Berg RL, Peissig P, *et al*. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin Med Res* 2007;**5**:1—7.
8. **Chapman WW,** Cohen KB. Current issues in biomedical text mining and natural language processing. *J Biomed Inform* 2009;**42**:757—9.
9. **Demner-Fushman D,** Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;**42**:760—72.
10. **U.S.** Department of Health and Human Services. *Standards for Privacy of Individually Identifiable Health Information; Final Rule*. Federal Register, 2002:53181—273.
11. **DE-ID Data Corp.** *DE_ID Health Data Safety Software*. 2012. http://www.de-idata.com/ (accessed 7 Apr 2012).
12. **Aberdeen J,** Bayer S, Yeniterzi R, *et al*. The MITRE identification Scrubber Toolkit: design, training, and assessment. *Int J Med Inform* 2010;**79**:849—59.
13. **Gardner J,** Xiong L, Lu J, eds. HIDE: Heterogeneous Information DE-identification. In: *Proceedings of the 12th International Conference on Extending Database Technology (EDBT)*; March 2009, St. Petersburg, Russia. New York: ACM, 2009.
14. **Sweeney L.** Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp* 1996:333—7.
15. **Meystre SM,** Friedlin FJ, South BR, *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol* 2010;**10**:70.
16. **Szarvas G,** Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc* 2007;**14**:574—80.
17. **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.
18. **Yeniterzi R,** Aberdeen J, Bayer S, *et al*. Effects of personal identifier resynthesis on clinical text de-identification. *J Am Med Inform Assoc* 2010;**17**:159—68.
19. **Taira RK,** Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp* 2002:757—61.
20. **Neamatullah I,** Douglass MM, Lehman LW, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008;**8**:32.
21. **Mayer J,** Shen S, South BR, *et al*. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes. *AMIA Annu Symp Proc* 2009:416—20.
22. **Dorr DA,** Phillips WF, Phansalkar S, *et al*. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods Inf Med* 2006;**45**:246—52.
23. **Gardner J,** Xiong L. An integrated framework for de-identifying unstructured medical data. *Data Knowl Eng* 2009;**68**:1441—51.
24. **Friedlin F,** McDonald C. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* 2008;**15**:601—10.
25. **Hirschman L,** Aberdeen J, eds. *Measuring Risk and Information Preservation: Toward New Metrics for De-identification of Clinical Texts. NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*. Los Angeles, California: Association for Computational Linguistics, 2010.
26. **Kirk SR,** Jenkins S. Information theory-based software metrics and obfuscation. *J Syst Software* 2004;**72**:179—86.
27. **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.
28. **U.S. Census.** *Frequently Occurring First Names and Surnames From the 1990 Census*. 2011. http://www.census.gov/genealogy/names/ (accessed 24 Sep 2011).