Journal of the American Medical Informatics Association, 27(9), 2020, 1374–1382
doi: 10.1093/jamia/ocaa095
Advance Access Publication Date: 15 September 2020
Research and Applications

Research and Applications

Resilience of clinical text de-identified with "hiding in plain sight" to hostile reidentification attacks by human readers

David S. Carrell,[1] Bradley A. Malin,[2] David J. Cronkite,[1] John S. Aberdeen,[3] Cheryl Clark,[3] Muqun (Rachel) Li,[4] Dikshya Bastakoty,[2] Steve Nyemba,[2] and Lynette Hirschman[3]

[1]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA, [2]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA, [3]Human Language Technology, MITRE Corporation, Bedford, Massachusetts, USA, and [4]Privacy Analytics Inc, Nashville, Tennessee, USA

Corresponding Author: David S. Carrell, PhD, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA; david.s.carrell@kp.org

Received 2 January 2020; Revised 2 April 2020; Editorial Decision 1 May 2020; Accepted 26 May 2020

ABSTRACT

Objective: Effective, scalable de-identification of personally identifying information (PII) for information-rich clinical text is critical to support secondary use, but no method is 100% effective. The hiding-in-plain-sight (HIPS) approach attempts to solve this "residual PII problem." HIPS replaces PII tagged by a de-identification system with realistic but fictitious (resynthesized) content, making it harder to detect remaining unredacted PII.

Materials and Methods: Using 2000 representative clinical documents from 2 healthcare settings (4000 total), we used a novel method to generate 2 de-identified 100-document corpora (200 documents total) in which PII tagged by a typical automated machine-learned tagger was replaced by HIPS-resynthesized content. Four readers conducted aggressive reidentification attacks to isolate leaked PII: 2 readers from within the originating institution and 2 external readers.

Results: Overall, mean recall of leaked PII was 26.8% and mean precision was 37.2%. Mean recall was 9% (mean precision = 37%) for patient ages, 32% (mean precision = 26%) for dates, 25% (mean precision = 37%) for doctor names, 45% (mean precision = 55%) for organization names, and 23% (mean precision = 57%) for patient names. Recall was 32% (precision = 40%) for internal and 22% (precision =33%) for external readers.

Discussion and Conclusions: Approximately 70% of leaked PII "hiding" in a corpus de-identified with HIPS resynthesis is resilient to detection by human readers in a realistic, aggressive reidentification attack scenario—more than double the rate reported in previous studies but less than the rate reported for an attack assisted by machine learning methods.

Key words: de-identification, privacy, confidentiality, electronic health records, natural language processing, biomedical research

INTRODUCTION

The increasing importance of clinical natural language processing in support of epidemiology and outcomes research,[1–5] clinical trials,[6,7] and medical products safety surveillance[8,9] is accelerating demand for access to patient clinical notes for secondary use. But information-rich, free-text chart notes also harbor sensitive personally identifiable infor-

A. Raw text with PII	POST OP: 1/14/2013, John Doe, MRN 19570319: An 84-year-old male
B. Traditional redaction	POST OP: **<DATE>**, **<PT-NAME>**, MRN 19570319: An **<AGE>**-year-old male
C. *HIPS resynthesis*	POST OP: *12.20.12*, *Bob Smith*, MRN 19570319: An *81*-year-old male

Figure 1. An illustration of personally identifiable information (PII) in clinical text (A) in its original form, (B) with traditional redaction of PII tagged by a de-identification system that overlooked the identifier "19570319," and (C) with HIPS resynthesis of PII tagged by the de-identification system, allowing the overlooked identifier "19570319" to hide in plain sight.

mation (PII). The moral[10] and legal[8] imperatives to preserve patient privacy and the limitations of manual de-identification[11] have spurred development of scalable, automated text de-identification technologies,[12] but even the best-performing methods are imperfect, overlooking 1%-6% of patient identifiers that should be concealed.[13–15]

Incremental improvements in de-identification technologies may shrink—but are not likely to eliminate entirely—the small fraction of identifiers that escape detection and redaction. One response to this "residual PII problem"[14] is the hiding-in-plain-sight (HIPS) approach.[16] Instead of redacting the identifiers, a de-identification system successfully detects by replacing them with nondescript symbols (eg, "*******") or placeholders (eg, "<PT-NAME>"), whereas the HIPS approach replaces all detected identifiers with realistic but fictitious (ie, resynthesized) content. The presence of this resynthesized content allows leaked PII to blend in—at least for readers willing to forego scrutinizing a corpus for the purpose of distinguishing leaked from resynthesized PII. HIPS resynthesis allows residual PII to hide in plain sight, as illustrated in Figure 1.

With traditional redaction of known PII, all or nearly all[11] leaked PII can be recognized upon reading (Figure 1B). If known PII is replaced with resynthesized content, leaked content appears unremarkable (Figure 1C). To a nonmalign reader, not interested in expending significant effort to detect leaked PII, the HIPS solution resolves the residual PII problem.

But a de-identification solution that depends on user behavior can break down. Prudence requires data stewards to consider such breakdowns and assess the corresponding reidentification risks.[17] To inform stakeholder decision making, to date, 3 studies have investigated reidentification risks for HIPS-resynthesized corpora.[14,18,19] Two studies using small corpora "attacked" by readers internal to the institution producing the corpus suggest that HIPS is quite resilient to human reader attacks, but these studies acknowledged needs for more robust evaluation.[14,19] In the third study, it was shown that an attack incorporating machine learning could re-expose two-thirds of leaked PII in a HIPS corpus.[18,20]

This study addresses gaps in existing knowledge by investigating the vulnerability of HIPS-resynthesized corpora to malicious human reader attacks using larger, more representative corpora, from 2 diverse institutions, subjected to a more realistic reidentification attack by both internal and external readers. We hypothesized that a realistic, aggressive attack wherein readers could use information appearing anywhere in the corpus (rather than just the immediate document), unrestricted Internet search, and information available within their respective healthcare organizations, would expose more PII leaks than previously reported, but still at lower detection rates than with traditional redaction, in which 100% of leaked PII is detectable upon examination of the corpus.

Clinical text de-identification is the process of removing identifying information from a corpus for the purpose of facilitating is use for purposes other than patient care (ie, secondary use). Identifiers removed include the 18 PII types enumerated in the Safe Harbor de-identification model of the Privacy Rule of the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA).[8] Several scalable, automated de-identification systems have been developed for this purpose,[13,15,21–26] but numerous studies have shown that it is virtually impossible for these systems to remove more than about 94%-99% of PII when applied at scale,[13,15,22,27–37] without sacrificing clinically useful information through over-redaction. So vexing is the residual PII problem that some researchers advocate for severe over-redaction to facilitate corpus sharing.[38]

We believe that scalable systems are unlikely to achieve 100% PII removal without noticeably sacrificing a corpus' utility, making HIPS resynthesis a potentially viable approach for information-preserving de-identification.

The HIPS approach aims to enhance de-identification by concealing both detected PII (PII tagged by a de-identification system) and undetected PII (PII in a document the system overlooks). Aware that the vast majority of PII-like content in a corpus is fictitious resynthesized content, a nonmalign reader will read documents assuming that all PII has been concealed.[39] In a previous study based on a small corpus we reported that up to 90% of leaked PII may escape detection by readers attempting to expose it.[14]

Evaluating the scenario in which a release corpus (a set of de-identified documents made available for secondary use) is subjected to a malicious reidentification attack requires clarity regarding which PII is actually at risk of detection and which is not.[18] Both traditional redaction and HIPS resynthesis attempt to mitigate exposure risk for detected PII instances tagged by a de-identification system. PII at risk of detection in a HIPS corpus is limited to that which the de-identification system failed to tag and resynthesize. Resynthesized content is randomly generated and therefore cannot be reverse engineered to expose the original PII. In the extreme case, in which a malicious attack isolated all actual leaks in a HIPS corpus, the exposure would be no worse than that in traditional redaction, in which all leaks are readily detected.

The objective of the present study is to investigate the extent to which the added protection of HIPS resynthesis may be undone by human readers in a realistically aggressive attack scenario.

MATERIALS AND METHODS

In these experiments, 2 readers from the originating site and 2 external readers attempted to identify leaked PII in 2 corpora—1 originating from Kaiser Permanente Washington (KPWA) and 1 from Vanderbilt University Medical Center (VUMC). These corpora were deidentified using HIPS resynthesis.

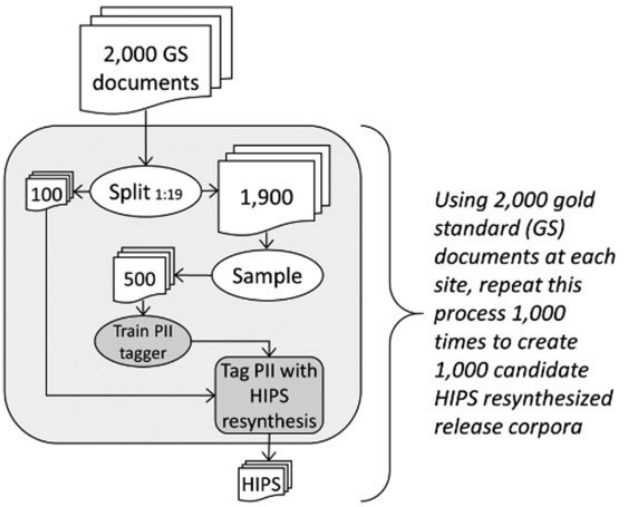

Figure 2. Process for creating 1000 candidate hiding in plain sight (HIPS) resynthesized release corpora based on 2000 gold standard (GS) annotated documents at each study site. PII: personally identifying information.

Creating a study release corpus at each site

A corpus suitable for use in conducting a reidentification attack experiment should be as realistic as possible (for generalizability) and of a size that facilitates conducting the experiment within time and cost constraints. As described subsequently, the experiment required multiple readers to rigorously scrutinize the entire corpus multiple times—a labor-intensive process that drives the cost of conducting such experiments. As noted, the highest-performing de-identification systems—including those used at the institutions participating in this experiment—yield corpora with very low (1%-6%) leak rates.[13,15,22,27–37] This creates an experimental design dilemma. A study corpus with state-of-the-art (and therefore very low) leak rates would have to be very large to contain enough leaked PII to allow stable estimation of leak detection rates. To illustrate, if 80% of documents contained a patient name and 5% of these patient names were leaked, a study corpus of 600 documents would be needed to yield 24 leaked names for possible detection experimentally ($600 \times 0.80 \times 0.05 = 24$). Study corpora including several hundred documents were infeasible given our budgetary constraints. We considered corpora of 100 documents each to be a reasonable compromise for these experiments. We then devised a novel method to create 100-document experimental corpora at each site that: (1) included broadly representative clinical documents, (2) were de-identified by a fully automated machine-learned system, and (3) contained enough leaked PII to facilitate stable estimation of vulnerability to reidentification attack. We included non-HIPAA institutional identifiers (clinician names and healthcare organization names) because our 2 institutions routinely require their redaction.

Specifically, we randomly selected 2000 outpatient encounter notes at KPWA, and 2000 history and physical assessments and hospital discharge summaries at VUMC, then created gold standard PII annotations in both corpora following a detailed protocol (Supplementary Appendix). We trained separate PII taggers for KPWA and VUMC corpora because training on documents like those being de-identified is an established best practice.[40] As illustrated in Figure 2, we used each site's 2000-document corpus to create 1000 candidate release corpora at each site, in which each candidate consisted of 100 documents randomly selected from the 2000-document corpus, de-identified using the MITRE Identification Scrubber Toolkit (MIST) version 2.0[41] trained on a nonoverlapping set of 500 documents randomly selected from the remaining 1900 documents. An earlier version of MIST was the highest-performing automated system in the Informatics for Integrating Biology and the Bedside (i2b2) de-identification challenge.[22,32]

We selected 500 as the optimal training set size after empirically assessing PII recall rates for alternative PII taggers trained on 200, 300, 400, 500, 700, and 1000 documents each. We judged training sets of 500 documents to be optimal for this experiment because they yielded somewhat degraded recall—thereby leaking enough PII to make the experiment feasible—while still achieving reasonably high recall (even if below state-of-the-art levels).

Because each of the 1000 candidate corpora contained a unique set of 100 documents, and each of their corresponding MIST systems was trained on a unique set of 500 documents (thereby yielding PII taggers with somewhat different performance characteristics), none of the 1000 candidate corpora contained identical sets of PII leaks, despite the fact that each of the 2000 documents appeared in approximately 50 of the 1000 candidate corpora because of random sampling.

We then summarized PII leak counts by type across the 1000 candidates at each site and selected as the site study corpus the one whose PII leak counts were closest to the median count of PII leaks for each PII type across all 1000 candidates (details in Supplementary Figure 1).

Reidentification attack by human readers

Separate reidentification attack experiments were conducted for the KPWA corpus (first) and the VUMC corpus (second). For each experiment each reader was given a binder with paper copies of the 100 HIPS-resynthesized documents for that corpus collated in the same random order. Prior to conducting their attacks, each reader received 3-4 hours of training explaining automated PII tagging, HIPS resynthesis, and rules for conducting the attack (below). In a practice session, readers practiced marking documents like those in the study corpus. During the experiment, readers marked content judged to be PII leaks by circling them with red ink and, separately, recording the document number and PII type of each leak on a separate paper tracking sheet. Following this, a research coordinator reviewed each circled instance in the study binder, compared the binder markings with the tracking sheet, resolved any discrepancies or ambiguities through discussion with the reader, then key-entered information about each marked leak into a study database. Summary data form the database were then verified against the tracking sheets and corrected as needed.

The reader attack on each corpus was conducted in 2 stages. The first stage (stage 1) was designed to reflect a real-world attack in which attackers may reasonably assume that a corpus contains PII leaks, but the actual number of such leaks is unknown. This creates the possibility that attackers may underestimate or overestimate the actual number of leaks, thereby yielding misleading information as to which content corresponds to actual patients. In judging which content was leaked, readers were:

- allowed to use Internet search (eg, Google, Facebook, or LinkedIn searches to discover information about patients, doctors, or organizations);
- allowed to use information readily accessible in an employee's institution (eg, clinic and clinician directories);
- allowed to use information appearing anywhere in the study corpus;

- allowed to use personal knowledge (eg, awareness of which local employers offered healthcare coverage at their institution);
- prohibited to use EHR and clinical databases because access to this information would not be available to most attackers and would violate institutional privacy rules; and
- prohibited from discussing the experiment with other readers or employees because this may introduce contamination across individual attackers.

Results from stage 1 were used to estimate vulnerability to detection of leaked PII under realistic conditions by individual attackers. We hypothesized the following:

> Hypothesis 1 (H1): Human reader leak detection rates will exceed previously published rates.
>
> Rationale: The present experiment permitted attackers to use more information to identify potential leaks than previous experiments.
>
> H2: Leak detection rates in internal corpora will exceed rates in external corpora.
>
> Rationale: Insider familiarity may help readers isolate actual leaks.

We designed stage 2 of the attack to reduce reader-specific differences in propensities to judge content to be leaked. We did this by revealing to readers how many PII leaks of each type were in the corpus (without revealing which content was leaked) and then requiring them to make judgments for each PII type corresponding to these counts (eg, if there were 16 leaked dates each reader had to judge 16 dates to be leaked). We hypothesized the following:

> H3: Providing information about the number of leaks in a corpus (stage 2) will increase leak detection recall (vs stage 1) without degrading precision.
>
> Rationale: Knowledge of actual leak counts may reduce attackers' self-imposed conservatism.

Readers

Two experienced chart abstractors who routinely read and abstract real patient charts for research at their respective institutions participated as readers in the experiments. Readers had access to information inside their respective institutional firewalls accessible to all employees. When reading the corpus from their own institution, readers were considered internal readers, and external readers when reading the other institution's corpus.

Evaluation metrics

We assessed readers' abilities to isolate leaked PII using standard measures of recall (equation 1) precision (equation 2), and F1 score (equation 3). We used F1 score to compare differences in leak detection rates across readers.

PII instances that were completely overlooked or partially tagged by MIST were considered leaks. For example, if only "E. Doe" was tagged and resynthesized for the patient name "Jane E. Doe" (yielding a resynthesized name "Jane K. Smith"), we considered this a leak because "Jane" conveyed information about the actual patient.

Equation 1: Leak detection recall

$$= \frac{\#\text{ of actual PII leaks in a corpus the reader judged to be leaked}}{\#\text{ actual PII leaks in the corpus}}$$

Equation 2: Leak detection precision

$$= \frac{\#\text{ of actual PII leaks in a corpus the reader judged to be leaked}}{\#\text{ PII instances (resynthesized or leaked) the reader judged to be leaked}}$$

Equation 3: F1 score

$$= 2 \times \frac{(P \times R)}{(P + R)}$$

To summarize recall, precision, and F1 score metrics for each PII type individually, for various subsets of results, and overall (ie, for all PII types, readers and corpora combined), we calculated the mean, median, minimum, and maximum of each metric across the relevant set of results. For example, to calculate recall for leaked patient names, we summarized the 8 results for detection of leaked patient names produced by 4 readers in each of 2 corpora ($4 \times 2 = 8$). Similarly, to calculate recall across multiple PII types, we summarized results for each of the PII type included in the summary from each of 4 readers in 2 corpora. Summaries for subgroups of results and the overall summaries are of interest to some audiences and are useful for comparison with previously published work. However, we believe that the most granular summaries—those for 1 reader detecting a specific leak type in 1 corpus—are most informative because they are less likely to incorporate unrecognized bias. For example, overall detection rates can be driven by the most common PII type(s), the most easily detected leak type(s), or the highest performing reader. The operator-dependence bias in each granular result is clearly evident in its label (eg, "KPWA reader 1").

Recall and precision impact the utility of reidentified information. Higher recall combined with higher precision indicates more meaningful leak detection, as a greater number of leaks are correctly recognized as such, leading to fewer erroneous judgments that may mislead the attacker. High precision indicates meaningful leak detection at any level of recall. Low precision confounds an attack, as resynthesized content is perceived to be real, thereby misleading the attacker.

Given that secondary use of patient data for research typically occurs among known, trusted collaborators with shared commitment to preserving patient privacy,[42] we believe that the most realistic attack scenario involves a lone attacker. However, previous studies have reported pooled results simulating 2 attackers working together. We therefore report individual results as well as pooled results. For pooled results, PII instances judged to be leaked by any reader in the pool count as judgments by the pool, and correct judgments by any reader count as correct judgments for the pool.

To facilitate hypothesis testing we calculated the chance leak recall rate (equation 4)—the expected detection rate for a reader judging PII leaks completely at random (and therefore without meaningful perception of actual leaked information)—as follows:

Equation 4> Leak recall by chance (Chance R)

$$= \frac{\#\text{ of actual PII leaks in a HIPS study corpus}}{\#\text{ PII instances (resynthesized or leaked) in a HIPS study corpus}}$$

We chose this metric because it adjusts for leak detection that may be due entirely to chance and because it avoids introducing a subjective element, as would be the case if we only counted correctly detected leaks that were accompanied by a "convincing explanation" for why the reader thought the content was leaked.

When a reader's observed recall rate (equation 1) for any of the 40 observations failed to surpass the corresponding chance rate (equation 4), we concluded that reader was unable to meaningfully

Table 1. Leaked PII in the 100-document KPWA study release corpus and the 100-document VUMC study release corpus, each selected from 1000 candidate corpora generated at each study site by de-identifying randomly-selected 100-document corpora using PII tagger models trained on separate, 500-document training corpora (also randomly selected)

		PII in the HIPS-resynthesized study release corpus			
Corpus	PII type	Resynthesized	Leaked	All	Chance leak recall (equation 4)[a]
KPWA	All 5 types	747	87	834	10%
	Age	76	8	84	10%
	Date	418	16	434	4%
	Doctor name	83	20	103	19%
	Org. name	105	21	126	17%
	Patient name	65	22	87	25%
VUMC	All 5 types	1158	125	1283	10%
	Age	36	2	38	5%
	Date	498	22	520	4%
	Doctor name	351	32	383	8%
	Org. name	103	20	123	16%
	Patient name	170	49	219	22%

HIPS: hiding in plain sight; KPWA: Kaiser Permanente Washington; Org. = organization; PII: personally identifying information; VUMC: Vanderbilt University Medical Center.

[a]Chance leak recall (equation 4) is calculated by dividing the number of leaked PII instances in the HIPS study corpus by the total number of PII instances in the same corpus (resynthesized plus leaked). As described in the Materials and Methods, we used the chance leak recall rate to determine whether a reader's attempt to detect leaked PII was meaningful (ie, better than the rate expected at random).

detect that type of leak in that corpus. However, if a reader's observed recall surpassed the chance rate, we calculated a chi-square test to determine whether this rate was statistically significantly better than the chance rate. This chi-square statistic tested the null hypothesis that, for a particular PII type in a particular corpus, a reader's observed true positive, false positive, true negative, and false negative counts were no different than the corresponding counts expected by chance, conditioned on the number of judgments the reader made (Supplementary Appendix). We used Bonferroni correction to conservatively preserve a type 1 error rate of ≤ 0.05 across all hypothesis tests.

This study was approved by the institutional review boards of KPWA and VUMC.

RESULTS

Our corpus selection method yielded corpora with enough leaks to investigate reidentification risk for 5 PII types: patient ages, dates, doctor names, organization names, and patient names (Supplementary Figure 1). Among these PII types, leak rates ranged from 4% for dates to 25% (KPWA) and 22% (VUMC) for patient names (Table 1).

As shown in the lower left quadrant (III) of Figure 3, 23 of 40 results in experiment stage 1 had recall $\leq 50\%$ (indicating most PII was not detected) and precision $\leq 50\%$ (indicating detected PII was mixed with at least as much misleading/resynthesized PII). The only 2 results in the upper left quadrant (I) of Figure 3, where leak detection is greatest, are for organization names (which are not HIPAA identifiers) detected by internal readers. As shown in Table 2 row A, mean recall overall (the 2 corpora, 4 readers, and 5 PII types) was 27% and mean precision was 37%. Readers' leak detection rates were significantly better than chance for 12 (30%) of the 40 results based on a Bonferroni-corrected threshold of $P < .00125$ (.05/40 = .00125) (see Supplementary Table 1).

Quadrant I in Figure 3 is the least densely populated, and quadrant III is the most densely populated, indicating HIPS resynthesis conceals many PII leaks that would have been exposed by traditional redaction approaches. Quadrant II is the second most densely populated, indicating scenarios in which readers meaningfully detect leaked PII but for a limited proportion of actual leaks.

Pooling the results for pairs of readers yielded somewhat higher recall (mean 44%) and somewhat lower precision (mean 31%) (Supplementary Table 3).

Leak detection rates varied considerably by PII type (Table 2, rows B-F). Mean recall and precision for patient names, which are particularly sensitive identifiers, were 22% and 57%, respectively. One reader isolated patient names at a statistically significant rate in an internal corpus (Supplementary Table 1). Organization names had the highest leak detection rate, with a mean recall of 45% and mean precision of 55%. Notably, 3 of the 4 readers detected leaked organization names in one or the other corpus at statistically significant rates (Supplementary Table 1). Mean recall and precision for dates were 32% and 26%, respectively, and 3 of 4 readers were able to detect leaked dates in one or the other corpus at statistically significant rates (Supplementary Table 1). Mean recall and precision for leaked doctor names were 25% and 37%, respectively; 2 readers detected these at a statistically significant rate in their local corpus and none did so in external corpora. Leaked ages had the lowest rates of detection, with mean recall of 9% and mean precision of 10. No attempt to isolate leaked ages was statistically significant (Supplementary Table 1).

The ability to isolate leaks also varied considerably by reader (Table 2, rows G-J; and Figure 3). Summarizing 10 results per reader, mean reader F1 score ranged from 18% to 34%. Readers' abilities to isolate specific types of PII leaks also varied, and superior ability to detect one type of leak did not carry over to other types. The 2 readers with the lowest mean F1 scores for overall leak detection (KPWA-1 at 18% and KPWA-2 at 24%) (Table 2, rows G, H)

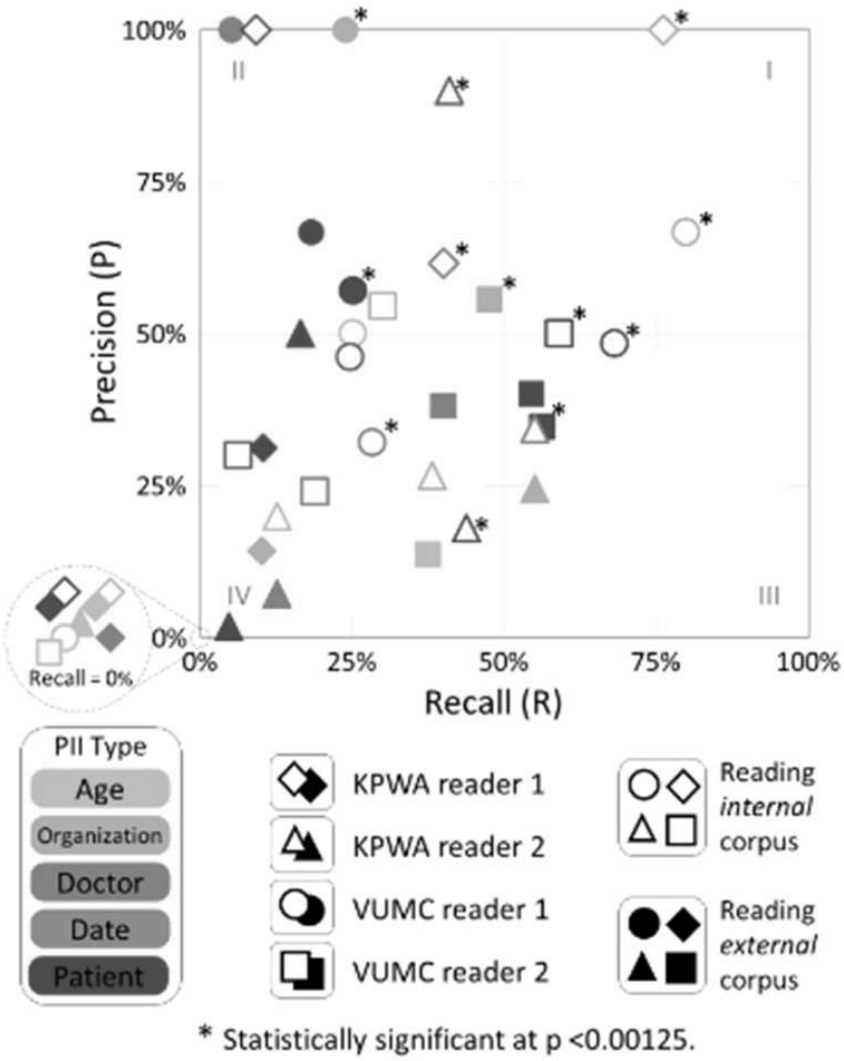

Figure 3. Scatterplot of precision and recall by personally identifying information (PII) type, reader, and corpus in experiment stage 1, without knowledge of actual leak count (inset shows results where recall = 0). KPWA: Kaiser Permanente Washington; PII: personally identifying information; VUMC: Vanderbilt University Medical Center.

had the highest F1 scores for detecting 3 of the 5 PII leak types (organization names: KPWA-1 F1 = 86%; patient names: KPWA-2 F1 = 56%; and doctor names: KPWA-1 F1 = 48%) (see Supplementary Table 1). Similarly, the reader with the second-highest mean F1 score for overall leak detection (VUMC-2: mean F1 score 32%) (Table 2, row J) was not best at detecting any of the 5 PII types individually (Supplementary Table 1).

Leak detection was somewhat higher for internal corpora vs external corpora, as hypothesized (H2). Mean F1 score for leak detection in internal corpora was 31% vs 22% for external corpora. Leak detection was also higher for the KPWA corpus (mean F1 score = 33%) than for the VUMC corpus (mean F1 score = 21%) (Table 2, rows M, N).

Giving readers information about the actual number of leaks in the corpora (experiment stage 2) increased recall rates without substantially degrading precision, as hypothesized (H3) and as illustrated in Figure 4. Median increase in number of leaks detected (stage 2 vs stage 1) was 4 (mean = 3.4), with 24 of 40 results increasing, 2 remaining unchanged, and 14 decreasing. Driven by the positive change in recall, F1 score also increased in stage 2. Detailed results from stage 2 are in Supplementary Table 2.

DISCUSSION

This study extends knowledge regarding the protective effect of HIPS resynthesis to conceal PII leaks that are readily exposed by traditional redaction methods. First, it provides estimates of human reader leak detection rates for more PII types than previously reported. Second, it is, to our knowledge, the first study to comparatively assess leak detection by readers who are internal vs external to the institution releasing the corpus. Third, this study evaluated a realistic attack scenario in which readers may leverage any public information, internal institutional information (except EHRs), or content elsewhere in a corpus as aids to leak detection. Fourth, the systematic approach to selecting a representative study corpus enhances the generalizability of the findings by reducing the chance that results are an artifact of the corpus used.

As hypothesized (H1), leak detection rates in this study (mean overall recall of 27%) (Table 2, row A) were considerably higher than those previously reported. One previous study used a smaller but comparable corpus to investigate detection of 2 PII types—patient names and dates—by 2 internal readers under a less aggressive attack scenario (permitting readers to only use information from the

Table 2. Summary statistics for leak detection metrics

					Recall				Precision				F1			
Row	PII type	Reader[a]	Corpora	Results	Mean	Median	Min	Max	Mean	Median	Min	Max	Mean	Median	Min	Max
A	All	All	Both	40	0.27	0.24	0.00	0.80	0.37	0.33	0.00	1.00	0.27	0.27	0.00	0.86
B	Age	All	Both	8	0.09	0.00	0.00	0.38	0.10	0.00	0.00	0.50	0.09	0.00	0.00	0.33
C	Date				0.32	0.34	0.00	0.68	0.26	0.26	0.00	0.57	0.27	0.30	0.00	0.57
D	Doc.				0.25	0.23	0.00	0.55	0.37	0.33	0.00	1.00	0.25	0.26	0.00	0.48
E	Org.				0.45	0.43	0.10	0.80	0.55	0.55	0.14	1.00	0.46	0.39	0.12	0.86
F	Pt.				0.22	0.17	0.06	0.55	0.57	0.48	0.30	1.00	0.29	0.27	0.10	0.56
G	ALL	KPWA-1	Both	10	0.15	0.05	0.00	0.76	0.31	0.07	0.00	1.00	0.18	0.06	0.00	0.86
H		KPWA-2			0.28	0.27	0.00	0.55	0.27	0.22	0.00	0.90	0.24	0.25	0.00	0.56
I		VUMC-1			0.30	0.25	0.00	0.80	0.57	0.54	0.00	1.00	0.34	0.33	0.00	0.73
J		VUMC-2			0.35	0.39	0.00	0.59	0.34	0.36	0.00	0.56	0.32	0.39	0.00	0.54
K	All	All	Internal[b]	20	0.32	0.29	0.00	0.80	0.40	0.33	0.00	1.00	0.32	0.31	0.00	0.87
L			External[b]		0.22	0.17	0.00	0.56	0.34	0.33	0.00	1.00	0.22	0.22	0.00	0.51
M	All	All	KPWA	20	0.32	0.38	0.00	0.76	0.50	0.45	0.00	1.00	0.33	0.34	0.00	0.87
N			VUMC		0.21	0.11	0.00	0.80	0.24	0.24	0.00	0.67	0.21	0.14	0.00	0.73

Data overall (row A), by PII type (rows B-F), by reader (rows G-J), by internal vs external corpora (rows K, L), and for the KPWA vs the VUMC corpus (rows M, N), with the number of results summarized in each set of statistics (N results) from experiment stage 1 in which actual leak counts are unknown to readers. These summaries can be calculated from detailed data reported in Supplementary Table 1.

Age = patient age; Date = calendar date/time; Doc. = doctor name; HIPS: hiding in plain sight; KPWA: Kaiser Permanente Washington; Org. = organization; PII: personally identifying information; Pt. = patient name; VUMC: Vanderbilt University Medical Center.

[a]KPWA-1 and KPWA-2 refer to the 2 readers from Kaiser Permanente Washington; VUMC-1 and VUMC-2 refer to the 2 readers from Vanderbilt University Medical Center.

[b]Internal and external are defined relative to the reader. For example, the KPWA corpus is internal when read by KPWA readers and external when read by VUMC readers.

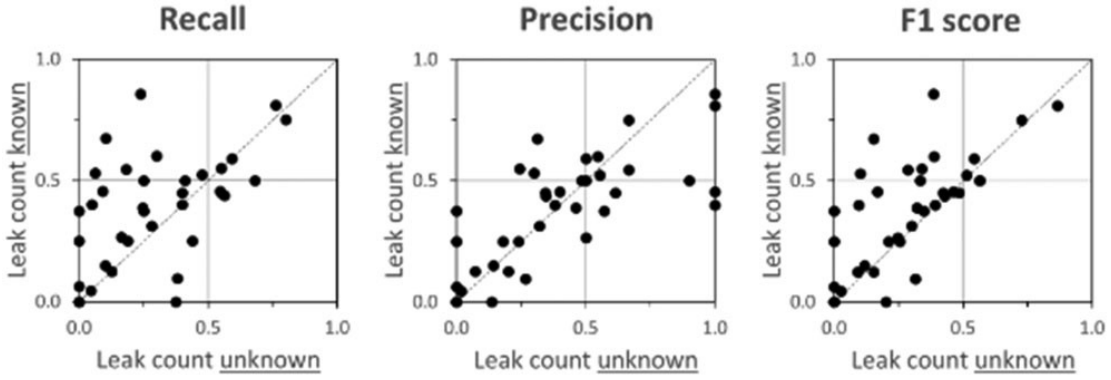

Figure 4. Scatterplot comparing (A) leak detection recall, (B) leak detection precision, and (C) leak detection F1 score when the actual leak count is unknown to readers (horizontal axes) vs when actual leak count is known (vertical axes) for 5 personally identifying information types, 4 readers, and 2 corpora (40 results total). Some plots may appear to have fewer than 40 points due to overlapping observations.

immediate document to isolate leaks). This study reported 2-reader pooled recall, precision, and F1 score for detecting leaked dates of 6%, 25%, and 10%, respectively (Supplementary Table 5). Comparable rates of detecting leaked dates in the present study were 54%, 27%, and 35%, respectively (Supplementary Table 3). The same previous study reported 2-reader pooled recall, precision, and F1 score for detecting leaked patient names of 13%, 67%, and 21%, respectively (Supplementary Table 5). Comparable rates detecting leaked patient names in the present study were 34%, 48%, and 39%, respectively (Supplementary Table 3). The present study's more aggressive (and realistic) attack scenario may account for much of this difference, but the previously reported rates may also have been artifacts of a smaller corpus or the particular PII tagger model used.

Examples of patient name leaks correctly detected by readers may suggest areas for future methods development. Some name leaks result from inadvertent loss of whitespace as in "CarrellMRN : 123456." Such irregularities are easily parsed by human readers but are challenging for machine-learned PII taggers. Other detected leaks have irregular capitalization or formatting, as in "CARRELL, David SCott"—which may confuse a PII tagger but not a human. The most commonly detected name leaks were repeat mentions of a patient's name mid-note following a correct tagging and resynthesis of the name at the beginning of the note. Augmenting machine-learned model with rules to propagate tagged names throughout a note may resolve many of these leaks.

Given data stewards' interests in worst case scenarios, it may be possible to infer from our results an approximate upper bound of overall leak detection ability for an exceptionally capable person conducting a reader attack. Let us assume that precision of at least 50% is required for meaningful leak detection (ie, when a set of sup-

posed leaks includes at least as many actual PII instances as misleading resynthesized instances) and, further, that an exceptional attacker could detect leaked PII at rates that matched the highest detection rate of any of our 4 readers for each PII type. Then, we may estimate the upper bound of meaningful leak recall by an exceptionally capable attacker to be 25% for ages (with 50% precision), 40% for doctor names (with 62% precision), 41% for patient names (with 90% precision), 59% for dates (with 50% precision), and 76% for organization names (with 100% precision) (Supplementary Table 1).

Notably, the leak detection rates reported in this study are considerably lower than those recently reported for attacks that leveraged machine learning.[18] Using a similar corpus, a similarly trained PII tagger, and identical HIPS resynthesis, that attack achieved mean recall of 71%, mean precision of 52%, and mean F1 score of 58% across the same 5 PII types (Supplementary Table 4). Comparable leak detection measures in the present study were 27%, 37%, and 27%, respectively. The machine attack's mean recall rates were even superior to the present study's pooled recall rates for 2-reader (44%), 3-reader (55%), or 4-reader (62%) pools (Supplementary Table 3, rows F, L, and R). Attacks leveraging machine learning appear to present the greater reidentification threat—at least when implemented at scale.

The higher overall recall rate in internal (mean = 32%) vs external (mean = 22%) corpora (Table 2, rows K, L) suggests that an attacker's local institutional knowledge affords some advantage in leak detection, driven largely by detection of leaked organization names (Figure 3). An internal employee's access to sensitive internal information systems may pose a greater institutional risk than that associated with imperfectly de-identified corpora.

In stage 2, in which readers used the actual leak counts to detect leaks, leak detection recall improved modestly (Figure 4), suggesting that leak detection in stage 1 (without knowledge) was somewhat diminished by reader conservatism. This conservatism, however, did not dominate stage 1 results.

Study limitations

This study has several limitations. First, though considerable effort was devoted to selecting representative corpora and PII tagger models, idiosyncrasies in either may have influenced results. Performance of the PII tagger models (median recall rates $\geq$75% and median precision rates $\geq$90%) was inferior to state-of-the-art performance in our 2 institutions. This was intentional and, we believe, acceptable because it allowed us to conduct the experiment with manageable size corpora. We believe that it was also superior to more idiosyncratic approaches in which a single tagger model and corpus combination is used,[14,19] or in which PII leaks were artificially introduced. The 10% overall PII leak rate in our corpus was comparable to those used in similar experiments.[19] Nevertheless, detection rates in real-world settings may be different. Second, our 4 readers may not be representative of all potential attackers, and we did not give readers financial incentives, which limits somewhat the experiment's fidelity to a real-world adversarial attack. Third, we were not able to meaningfully assess leak detection rates for all HIPAA identifiers due to an inability to naturally generate some types of leaks in sufficient volumes. Highly sensitive numerical identifiers such as medical record numbers, social security numbers, and patient phone numbers tend to be rare in clinical notes, and when they do appear, their regularity makes them relatively easy for de-identification systems to redact.

CONCLUSION

The HIPS approach to concealing leaked PII in an otherwise de-identified clinical natural language corpus offers substantial benefit over traditional redaction. We specifically showed that an estimated 70% of leaked PII remained undetected by human readers at 2 different healthcare institutions under realistically aggressive attack scenarios. While still advantageous, the added protection compared with traditional redaction is lower than previously reported. Additionally, the leak detection capability of human readers appears to fall short of that reported for an adversary assisted by machine learning. Future research should investigate whether alternative PII resynthesis strategies may impact leak detection rates and how novel PII tagger methods may reduce leak rates.

FUNDING

This work was supported by the National Library of Medicine grant number R01LM011366 (to DSC); National Human Genome Research Institute grant numbers RM1HG009034 (to BAM), U01HG008657, and U01HG008701; and the MITRE Corporation (to JSA and LH).

AUTHOR CONTRIBUTIONS

DSC, BAM, JSA, CC, ML, and LH contributed to the conception and design of the project. DSC, BAM, DJC, ML, and SN implemented selection of study corpora at their respective institutions. All authors participated in creation of gold standard annotation of the corpora and implementation of the experiments. All authors contributed to the interpretation of results. DSC wrote the first draft of the article and all authors provided critical feedback on and approved the final version of the article.

SUPPLEMENTARY MATERIAL

Supplementary is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

ML performed the work reported in this article while a student at Vanderbilt University; she is currently employed by Privacy Analytics, Inc, a company that does business in the de-identification space. Other authors have no competing interests to declare.

REFERENCES

1. Velupillai S, Suominen H, Liakata M, *et al*. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *J Biomed Inform* 2018; 88: 11–9.
2. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019; 26 (4): 364–79.
3. Wang Y, Sohn S, Liu S, *et al*. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019; 19 (1): 1.
4. Yu S, Ma Y, Gronsbell J, *et al*. Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc* 2018; 25 (1): 54–60.

5. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019; 7 (2): e12239.
6. European Medicines Agency. Data anonymization-a key enabler for clinical data sharing (Workshop report, 30 November–1 December 2017, European Medicines Agency, London). 2018. https://www.ema.europa.eu/en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing_en.pdf Accessed October 29, 2019
7. Tucker K, Branson J, Dilleen M, *et al*. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med Res Methodol* 2016; 16 (S1): 77.
8. U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information; final rule. *Fed Regist* 2002; 67: 53181–273.
9. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform* 2019; 132: 103971.
10. Rothstein MA. The Hippocratic bargain and health information technology. *J Law Med Ethics* 2010; 38 (1): 7–13.
11. Carrell DS, Cronkite DJ, Malin BA, Aberdeen JS, Hirschman L. Is the worth the squeeze? Costs and benefits of multiple human annotators for clinical text de-identification. *Methods Inf Med* 2016; 55 (4): 356–64.
12. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care* 2012 Jul; 50 Suppl: S82–101.
13. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol* 2010; 10 (1): 70.
14. Carrell D, Malin B, Aberdeen J, *et al*. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *J Am Med Inform Assoc* 2013; 20 (2): 342–8.
15. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc* 2017; 24 (3): 596–606.
16. Hirschman L, Aberdeen J. Measuring risk and information preservation: toward new metrics for de-identification of clinical texts. In: proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents; 2010: 72–5.
17. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. *J Am Med Inform Assoc* 2015; 22 (5): 1029–41.
18. Carrell DS, Cronkite DJ, Li MR, *et al*. The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight. *J Am Med Inform Assoc* 2020; 26 (12): 1536–44.
19. Grouin C, Griffon N, Neveol A. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In: proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi); 2015: 31–9.
20. Li M, Carrell D, Aberdeen J, *et al*. Optimizing annotation resources for natural language de-identification via a game theoretic framework. *J Biomed Inform* 2016; 61: 97–109.
21. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. *J Biomed Inform* 2015; 58: S53–9.
22. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007; 14 (5): 550–63.
23. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. *J Biomed Inform* 2015; 58: S11–9.
24. Aberdeen J, Bayer S, Yeniterzi R, *et al*. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *Int J Med Inform* 2010; 79 (12): 849–59.
25. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *J Am Med Inform Assoc* 2013; 20 (1): 77–83.
26. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp* 1996; 1996: 333–7.
27. Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods Inf Med* 2006; 45 (3): 246–52.
28. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* 2008; 15 (5): 601–10.
29. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008; 17 (1): 128–44.
30. Morrison FP, Li L, Lai AM, Hripcsak G. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *J Am Med Inform Assoc* 2009; 16 (1): 37–9.
31. 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 (5): 574–80.
32. Wellner B, Huyck M, Mardis S, *et al*. Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc* 2007; 14 (5): 564–73.
33. Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. Effects of personal identifier resynthesis on clinical text de-identification. *J Am Med Inform Assoc* 2010; 17 (2): 159–68.
34. Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp* 2002; 2002: 757–61.
35. Neamatullah I, Douglass MM, Lehman LW, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008; 8 (1): 32.
36. 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. *Proc AMIA Symp* 2009; 2009: 416–20.
37. Gardner J, Xiong L. An integrated framework for de-identifying unstructured medical data. *Data Know Eng* 2009; 68 (12): 1441–51.
38. Sadat MN, Aziz MMA, Mohammed N, Pakhomov S, Liu H, Jiang X. A privacy-preserving distributed filtering framework for NLP artifacts. *BMC Med Inform Decis Mak* 2019; 19 (1): 183.
39. Li B, Vorobeychik Y, Li M, Malin B. Scalable iterative classification for sanitizing large-scale datasets. *IEEE Trans Knowl Data Eng* 2017; 29 (3): 698–711.
40. Li M, Carrell D, Aberdeen J, Hirschman L, Malin BA. De-identification of clinical narratives through writing complexity measures. *Int J Med Inform* 2014; 83 (10): 750–67.
41. MITRE. MITRE Identification Scrubber Toolkit (MIST); 2011. http://mist-deid.sourceforge.net/ Accessed September 3, 2018.
42. Mazor KM, Richards A, Gallagher M, *et al*. Stakeholders' views on data sharing in multicenter studies. *J Comp Eff Res* 2017; 6 (6): 537–47.