

# Assessing the Difficulty and Time Cost of De-identification in Clinical Narratives

D. A. Dorr<sup>1</sup>, W. F. Phillips<sup>2</sup>, S. Phansalkar<sup>3,4</sup>, S. A. Sims<sup>3,4</sup>, J. F. Hurdle<sup>3,4</sup>

<sup>1</sup>Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, USA

<sup>2</sup>School of Computing, University of Utah, Salt Lake City, UT, USA

<sup>3</sup>Department of Medical Informatics, University of Utah, Salt Lake City, UT, USA

<sup>4</sup>Geriatric Research, Education, and Clinical Center (GRECC), George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT, USA

## Summary

**Objective:** To characterize the difficulty confronting investigators in removing protected health information (PHI) from cross-discipline, free-text clinical notes, an important challenge to clinical informatics research as recalibrated by the introduction of the US Health Insurance Portability and Accountability Act (HIPAA) and similar regulations.

**Methods:** Randomized selection of clinical narratives from complete admissions written by diverse providers, reviewed using a two-tiered rater system and simple automated regular expression tools. For manual review, two independent reviewers used simple search and replace algorithms and visual scanning to find PHI as defined by HIPAA, followed by an independent second review to detect any missed PHI. Simple automated review was also performed for the "easy" PHI that are number- or date-based.

**Results:** From 262 notes, 2074 PHI, or 7.9 ± 6.1 per note, were found. The average recall (or sensitivity) was 95.9% while precision was 99.6% for single reviewers. Agreement between individual reviewers was strong (ICC = 0.99), although some asymmetry in errors was seen between reviewers (p = 0.001). The automated technique had better recall (98.5%) but worse precision (88.4%) for its subset of identifiers. Manually de-identifying a note took 87.3 ± 61 seconds on average.

**Conclusions:** Manual de-identification of free-text notes is tedious and time-consuming, but even simple PHI is difficult to automatically identify with the exactitude required under HIPAA.

## Keywords

Health Insurance Portability and Accountability Act, Computerized medical records systems, medical informatics computing, natural language processing, de-identification

Methods Inf Med 2006; 45: 246–52

## 1. Introduction

A key issue in performing research using clinical information systems is to protect the confidentiality of participants, either through a consent process or by removing personal information (de-identification). In the case of epidemiologic or other large-scale research projects, asking individual patients for consent may be prohibitively expensive or simply impractical, making de-identification an attractive option. However, the majority of clinical data is stored as narrative or free-text data, even in electronic health records [1]. This narrative text, although unstructured, is useful for many research purposes such as qualitative studies, understanding of health care process, retrospective chart reviews, and advanced computer analytic techniques such as natural language understanding (NLU). The sharing of textual databases between research groups either to develop automated extraction techniques or to increase the dissemination and generalizability of results has been endorsed by multiple groups, including by technique developers and governmental and non-profit agencies [2-4]. Clinical researchers who wish to collaborate using narrative data would greatly benefit from the development of effective techniques for de-identification of narrative datasets.

For narrative data in electronic format, previous research has focused on advanced techniques to automatically de-identify, tag, or impose structure for future de-identification on free-text clinical data [5-8]. The process of complete de-identification requires removing several kinds of protected health information (PHI). Some types of PHI such as dates or telephone

numbers have been assumed to have a relatively constant format, lending themselves to simple stripping techniques. Other PHI, such as names, seem much harder, requiring significant technological expertise to remove automatically.

A typical clinical research group likely lacks the experience or the inclination to develop its own advanced de-identification techniques, but unfortunately most published de-identification techniques are difficult to generalize [9]. This owes to the fact that implementation of an advanced technique requires considerable computer science expertise and a deep understanding of site-specific datasets. This study attempts to anticipate the approach of investigators with limited experience in de-identification who need to remove PHI from an electronic narrative dataset. These investigators are presumed to have limited experience in de-identification, although we do assume that they have the expertise to use simple, automated pattern matching techniques to remove simple PHI types from their datasets. For the more complicated PHI types, like names, we assume that they will have to rely on manual identification/removal of PHI. We define diagnostics for "simple automated" and manual removal of PHI, and describe the time incurred to ensure all PHI is removed.

### 1.1 The Impact of HIPAA and Similar Regulations: Redefining De-identification

The legal definition of what constitutes appropriate de-identification of data to ensure confidentiality and security of PHI has been recalibrated in the US with the passage of

the Health Insurance Portability and Accountability Act (HIPAA, PHI is defined specifically in 45CFR§164.514). The European Union Data Protection Directive has also increased the importance and refined the definition of de-identification (used synonymously with anonymization here), although the precise definition is dependent on the data content and context [4]. Other countries have also been advancing privacy laws, increasing the importance of de-identification [10]. We choose HIPAA as test case since it uses an internationally developed standard (the “Safe Harbor” technique) as its base, and as the relatively structured definition allows for careful consideration of anonymization.

In cases where a dataset is to be used for research purposes but obtaining patient consent is impractical, HIPAA provides two alternatives regarding PHI: 1) a statistical proof that any shared PHI cannot be linked to an individual; or 2) the “Safe Harbor” technique where all PHI are removed before data sharing. The former is defined quite vaguely in the HIPAA rules and can never be safely applied to unfiltered clinical notes, where the inclusion of single name results in a near certain probability of identification. The more straightforward Safe Harbor technique requires the complete removal of 18 types of identifiers. These are displayed in Table 1. Although HIPAA has only recently redefined the legal standard for de-identification in the US, techniques for the removal of named entities from clinical text have been well studied in the past [11-13].

## 1.2 Review of Previous Work: Named Entity Extraction/Tagging and Alternatives to De-identification

A clinical research group has several options when faced with the task of de-identifying a narrative dataset. The first option is to transform free text into structured data, with the assumption that identification of PHI is easier in a structured database. Despite some success with this technique using pathology reports [14], it requires considerable local informatics expertise and renders the data unusable for almost all NLU tech-

niques, since syntactic and semantic information is lost. Even in the narrow clinical context of pathology, the concepts detected are subject to systematic bias, usually resulting from the specific structuring technique (e.g., the structure of the standardized concept lexicon) [14]. Benchmarking procedures have been advocated to identify anonymized data to improve the comparability of such research efforts [15, 16].

A second approach is to create de-identified narrative datasets directly. Early innovators have already begun sharing restricted-domain narrative datasets. For example, the Scrub system and the Shared Pathology Information Network (SPIN) have had success with a subset of narrative data types [11, 17, 18]. The Scrub system compared automated and manual removal of names from letters between physicians, as these had a semi-structured format. Notably, the manual efforts in Scrub had a non-trivial error rate with “irregular” instances of PHI (5-6%), while the automated system was trained to catch 99% of PHI instances [11]. Similarly, the SPIN efforts have led to a working system which shares

certain de-identified information from pathology notes after extensive training to reduce the error rate from 6.3% to 1% [18].

Multiple other studies have experimented with different subsets of clinical data using a variety of advanced technology, with varying success. Thomas et al. used a database of names and non-name words to raise or lower the probability of removing patient and provider names in pathology reports, and achieved a sensitivity of 92.7% for complete reports and 98% in just the report narratives themselves [6]. Ruch et al. used a medical tagging program and achieved a sensitivity of 96.8% in clinical narratives [7]. Taira et al. used semantic selection techniques to remove names from urology reports and achieved a sensitivity of 93.9% [8]. Comparing sensitivity scores between these studies is not possible since the total number of PHI was not reported. The applicability of these advanced techniques to other clinical domains’ narrative data, and to less-informatics savvy institutions, remains unclear.

Outside of healthcare the task of name identification, the most difficult PHI type,

**Table 1** HIPAA’s 18 protected health identifiers (PHIs) and their use in this study

PHI type	Manual removal	Automated removal
Names	✓	
Dates (except year)	✓	✓
Geographic identifiers	✓	
Telephone numbers*	✓	✓
Fax numbers	✓	✓
Electronic mail addresses	✓	
Social security numbers (SSNs)	✓	✓
Medical record numbers <sup>†</sup>	✓	✓
Health plan beneficiary numbers	✓	
Account numbers	✓	
Certificate/license numbers	✓	
Device identifiers and serial numbers	✓	
Web Universal Resource Locators (URLs)	✓	
Internet Protocol (IP) address numbers	✓	
Vehicle identifiers and serial numbers, including license plate numbers	✓	
Biometric identifiers, including finger and voice prints	✓	
Full face photographic images and any comparable images	✓	
Any other unique identifying number, characteristic, or code	✓	

\* includes multiple types of pagers; <sup>†</sup> includes the internal VA identifier of first initial of last name plus the last SSN digits only.

has been extensively investigated. The Message Understanding Conferences (MUCs), sponsored by the National Institute of Science and Technology, evaluated the ability of various systems to extract information and recognize conceptual entities from narrative text coming from various sources. Named-entity extraction was one of the evaluation domains in MUC-6 and 7 [19]. Using newswire reports the various systems had F-scores, which summarize both recall and precision, as high as 94-97%, essentially equivalent to single human review. One of the caveats from the results, however, is that detection by the various tools was limited when the domain of the source texts changed; suggesting that de-identification techniques must always be re-evaluated when moved to another setting like healthcare [20].

Even single-person manual de-identification was not perfect in the MUC trials, with 3-6% of identifiers missed. To fully comply with HIPAA under the Safe Harbor standard, one must remove all names and other types of PHI. In its evaluation of the HIPAA rules for de-identification, the Workgroup for Electronic Data Interchange highlighted the difficulty of applying the rules to narrative data when it recommended that “[u]ntil there is a cost effective technological solution [to de-identification of free text], data within unstructured free text should be removed” [9]. We explore the feasibility of performing manual and simple automated de-identification of narrative text from several different clinical domains with an attempt to evaluate the feasibility and cost-effectiveness (in terms of personnel time) of such an effort. As shown in Table 1, we define complete de-identification through the HIPAA subtypes. To more closely approximate the work of investigators unskilled in advanced text mining, we limit our automated technique to simple patterns, relying primarily on manual review.

## 2. Methods

### 2.1 Data and Setting

The corpus used for de-identification consisted of 262 notes randomly selected from a

research database containing 88,000 distinct patient notes. The 262 notes were all the notes associated with a small number of admissions from the year 2000; the admissions themselves were culled from a series of consecutive inpatient hospitalizations at the Veteran’s Administration (VA) hospital in Salt Lake City chosen at random for a study on adverse drug events (for details on that study, see the article by Nebeker et al. [21]). As part of the study, the researchers made an electronic copy of the clinical notes from the VA electronic medical record pertinent to the hospitalization. Each copied set included a wide variety of inpatient, outpatient, and administrative notes, since providers are responsible for typing their own progress and consultation notes at this site (with the exception of emergency care and a few surgical locations). This database of pertinent notes, used for adverse event detection in the primary study, were used for de-identification in the current study. The research described here was approved by the local Institutional Review Board.

### 2.2 Training

In the development of the gold standard for manual tagging, this study attempted to replicate the manual tagging techniques in the Message Understanding Conferences (MUCs) [19], as well as the illustrative work in subjective manual tagging of Bruce and Wiebe [22]. An instruction manual was prepared (available on request). This manual consisted of the text of the Code of Federal Regulations for de-identification under HIPAA and instructions on how to mark the text under review. The reviewers were given additional verbal instruction for the most difficult domains (names, geographic locations).

### 2.3 Study Design

Tagging the texts to build the gold standard consisted of two parts: a manual review and a straightforward, simple automated scan using regular expressions. These techniques were picked to simulate the resources available to most research teams interested in de-identification. For the manual tagging, two

independent reviewers (one pharmacist and one non-clinical student; hereafter, initial reviewers) marked up the entire 262 target note corpus and removed all PHI by: 1) visually scanning the text to detect and then code PHI by HIPAA type; and 2) using word-processor based search and replace methods to detect and code PHI when a common pattern became apparent to them (e.g., searching for the word “Date” or a specific patient’s name, once identified). The PHI was replaced with the token <PHI\_type>, where type was any of the HIPAA types. The initial reviewers were chosen so as to represent people with a clinical and non-clinical background who might be hired to complete such tasks in the real world. Using two independent reviewers with and without clinical expertise also allowed us to get a more accurate understanding of diagnostic errors. To assess time cost, the two initial reviewers timed themselves while manually de-identifying a randomly chosen 70 of the 262 notes in the corpus.

A separate program counted words, PHI, and PHI types in each initial reviewer’s marked-up documents. A complete example of PHI (e.g., “May 18, 2005”) was counted as single token (<PHI\_date>) rather than three words; this provided a conservative analysis for time and specificity calculations. Two physician reviewers then examined the two de-identified documents from the initial reviewers against the original, identified version for missed and wrongly selected elements, corrected mistakes, and created a final de-identification document. This final document was used as the gold standard against which the tagging by the initial two reviewers and by the regular expression processor were compared. Discrepancies were resolved using consensus methods. A simple automated tagging program was created for the highly structured PHI types (enumerated in Table 1), since the general perception is that these types are easy to remove [17]. Table 2 displays the regular expressions used by the simple automated system. The automated technique searched for text strings using a simple regular expression engine, matching and tagging them using the same annotation syntax as the initial, manual reviewers. These results were

compared to the gold standard for the subset of PHI types.

## 2.4 Statistical Analysis

Sensitivity (recall), specificity, and positive predictive value (precision) were calculated for each initial reviewer against the final gold standard; simple averages were used for overall single rater diagnostics. To assess inter-rater agreement, the number of PHI detected per note is assumed to be a continuous measurement scale. Therefore, reliability (inter-rater agreement) between initial reviewers was measured using the intra-class correlation coefficient [23]. The Cohen's kappa ( $\kappa$ ) statistic was used to detect differences in classification of PHI by the initial manual reviewers. Asymmetry of errors was assessed using Bowker's extension of McNemar's test. Means for time to manual de-identify notes were compared with Student's *t* between initial reviewers prior to combination in a single overall mean, and correlations between time and number of tokens in a note were obtained using Pearson's correlation coefficient. SAS<sup>®</sup> version 8.1 was used for all statistical analysis.

## 3. Results

The 262 notes in the de-identification corpus came from authors in several clinical domains (including Internal Medicine, Hematology, Emergency Care, Psychiatry, Acute Care and ICU Nursing, Inpatient Pharmacy, and Mental Health Group Education), and multiple healthcare professions (physician or physician extender, nurse, social worker, and clinical pharmacist). Table 3 describes the source and concentration of PHI in the note corpus. A total of 2074 PHI were found, an average of  $7.9 \pm 6.1$  identifiers per note, with an average note length of  $261 \pm 352$  words. The majority of identifiers were specific dates of treatment or admission ( $4.0 \pm 3.9$  per note), with provider name second ( $1.9 \pm 1.9$  per note). The prevalence of named entities (1.00/100 words, where a PHI is conservatively counted as one word) is substantially less

than in the work by Thomas et al. (7.1/100 words) [6] but was quite similar to the work by Ruch et al. (1.01 names/100 words) [7] which covered a similar narrative domain as ours.

## 3.1 Initial Manual Reviewer Performance

The average performance of the single initial reviewers is given in Table 4. The average sensitivity (recall) was 95.9%, and the average precision (positive predictive value) was 99.6% with seven incorrect PHI assignments out of 1996. For names only, the average recall was 96.4% and precision was 98.9%. The false positives from names were reviewed and found to be due to the incorrect tagging of medical devices or proce-

dures as names, and were more likely to come from the non-clinical initial reviewer. Sub-analyses of the other areas of PHI revealed similar sensitivities throughout. An exception was phone numbers, which had a recall of 99.3% and a precision of 99.8%. If one considers each non-PHI word as a true negative, the recall was 99.99%. A denominator using every word in the gold standard dataset (where each instance of PHI was only represented as a one-word token) was chosen for comparison against automated techniques, which necessarily consider every word.

The inter-rater agreement between the two single raters for the number and type of PHI per note was extremely high (ICC = 0.99, 95% CI 0.98–1.00). Phone numbers had the lowest agreement (ICC = 0.95, 95% CI 0.94–0.96). In notes where the ini-

**Table 2** Regular expressions for the highly structured protected health identifier (PHI) types

PHI type	Description	Pattern	Example
Phone numbers	A fully qualified US long-distance or local number*	1 ###-###-#### 1-###-###-#### 1(###)###-#### (###)###-####, (###)-###-#### ###-###-#### ###-###-####	1 801-582-1565 1-801-582-1565 1(801) 582-1565 (801) 582-1565 (801)-582-1565 801-582-1565 582-1565
	A stand alone extension (University and VAMC variants)*	####, #-####, x####, x-####, where x can be 'x', 'ex', 'ex.', 'ext', 'ext.'	2468, n2468, ex 2468, ex. n2468, ext. 2468, ext. n2468, etc
	A pager number	page <any number type above> pager <any number type above>	
Dates	Simple numeric dates	MM/DD/YY, MM-DD-YY MM/DD/YYYY, MM-DD-YYYY MM/DD, MM-DD	01/01/05, 01-01-05
	Mixed alphanumeric dates	month_token DD month_token DD, YY month_token DD, YYYY DD month_token DD month_token YY DD month_token YY DD month_token YYYY	Nov 02 Nov. 2, 04 Nov. 2, 2004 2 November 02 November 04 2 November 04 02 November 2004
Patient identifier codes	Regular SSNs	###-##-#### #####	543-34-5678 543345678
	Local Medical Record Number	?####	D5432

\* Full numbers and extensions, when combined, count as one number only.

SSN is Social Security Number, VAMC is Veterans Administration Medical Center.

In all cases, the number of white space characters, such as <space> or <carriage return> may be arbitrarily large; internal commas and dashes are optional; the symbol # stands for any digit and ? for any letter; the following are literals: (, ), page, pager, -, the standard day-month-year symbols of DD, MM, YYYY refer to digits, the leading 0 or 1 is optional (e.g., 11/2/04 is the same as 11/02/2004); month\_tokens are (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Sept, Oct, Nov, Dec, January, February, March, April, May, June, July, August, September, October, November, December) – note that any abbreviation may contain an optional period, e.g., "Jan." Perl code for these regular expressions are available on request.

**Table 3** Source of notes and concentration of PHI each contributed

Location	N	%
Outpatient	109	41.8%
Inpatient	147	56.3%
Other*	6	2.0%
<b>Author</b>		
Physician or extender	177	67.6%
Nurse	54	20.6%
Social Worker	9	3.4%
Pharmacist	21	8.0%
Other	1	3.8%
<b>PHI content (262 notes)</b>		
	Mean	SD
Total PHI	n = 2074 PHI	
Average per note	7.9	6.1
Provider name	1.9	1.9
Patient name	0.8	1.9
Date	4	3.9
Phone number	0.9	0.7
Geographic location	0.2	0.6
Other	0.1	0.1
Total word count	70,552	
Average words per note	269.3	352.5
PHI prevalence	2.94/100 words	
Name prevalence	1.00/100 words	

\* Other includes nursing home, home health, and emergency department. PHI = protected health information.

**Table 4** Average diagnostic performance of initial reviewers versus gold standard (two stage review)

Average of initial reviewers	Gold standard			Prevalence	2.94/100		
		PHI +	PHI -			Total	
	PHI +	1989	7			1996	
	PHI -	85	68471			68557	Sensitivity (recall) 95.9%
	Total	2074	68478			70552	Specificity 99.99%
				Precision 99.6%			

**Table 5** Five subtypes of PHI found by a simple automated technique

Automated	Gold standard			Prevalence	1.83%		
		PHI +	PHI -			Total	
	PHI +	1274	168			1442	
	PHI -	19	69091			69110	Sensitivity (recall) 98.5%
	Total	1293	69259			70552	Specificity 99.8%
				Precision 88.4%			

tial reviewers did not agree, they differed by one element of PHI 56% of the time. Analyzing the few mistakes the initial reviewers made versus the gold standard (by the physician reviewers) did reveal some asymmetry (Bowker's extension of McNemar's test,  $p = 0.001$ ); the clinician reviewer had 14 of the 15 false positives. Classification agreement for shared signals (e.g., where both initial reviewers found a true positive patient identifier) was strong ( $\kappa = 0.89$ , 95% CI 0.88-0.91,  $N = 1925$  shared matches over five PHI categories). The relatively few classification errors were between patient and provider names. Since the HIPAA standard is vague about whether to exclude provider names, these errors may or may not be important.

### 3.2 Simple Automated Performance

Table 5 contains the overall results of the automated technique versus the same five identifiers (date, phone, fax, Social Security Number or SSN, and medical record number) in the gold standard. The technique had higher recall, lower precision, and many more classification errors of PHI type than the initial human reviewers. The lower precision (88.4%) was due to the automated technique producing many more false positives than the initial manual reviewers (e.g., misclassifying a laboratory value – not a PHI type – as a pager number). The recall, however, was slightly better than the initial human reviewers for these three subtypes (automated: 98.5%, human average: 96.7%).

A similarity between the regular expressions for some telephone numbers and dates led to misclassification by the automated system (Bowker's extension of McNemar's: 44.26,  $p < 0.0001$ ). The actual misclassification depended on the search order; that is, if dates were sought first, the ambiguous identifiers (67 in total) were classified as dates and vice versa for telephone numbers. These misclassifications did not hurt performance (all were true positive PHI). However, the inability to accurately classify type would worsen pseudo-anonymization performance (where sen-

tence structure and syntax is maintained by replacing an identifier with another token of the same PHI type) by replacing PHI with the wrong type.

### 3.3 Time Costs

Table 6 describes the time costs and correlations for notes based on number of words and PHI. Individual initial reviewer times were not significantly different ( $t = 0.320$ ,  $p = 0.86$ ), and are combined on the Table. The time per note was most highly correlated with the number of PHI per note ( $r = 0.90$ ), although word count and time were also highly correlated ( $r = 0.83$ ). Manually de-identifying a note took  $87.3 \pm 61$  seconds on average or  $11.3 \pm 3.6$  seconds per PHI. By extrapolation, de-identifying the entire master dataset (88,000 notes) would take 58 weeks and 1 day (95% CI 57 weeks 6 days to 58 weeks 3 days), assuming de-identification took place 7.5 hours per day during a 5-day work week. It is unlikely that manual reviewers could maintain consistency tagging text as a full-time job due to the tedium and repetitiveness of the task, so this represents a lower bound.

## 4. Discussion

Our manual and simple automated techniques have similar error rates to studies conducted on non-healthcare narrative data, and our results suggest that these techniques have significant problems when attempting to meet the performance requirements imposed by HIPAA. Thus, investigators should be cautious in applying both manual and simple automated methods which lack HIPAA-compliant recall and precision levels, as well as be cautious with more advanced de-identification techniques which may not be appropriate, or even possible, to generalize from other sites or domains.

In our analysis, agreement between the initial reviewers for the manual technique was high. In addition, their precision averaged 99.6%. Thus, the errors they did make were largely missed true positives rather than misassignment as false positives. While this trend is problematic from a pure

de-identification perspective as defined by HIPAA, it does imply that inadvertent information loss with a manual technique would be relatively low, since pertinent names (e.g., a procedure name) would not be eliminated. However, the analysis did detect asymmetry, with one initial reviewer producing more false positives. This was determined to be from a faulty manual search-and-replace algorithm based on one of the patients' names which was also a verb in common usage; the initial reviewer attempted to correct these introduced errors, but missed some in the visual review.

The simple, automated search-and-replace technique was deemed an appropriate approach for only five of the 18 Safe Harbor identifiers, but these identifiers (faxes, phone numbers, dates, SSNs, and medical record numbers) made up 62% (1293/2074) of the PHI in the studied subset. The performance of the regular expressions was good (recall of 98.5%). However, the information loss was greater as the precision was only 88.4%, and the rate of misclassification errors was high due to the lexical similarity of the expressions involved. Assuming the workload for manual de-identification is reduced by the amount of PHI found automatically (a reduction of 61.4%), a single reviewer would still take over 22 weeks to de-identify the rest of the master dataset of 88,000 notes. This assumption is likely too conservative, since the majority of the time is spent scanning the text, not simply removing PHI. If the published, advanced automated techniques described in the introduction had a similar sensitivity on these narrative data as they reported (93-96%), finding and eliminating the remaining PHI would still take many weeks of effort.

The data reported here used both notes and word tokens (where a date is one token) as the unit of measure. HIPAA sanctions for revealing PHI apply at the patient level, and increase directly with the number of PHI breaches; thus, unit of analysis for practical purposes is patient level. Word tokens are an awkward unit, but meaningful values for recall and precision can only be calculated at this basic level. Each word is considered as to the presence or absence of PHI; the definitions of true positive and negative is based on this method. In all, 193 of the 262 notes would have been de-identified correctly (73.6%), but none of the patients would have been completely de-identified using the methods described here.

The study has several limitations. The number of notes ( $N = 262$ ) and their selection (patient hospitalizations) might miss significant characteristics of the dataset. Further de-identification is being completed on the rest of the dataset using more advanced techniques and checked manually; this work helps us to understand how often the manual checking might also propagate errors. Similarly, the time costs are extrapolated from a very small subset (70 notes); however, the equivalence between initial reviewers indicates some stability of measurement. The 70 notes were chosen to minimize the increased time needed at the beginning of de-identification due to the learning effect. Finally, others have created excellent advanced techniques in this area; although such advanced techniques often yield better results, the definition of PHI, the sophistication needed for their execution, and the lack of a reasonable gold standard led us to examine the underlying diagnostic error inherent in even simple and manual techniques.

**Table 6** Time costs for manual de-identification of narrative notes by initial reviewers ( $n = 70$ )

Counts	Total	Mean per note	SD
Word count	22,232	317.6	354.7
PHI count	553	7.9	5.3
Times (in seconds)	(N = 70)	Mean	SD
Time per note		87.3	61.0
Time per word		.48	.36
Time per PHI		11.3	3.6

Our assessment of time and personnel cost outcomes in a diversity of clinical note types (eight different clinical domains) is the first study of its kind in the healthcare literature. Although much work has been done with other, advanced NLU techniques, this work provides time-costs for the effort required by a relatively naive investigator group to use simple techniques to de-identify data. Our work suggests that the time and personnel costs associated with manual de-identification represent a significant barrier to clinical informatics research involving narrative texts.

## 5. Conclusions

The US federal government estimates it will cost 55 million dollars annually to comply with the de-identification standards within HIPAA [6]. However, this estimate specifically excludes the costs of de-identifying narrative text, the very format in which the vast majority of clinical information is stored. As advanced analytic techniques such as NLU (which can mine the knowledge from such narrative data) increase in power and scope, so do the benefits of de-identifying and sharing this information. This paper is presented as a first step towards assessing the difficulties and costs of such de-identification and, by using simple techniques, creating a baseline by which potential solutions can be compared.

Our study demonstrates that manual de-identification of narratives based on HIPAA standards is a difficult and time-consuming task. Despite careful training and a small corpus of notes to review, individual manual reviewers achieved only 96% recall (or sensitivity). Simple automated, pattern-matching search-and-replace techniques did not perform as well as human reviewers in this dataset, raising the possibility that naive investigators should pause before adopting this strategy. Since HIPAA requires removal of all PHI, our findings suggest new techniques will be necessary for the typical researcher to meet the HIPAA mandate for de-identification. Future directions for de-identification may include the use of demographic meta-data (e.g., the name of the patient and provider) associated with each narrative note to assist with auto-

mated removal strategies or complex natural language software specifically trained in the medical domain.

Finally, the time costs were almost 90 seconds per note. To analyze the entire master research corpus (88,000 notes) would, by extrapolation, take over 58 reviewer-weeks of time under absolutely ideal conditions. Researchers should be aware that there may be significant costs associated with manually de-identifying large data sets.

### Acknowledgments

D. A. Dorr and S. A. Sims were supported in part through a grant from the National Library of Medicine (Training Grant 5T15LM007124-07); S. S. Phansalkar and J. F. Hurdle were supported in part from by the Foundation of the American Society of Health-systems Pharmacists (ASHP) and VA HSR&D grant TRP 02-147. W. Phillips received partial support through the National Library of Medicine's informatics summer internship program. The authors are grateful for the data entry and review by Jared Anderson.

## References

1. Chute CG, et al. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. *J Am Med Inform Assoc* 1998; 5 (6): 503-10.
2. NIH Draft Statement on Sharing Research Data. Bethesda, MD: NIH; 2002 [accessed April 24, 2005]. Available from: <http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html>.
3. Berman JJ. Confidentiality issues for medical data miners. *Artif Intell Med* 2002; 26 (1-2): 25-36.
4. Initiative on Privacy Standardisation in Europe (IPSE): Final Report. Brussels, Belgium: CEN/ISSS; 2002 [Accessed Aug 25, 2005]. Available at: <http://www.cenorm.be/cenorm/businessdomains/businessdomains/iss/activity/ipsefinalreport.pdf>
5. Fielstein E, et al. Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings. *Medinfo* 2004; 11 (Pt 1): 1590.
6. Thomas SM, et al. A successful technique for removing names in pathology reports using an augmented search and replace method. *Proc AMIA Symp* 2002; pp 777-81.
7. Ruch P, et al. Medical document anonymization with a semantic lexicon. *Proc AMIA Symp* 2000; pp 729-33.
8. Taira RK, et al. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp* 2002; pp 757-61.
9. Security and Privacy Workgroup (WEDI). De-identification white paper (version 3.1) (monograph on the internet). Reston, VA: Workgroup for Electronic Data Interchange; 2001 /cited Aug 25, 2005): 1-12. Available from: <http://privacy.cs.cmu.edu/dataprivacy/HIPAA/SNIPdeidv31.pdf>
10. Baker S, et al. Anonymization, data-matching and privacy: a case study. Final report. Washington, DC: Steptoe & Johnson, LLP; 2003.
11. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp* 1996; pp 333-7.
12. Quantin C, et al. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods Inf Med* 1998; 37 (3): 271-7.
13. Quantin C, et al. Decision analysis for the assessment of a record linkage procedure: application to a perinatal network. *Methods Inf Med* 2005; 44 (1): 72-9.
14. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. *Arch Pathol Lab Med* 2003; 127 (6): 680-6.
15. Ohno-Machado L, et al. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. *Int J Med Inform* 2004; 73 (7-8): 599-606.
16. Berman JJ. Zero-check: a zero-knowledge protocol for reconciling patient identities across institutions. *Arch Pathol Lab Med* 2004; 128 (3): 344-6.
17. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. *Proc AMIA Annu Fall Symp* 1997; pp 51-5.
18. Gupta D, et al. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol* 2004; 121 (2): 176-86.
19. Chinchor N. Overview of MUC-7/MET-2 [monograph on the internet]. Science Applications International Corporation; 1998 [accessed Aug 25, 2005]. Available at: [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/proceedings/muc\\_7\\_proceedings/overview.html](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html)
20. Hirschman L, et al. Rutabaga by any other name: extracting biological names. *J Biomed Inform* 2002; 35 (4): 247-59.
21. Nebeker J, et al. High rates of adverse drug events in a highly computerized hospital. *Arch Intern Med* 2005; 165 (10): 1111-6.
22. Bruce R, Wiebe J. Recognizing Subjectivity: A case study of manual tagging. *Natural Language Engineering* 1999; 5 (2): 1-16.
23. Streiner D, Norman G. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 2nd ed. New York: Oxford University Press; 1995.

### Correspondence to:

David A. Dorr, Assistant Professor  
OHSU  
3181 SW Sam Jackson Park Rd.  
Mailcode BICC  
Portland, OR 97239  
USA  
E-mail: [dorrd@ohsu.edu](mailto:dorrd@ohsu.edu)