

*Technical Brief* ■

# Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes?

FRANCES P. MORRISON, MD, MPH, MA, LI LI, MS, ALBERT M. LAI, PhD, GEORGE HRIPCSAK, MD, MS

**Abstract** Electronic clinical documentation can be useful for activities such as public health surveillance, quality improvement, and research, but existing methods of de-identification may not provide sufficient protection of patient data. The general-purpose natural language processor MedLEE retains medical concepts while excluding the remaining text so, in addition to processing text into structured data, it may be able provide a secondary benefit of de-identification. Without modifying the system, the authors tested the ability of MedLEE to remove protected health information (PHI) by comparing 100 outpatient clinical notes with the corresponding XML-tagged output. Of 809 instances of PHI, 26 (3.2%) were detected in output as a result of processing and identification errors. However, PHI in the output was highly transformed, much appearing as normalized terms for medical concepts, potentially making re-identification more difficult. The MedLEE processor may be a good enhancement to other de-identification systems, both removing PHI and providing coded data from clinical text.

■ *J Am Med Inform Assoc.* 2009;16:37–39. DOI 10.1197/jamia.M2862.

## Introduction

Increasing adoption of electronic health records has intensified interest in making clinical data available for multiple purposes in addition to clinical care. Many useful data are contained in narrative clinical text, an important form of clinician communication. Natural language processing (NLP) systems have the potential to increase the usefulness of clinical text for multiple purposes, such as public health, research, and quality improvement, by transforming a large amount of text into computable data in a rapid and automated way. This has been demonstrated in adverse event detection,<sup>1</sup> surveillance during large events,<sup>2</sup> and detecting smoking status.<sup>3</sup> Data contained in clinical text may offer more information than is available in structured data such as lab results or billing codes. This includes data required for automated notifiable disease reporting, determining adherence to quality measures, and detecting risk factors for disease. Various de-identification systems, some of which use NLP techniques, have shown promise; most of them focus not on providing structured data from text but on replacing or removing identifying information while keeping the notes otherwise intact.<sup>4</sup> De-identification strategies have included pattern recognition or rules,<sup>5</sup> conditional

random fields,<sup>6</sup> and support vector machines.<sup>7</sup> These have been successful to varying degrees, with the one of the best performing systems missing 0.18% of PHI.<sup>8</sup>

To use a physiologic metaphor, many of the de-identification methods published thus far function like the liver, where specific toxins are recognized and filtered out of the blood; the systems leave all clinical data and remove or replace only PHI. Another option is to use a system that functions more like a kidney, which allows all blood into renal tubules and then selectively reclaims required components. The analogous process for notes is to recognize and retain only the important information and discard the rest, presumably including the PHI. Used in this way, NLP has the potential to offer more descriptive, detailed information than billing data or chief complaints while reducing the likelihood of patient identification.

The MedLEE processor has proved its usefulness as a general purpose NLP system in a variety of clinical settings and document types including radiology reports, discharge summaries, and visit notes with good performance in identifying concepts.<sup>9–11</sup> We were interested in determining whether we could use MedLEE to provide some amount of de-identification of clinical text. The PHI may be retained if names match terms in the MedLEE lexicon (a collision) or if information is incorrectly processed and retained as another type of data that is unprotected. Our goal is to determine how often this happens and how. If MedLEE can both remove PHI and provide structured data from text, it may be a good tool to complement existing de-identification systems for the many activities that require the use of large amounts of clinical text.

## Methods

We used electronic outpatient clinical follow-up notes written between November 2004 and April 2005 by internal medicine practitioners, including oncology and neurology

Affiliation of the authors: Columbia University Department of Biomedical Informatics, New York, NY.

Research for natural language processing and continuing development of MedLEE supported by R01 LM007659 and R01 LM008635 from the National Library of Medicine.

Research for evaluation of MedLEE as a de-identification and syndromic surveillance tool supported by RO1 LM06910 and PO1 HK000029 from the Centers for Disease Control and Prevention.

Correspondence: Frances Morrison, 622 West 168th Street, Vanderbilt Clinic, 5th Floor, New York, New York 10032; e-mail: <[frances.morrison@dbmi.columbia.edu](mailto:frances.morrison@dbmi.columbia.edu)>.

Received for review: 05/16/08; accepted for publication: 09/30/08.

subspecialists. Notes are unstructured, and their format ranges from typical outpatient SOAP notes to letters providing results of a referral. After preprocessing, which involved approximately 10–15 hours of programming to extract the text from the database and remove unrecognizable characters, the text was run through MedLEE to obtain output containing only parsed concepts tagged with XML.

A board-certified Preventive Medicine physician who has formal training in Public Health and Biomedical Informatics manually reviewed the notes and output. The PHI in the notes was characterized and summed by the eight types of PHI identified by Uzuner et al.<sup>4</sup>: patient name, clinician name, hospital, identifiers (e.g., social security numbers, medical record numbers), date (except for year), location, phone number, and age >89.

We sought to identify processing errors that allowed PHI into the output in any form by comparing 100 of the original notes with the corresponding XML-tagged output. We treated first and last names as separate units but locations and hospital names as a single unit. If any part of an identifier was allowed into output, we considered it an error. For example, if the apartment number of an address was erroneously allowed into output as a lab result value, it counted as a PHI leak.

We calculated the proportion of PHI in the original notes that ended up in the output. This is equivalent to the false negative rate (1–sensitivity) of a de-identification system (i.e., the PHI that is not identified and is therefore inappropriately left in the note). There was no analog for specificity in the experiment because MedLEE was not specifically identifying PHI. However, because of the importance of excluding names from output, we compared the 1000 most common names (first and last) in the CUMC patient name database to the MedLEE lexicon to estimate of the level of name collisions.

## Results

Of the 100 outpatient notes examined, most (81) were initial and follow-up notes from general internal medicine clinics. The remaining 19 notes originated from other internal medicine subspecialties, including infectious disease, oncology, hematology, neurology, and cardiology. Three of those documents were in the form of letters reporting the findings of an outpatient consult; the rest were clinic notes.

Detailed manual review of the notes began with summarization of the type and frequency of the various types of PHI found in the notes as seen in Table 1. After comparing each note with the corresponding XML tagged output, we found 26 pieces of PHI that slipped through and appeared in MedLEE's output. Most errors (21, or 81%) were due to collisions with medical terms; the remainder of errors resulted from the misclassification of numbers, mainly ages. Errors resulted in transformation of PHI into a variety of MedLEE data types, including: *Disease/syndrome* (5), *Measurement* (7), *Finding* (5), *Lab test* (2), *Medication* (1), and *Procedure* (1). Table 2, providing specific examples of input phrases and tagged output, is available as a supplement at [www.jamia.org](http://www.jamia.org).

Certain abbreviations associated with PHI proved difficult for MedLEE to handle. The MedLEE processor misinter-

**Table 1** ■ Number of Various Types of PHI Misclassified as Non-PHI by MedLEE and Retained in Output

PHI Type	Instances of PHI in the Notes	Instances of PHI Allowed into Output	Proportion of PHI that Leaked Through
Age >89	7	5	71%
Clinician	157	6	3.8%
Date	300	0	0.0%
Hospital	100	7	7.0%
Location	45	3	6.7%
Patient	126	4	3.2%
Telephone	33	1	3.0%
Identifiers	41	0	0.0%
Total	809	26	3.2%

PHI = protected health informaion.

preted the abbreviation “*st*,” seven times. In these instances, *st* denoted either *Street* or *Saint* (as a part of a hospital name) in the notes but was interpreted as an ST segment measurement (as on an EKG); this caused retention of part of the address as a finding related to a ST segment. As there were 22 occurrences of “*st*” in the context of street or hospital name, nearly one third of these were parsed incorrectly (only one instance occurred in the context of “ST segment”).

Processing of the 1000 most common names in the patient database resulted in 48 names (5%) that were normalized to medical terms, including problems, findings, medications, and lab tests. Examples include colors that could be taken as a physical finding, such as “*Green*” and “*Brown*,” abbreviations for lab tests and eponymous disease names, such as the surname “*Diaz*” interpreted as “*Diaz Disease*” (osteochondritis of the talus), as well as common English terms such as “*Rose*” meaning “increased.”

## Discussion

We were able to use an existing NLP system without modification to process a heterogeneous set of outpatient clinical notes into XML-tagged clinical data with a moderate level of de-identification, with 3.2% of total PHI allowed into output. Errors were mainly due to collisions between names of people and places with medical concepts or English terms. Errors were also a result of misclassification of ages as quantities or measurements (such as lab values).

Large numbers of clinical documents are likely to be used in the future for quality improvement, public health surveillance, and research; specific activities may include performing automated reporting using text, quantifying guideline adherence, implementing quality measures, and detecting adverse events. Using these clinical data for a multitude of purposes increases the likelihood that PHI will leave the protection of a health care facility and transfer to a variety of institutions, such as health departments, research facilities, and other organizations focused on quality and safety. One of the best performing de-identification systems that leave text intact did not recognize approximately 0.18% of PHI;<sup>8</sup> this means that in a practice that produces 2000 notes per month of 250 words per note with 4% being PHI, approximately 450 pieces of PHI could be missed per year in one practice alone. Improvement is likely necessary, but continu-

ing to perfect these existing systems may not be the optimal solution. It may be necessary to approach de-identification by using several methods to complement each other. If the ultimate goal is producing useful de-identified data from clinical text, then combining a traditional de-identification system with MedLEE may afford a solution.

Pipelining two systems that use different strategies in a series may produce better results than achieved by either alone. For example, one could use a system that tags potential PHI, followed by MedLEE processing. This strategy would have the advantage of transforming text to structured data, although how the systems interact is untested. If two systems were used in series, higher PHI removal with may result; by processing PHI differently, each system may catch identifiers that the other misses. We did not estimate the possible impact on the text processing performance of MedLEE due to misclassifying PHI, but implementing this strategy may reduce the problem. For example, if ages are marked as such, then MedLEE will not misinterpret it as a laboratory value. The MedLEE processor does have an advantage over other systems in its ability to convert text to computable data; it has proven its usefulness in other contexts and is likely to perform similarly in the future.

The rate of PHI that was allowed into output using MedLEE is higher than other systems but the PHI that remains in output is often transformed, with the actual text changed to normalized medical terms. The processing errors caused by MedLEE are of a different variety than the type of error that occurs in a system that removes identifiers but leaves the text intact. The context of the PHI may be important; PHI that remains in its original text may have a higher potential for identification than a piece of data that has been tagged incorrectly in structured output. It is also likely true that all identifiers are not equal—allowing a lab date test date into output is very different than a patient's last name, but for ease of quantification, these are considered the same. Regardless, we have learned the important lesson that MedLEE output is not necessarily de-identified and should not be treated as such.

We have demonstrated that an existing NLP system can de-identify clinical notes to some degree with the same tagged, structured output that has demonstrated utility in other contexts. The combination of de-identification of PHI

with identification of medical concepts may be useful in a variety of activities, such as research, quality improvement, and public health, or any other task which requires a large amount of detailed clinical data. In the future, we would like to improve the system to reduce the types of errors that allow PHI in output and test out the performance of MedLEE when used in conjunction with an existing de-identification system.

#### References ■

1. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc.* 2005;12(4):448–57.
2. Gundlapalli AV, Olson J, Smith SP, Baza M, Hausam RR, Eutropius LJ, et al. Hospital electronic medical record-based public health surveillance system deployed during the 2002 Winter Olympic Games. *Am J Infect Control.* 2007;35(3):163–71.
3. Heinze DT, Morsch ML, Potter BC, Sheffer RE, Jr. Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. *J Am Med Inform Assoc.* 2008;15(1):40–3.
4. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc.* 2007 Sep–Oct;14(5):550–63.
5. Gupta D, Saul M, Gilbertson J. 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.
6. Aramaki E, Miyo K. Automatic Deidentification by Using Sentence Features and Label Consistency. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data; 2006.*
7. Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple R. Identifying Personal Health Information Using Support Vector Machines. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.*
8. 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 Sep–Oct; 14(5):574–80.
9. Barrows Jr RC, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Am Med Inform Assoc Proc.* 2000:51–5.
10. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Am Med Inform Assoc Proc.* 1999:256–60.
11. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology.* 2002 Jul;224(1):157–63.



## Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes?

Frances P Morrison, Li Li, Albert M Lai, et al.

*J Am Med Inform Assoc* 2009 16: 37-39

doi: 10.1197/jamia.M2862

---

Updated information and services can be found at:

<http://jamia.bmj.com/content/16/1/37.full.html>

---

*These include:*

### Data Supplement

*"Data Supplement"*

<http://jamia.bmj.com/content/suppl/2009/11/20/16.1.37.DC1.html>

### References

This article cites 7 articles, 2 of which can be accessed free at:

<http://jamia.bmj.com/content/16/1/37.full.html#ref-list-1>

Article cited in:

<http://jamia.bmj.com/content/16/1/37.full.html#related-urls>

### Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

---

### Notes

---

To request permissions go to:

<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:

<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:

<http://group.bmj.com/subscribe/>