Hiding in Plain Sight: Use of Realistic Surrogates to Reduce Exposure of Protected Health Information in Clinical Text

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

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

Appendix A:
IRB Survey Regarding HIPS De-identification

This appendix reports the results of a Web-based survey of Institutional Review Board (IRB) managers in the HMO Research Network regarding clinical text de-identification practices and attitudes toward the Hiding In Plain Sight (HIPS) approach. It includes a brief summary of the survey methodology followed by a summary of responses to each survey question.

Survey Methods

To preliminarily assess attitudes toward the Hiding in Plain Sight (HIPS) approach to clinical text de-identification we surveyed Institutional Review Board (IRB) and Human Subjects Review Committee (HSRC) managers from institutions affiliated with the HMO Research Network (HMORN). Sixteen managers were invited to respond to a ten-question Web-based survey. The HMORN is a network of research institutions that individually and collectively conduct a wide range of health-related studies using patient data, some of which include clinical text, extracted from the electronic medical record systems of their affiliated care delivery organizations. Survey questions addressed local methods used to de-identify text, and elicited managers' opinions about the acceptability of automated methods to de-identify clinical text, including the HIPS approach. The HIPS approach was described and illustrated by an example similar to that provided in Figure

1 of the manuscript. Each potential respondent received one email invitation containing a link to an anonymous online survey, one follow up voice mail reminder, and two additional follow-up e-mail reminders encouraging participation in the survey. The full text of the survey and tabulated results follow. Eight of the managers responded to the survey, for a response rate of 50%. The percentage of respondents completing all questions was 100% (8/8).

A summary of survey responses follows.

Summary of Survey Responses

The full text of survey questions and a summary of individual responses to each question, including “Other (please specify)” responses, is provided below.

Question 1 of 10

This question was used for planning purposes. It is unrelated to the substantive issues addressed in the survey and has been omitted.

Question 2 of 10

2. On average HOW OFTEN do projects at your site de-identify clinical text for research purposes? Create Chart Download

	Response Percent	Response Count
Never	0.0%	0
Less than 1 project per year	0.0%	0
About 1 project per year	0.0%	0
2-3 projects per year	14.3%	1
More than 3 projects per year	57.1%	4
(Not sure/No answer)	28.6%	2
	Other (please specify) Show Responses	1
	answered question	7
	skipped question	1

Question 2 “other” responses:

- I'm not sure but I don't think it is common. HIPAA waiver or LDS more often used.

Question 3 of 10

3. What METHODS are currently used at your site to de-identify clinical text? CHECK ALL THAT APPLY		Response Percent	Response Count
Manual removal of PHI with at least 1 reviewer		12.5%	1
Manual removal of PHI with at least 2 reviewers		25.0%	2
Manual removal of PHI with 3 or more reviewers		0.0%	0
Automated software		25.0%	2
A combination of manual and automated methods		50.0%	4
(Don't know/not sure)		25.0%	2
Other (please specify) Show Responses		12.5%	1
		answered question	8
		skipped question	0

Create Chart Download

Question 3 “other” responses:

- For many studies, we have HIPAA-trained investigators query clinical datasets for specific fields of interest, so as to create a de-identified "research data set." This data set is all that the investigator works with. I am not sure if this is "manual" or "automated," but in any event, no paper records are involved.

Question 4 of 10

4. When sharing de-identified clinical text with outside collaborators do you typically use DATA USE AGREEMENTS (DUAs)?		Response Percent	Response Count
No		12.5%	1
Yes, usually		12.5%	1
Yes, always		37.5%	3
(Don't know/not sure)		12.5%	1
Other (please specify) Show Responses		25.0%	2
		answered question	8
		skipped question	0

Create Chart Download

Question 4 “other” responses:

- It depends on the source and the nature of the data. Outside collaborators do not ever see raw clinical data, of course, and many only are involved in data analysis (they do not even see the de-identified data set).
- It's required, but I don't know that it's executed.

Question 5 of 10

5. Please indicate your level of agreement or disagreement with this statement: "Most IRBs WOULD APPROVE use of AUTOMATED de-identification software as a substitute for manual approaches if the automated system REMOVED AT LEAST 99% OF PHI, leaving behind at most 1% of PHI."

Create Chart | Download

	Response Percent	Response Count
Strongly DISAGREE	0.0%	0
Disagree	37.5%	3
Neither agree nor disagree	25.0%	2
Agree	12.5%	1
Strongly AGREE	12.5%	1
Other (please specify) Show Responses	12.5%	1
	answered question	8
	skipped question	0

Question 5 “other” responses:

- Can't speak for most IRBs

Question 6 of 10

6. In your opinion, what is the SINGLE MOST COMPELLING REASON for using MANUAL approaches to de-identify clinical text?		Create Chart	Download
		Response Percent	Response Count
Manual methods are LESS EXPENSIVE than other methods		0.0%	0
Manual methods are EASY TO UNDERSTAND & explain		25.0%	2
Manual methods are the MOST EFFECTIVE		25.0%	2
Manual methods are EASY TO IMPLEMENT		0.0%	0
(Don't know/not sure)		12.5%	1
Other (please specify) Show Responses		37.5%	3
		answered question	8
		skipped question	0

Question 6 “other” responses:

- The most compelling reason is that we do not have automatic methods for producing de-identified data. That means that the manual methods automatically become easier to implement, most effective, less expensive, etc.
- Accountability trail
- Lack of training on electronic systems for de-identification

Question 7 of 10

7. In your opinion, what is the SECOND MOST COMPELLING reason for using MANUAL approaches to de-identify clinical text?		Create Chart	Download
		Response Percent	Response Count
Manual methods are LESS EXPENSIVE than other methods		0.0%	0
Manual methods are EASY TO UNDERSTAND & explain		12.5%	1
Manual methods are the MOST EFFECTIVE		0.0%	0
Manual methods are EASY TO IMPLEMENT		37.5%	3
(Don't know/not sure)		25.0%	2
Other (please specify) Show Responses		25.0%	2
		answered question	8
		skipped question	0

Question 7 “other” responses:

- See above
- Limits volume of materials released

Question 8 of 10

The HIPS Method:

There is a method of de-identifying clinical text called the "Hiding In Plain Sight" (HIPS) method. Instead of removing PHI, HIPS replaces it with realistic-looking "fake" PHI. This fake or "surrogate" PHI resembles real patient information but does not represent any actual patient.

When ~95% of real PHI in a document is replaced via the HIPS method, it is hypothesized that the remaining ~5% cannot be distinguished from the 95% that is fake. If this is true, the real PHI is said to be "hiding in plain sight."

Illustration:

Real clinical text containing REAL PHI might look like this:

==========
LINE 01: James Smith is a 61 year old ...
LINE 12: at the request of Dr Mary Johnson ...
LINE 19: PLAN: Dr Williams and I discussed ...
LINE 26: reviewed with Mr. Smith on 5/11/2010 ...
LINE 29: Mr. Smith has a consultation ...
==========

A HIPS de-identified version of the above is shown below, where all PHI except the name "Williams" (LINE 19) has been replaced by fake/surrogate PHI:

==========
NOTE: PERSONAL IDENTIFYING INFORMATION IN THIS
DOCUMENT HAS BEEN REPLACED WITH REALISTIC
APPEARING SURROGATES. USE OF THIS DOCUMENT IS
SUBJECT TO THE TERMS OF A DATA USE AGREEMENT.

LINE 01: Robert Davis is a 63 year old ...
LINE 12: at the request of Dr Patricia Brown ...
LINE 19: PLAN: Dr Williams and I discussed ...
LINE 26: reviewed with Mr. Davis on 4/27/2010 ...
LINE 29: Mr. Davis has a consultation ...
==========

In the above example of the HIPS method, the name "Dr Williams" is "hiding in plain sight." All other PHI has been replaced with fake/surrogate PHI.

Please consider this example of the HIPS method when answering the following questions.

Question 8 of 10 (continued)

8. Have you ever heard of the HIPS method of de-identification (described above)?		Create Chart	Download
		Response Percent	Response Count
No		75.0%	6
Yes		25.0%	2
(Don't know/not sure)		0.0%	0
		answered question	8
		skipped question	0

Question 9 of 10

9. Please indicate your level of agreement or disagreement with this statement: "If rigorous peer-reviewed research based on actual clinical text showed that trained chart abstractors were UNABLE TO IDENTIFY "hiding" PHI in documents de-identified by the HIPS method, most IRBs would CONSIDER APPROVING THE HIPS METHOD for research purposes when accompanied by standard data use agreements." Create Chart Download

	Response Percent	Response Count
Strongly DISAGREE	0.0%	0
Disagree	12.5%	1
Neither agree nor disagree	0.0%	0
Agree	37.5%	3
Strongly AGREE	12.5%	1
(Don't know/not sure)	25.0%	2
Other (please specify) Show Responses	12.5%	1
	answered question	8
	skipped question	0

Question 9 “other” responses:

- I think this is great. I am not sure that most IRBs would think it is HIPAA compliant to purposefully leave some PHI in place.

Question 10X of 10

10. Please indicate your level of agreement or disagreement with this statement: "If rigorous peer-reviewed research based on actual clinical text showed that clinical text de-identified by the HIPS method presented LESS RISK TO PATIENT PRIVACY THAN MANUALLY DE-IDENTIFIED TEXT, most IRBs would consider approving the HIPS method for research purposes when accompanied by standard data use agreements." Create Chart Download

	Response Percent	Response Count
Strongly DISAGREE	0.0%	0
Disagree	0.0%	0
Neither agree nor disagree	12.5%	1
Agree	50.0%	4
Strongly AGREE	12.5%	1
(Don't know/not sure)	12.5%	1
Other (please specify) Show Responses	12.5%	1
answered question		8
skipped question		0

Question 10 “other” responses:

- They would consider, but whether they would approve depends on other protections, size & scope of release, etc.