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*
[3]*Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA*
[4]*The MITRE Corporation, Bedford, MA, USA*

Appendix B:

Clinical Text De-identification Annotation Guidelines

This appendix contains the written guidelines used to create the gold standard clinical corpora used in the Hiding In Plain Sight (HIPS) pilot experiments.

Introduction

These guidelines provide detailed information about identifiers that should and should not be annotated for the purpose of de-identifying clinical text. While they are intended to be exhaustive, other identifiers not unanticipated by these guidelines may be found. Annotators are encouraged to contact the project manager with any questions about possible identifiers not covered by these guidelines.

These guidelines also describe the process by which documents are subjected to independent, successive reviews until the steps for achieving de-identification according to this protocol are completed. As noted below, each person annotating documents should never review a document more than once.

Annotation Guidelines

Identifiers to be annotated

Identifying information found in clinical notes that should be annotated include—but are not necessarily limited to—the following.

1. First and last names of patients, health proxies, family members and other non-practitioner persons mentioned. Honorifics and titles should not be annotated as part of a name. For example, "Mr.," "Mrs.," "Dr.," and "ARNP" should not be annotated. When a name is represented as "Lastname MD, Firstname," the "Firstname" and "Lastname" should be annotated as two different annotations, and "MD" should not be annotated. When a name is represented as "Lastname, Firstname" the entire string should be annotated as a single name annotation.

2. Ages of all people that indicate current age or which can be use to calculate a current age given other information in the document. For example, in "He was age 40 in 2009" the age "40" would be annotated as an age and the year "2009" would be annotated as a date.

3. Geographic locations including country names, addresses, street names, state names, zip codes, and other geographic identifiers.

4. All dates or elements of dates, including year. These include birth dates, admission dates, discharge dates, death dates, and dates of medical or other events (e.g., in "surgery in 2010" the date "2010" would be annotated as a date).

5. Phone numbers.

6. Fax numbers.

7. Electronic mail addresses.

8. Social Security numbers.

9. Medical record numbers.

10. Health plan beneficiary numbers.

11. Account numbers.

12. Certificate/license numbers.

13. Vehicle identifiers and serial numbers, including license plate numbers.

14. Device and equipment serial numbers.

15. Web Universal Resource Locators (URLs).

16. Internet Protocol (IP) address numbers.

17. Clinical trial names and numbers. For example, "SWOG 2217," which references a particular clinical trial protocol, should be marked with an other-identifier annotation.

18. Any other unique identifying number, characteristic, or code. For example, in "CLIA-88 certification number 123456789" the certification number must be annotated.

19. Institution and organization names, including health care organizations such as "Mass General Hospital" (including the word "hospital"), company names such as "Lockheed Martin," and other institution names, even if they are mentioned unrelated to a place of care. For example, in "Patient's son graduated from UW" the abbreviation "UW" should be annotated as an institution name. Always annotate the full proper name of the institution, but only if it has the appropriate title case (e.g., annotate all of "Roosevelt High School," but only annotate "Liberty" in "... she attends Liberty school."

20. Practitioner and other health care personnel names. Transcription initials and code numbers should also be annotated. For example, a transcription signature such as "ABC:DEF 12345" should receive two practitioner name annotations (one for "ABC" and another for "DEF") and an other-identifier annotation for the unique transcription number "1235."

21. Any extremely rare attribute of a person, including things like "retired chair of OB/GYN department at university," and "immigrant from Tierra del Fuego," and "pitcher for the Red Sox." Marking such identifiers with the other-identifier annotation.

<u>Entities *not* to be annotated</u>

The following entities should not be annotated for de-identification.

a. Times of day should not be annotated.

b. Periods of time associated with treatment plans (e.g., "follow up in 8 months") should not be annotated.

c. Public entities like "US Food and Drug Administration" should not be annotated, but entities like "University of Washington" are because they identify a place where care may have been received or a patient may work or be enrolled.

d. Common suppliers of medical equipment (e.g., "Boston Scientific") should not be annotated.

e. Model numbers of devices because these are not unique to patients receiving them. However, do annotate equipment serial numbers as noted above.

f. "CLIA-88," which is a generic type of certification level for a lab facility, should not be annotated.

g. Race and/or ethnicity (of anybody) should not be annotated.

h. Number of children in family should not be annotated.

i. References to years married or years employed should not be annotated, unless the number of years is large and indicates the person is likely to be more than 80 years old, or unless other information in the same note makes it possible to calculate age exactly (as noted above).

j. Temporally indefinite ages, which are ages that cannot be translated into current age (e.g., "Father died when the patient was 16.") should not be annotated.

k. Height and weight should not be annotated.

Multiple Review Process

Documents will be considered de-identified when they have completed all stages of the review process described below. All personnel participating in the annotation process will have received training in the proper use of the annotation guidelines. Each annotator will review a document at most one time.

<u>Stage 1</u>: Documents will be randomly assembled into sets of at least 5 documents.

<u>Stage 2</u>: All documents within a set (defined by Stage 1) will be reviewed by an annotator who will mark all identifiers found in the document according to the above Annotation Guidelines. Any questions the annotator may have will be discussed and resolved with a project manager and/or study investigator.

<u>Stage 3</u>: Sets of documents produced by Stage 2 will be reviewed by another annotator who has not previously reviewed any of the documents in the set. Annotations made in the prior stage will be clearly visible in the documents. The Stage 3 reviewer will apply the same Annotation Guidelines described above to mark any additional identifiers overlooked by the prior stage of review.

- If no additional ("new") identifiers are found in Stage 3 a label will be added to the document set that indicates a "clean review" has been completed.

- If *any* additional identifiers are discovered *any* of the documents in the set no label will be added to the document set.

Stage 4: Sets of documents produced by Stage 3 will be reviewed by another annotator who has not previously reviewed any of the documents in the set. Annotations made in the prior stage will be clearly visible in the documents. The Stage 4 reviewer will apply the same Annotation Guidelines described above to mark any additional identifiers overlooked by the prior stage of review.

- If no additional ("new") identifiers are found in Stage 4 a label will be added to the document set that indicates a "clean review" has been completed. If this is the second "clean review" added to the document set the document set is considered de-identified and will receive no further review

- If *any* additional identifiers are discovered *any* of the documents in the set no label will be added to the document set.

Stages 5-N: Stages 5-N represent additional review stages that each repeat the process described in Stage 4 until all sets of documents contain two "clean review" labels, at which point the entire corpus is considered de-identified.