

Annotation guidelines for de-identification of medical records

Amber Stubbs and Ozlem Uzuner

Last updated: April 1, 2014

TASK DESCRIPTION

HIPAA requires that patient medical records have all identifying information removed in order to protect patient privacy. There are 18 categories of Protected Health Information (PHI) identifiers of the patient or of relatives, employers, or household members of the patient that must be removed in order for a file to be considered de-identified.

In order to de-identify the records, each file must have the PHI marked up so that it can be removed/replaced later. This will be done using the annotation software MAE, and all PHI will be given an XML tag indicating its category and type, where applicable.

For the purposes of this annotation project, the 18 HIPAA categories have been grouped into 6 main categories and 25 sub categories:

PHI category	Sub-category
NAME	PATIENT, DOCTOR, USERNAME
PROFESSION	(none)
LOCATION	HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER
AGE	(none)
DATE	(none)
CONTACT	PHONE, FAX, EMAIL, URL, IPADDRESS
IDs	SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER

Table 1: i2b2 annotation categories

ANNOTATION NOTES

Names:

- annotate names/initials at end of documents-- these belong to people who modified the file
- Titles (Dr., Mr., Ms., etc.) do not have to be annotated.
- Information such as "M.D.", "R.N." do not have to be annotated
- If a name is possessive (e.g., Sam's) do not annotate the 's
- USERNAME category be used for PHI that are initials followed by numbers (i.e., as4)

Profession:

- any job that is mentioned that is not held by someone on the medical staff should be tagged

Dates:

- Any calendar date, including years, seasons, months, and holidays, should be annotated
- Days of the week should also be annotated
- Do not include time of day
- If the phrase has 's (i.e., "in the '90's), annotate "'90's"
- Include annotations of seasons ("Fall '02")
- Include quote marks that stand in for years ("92")

Locations:

- annotate state/country names as well as addresses and cities. Each part of an address will get its own tag. For example:
32 Vassar Street -- Street
Cambridge -- City
MA -- State
02142 -- ZIP
USA -- Country
- Generic locations like "hair salon" do not need to be annotated, but named organizations (i.e., "Harvard University") do

Age:

- annotate all ages, not just those over 90, including those for patient's families if they are mentioned

Contact:

- pager numbers should be annotated as phone numbers

IDs:

- If you aren't sure what type of an ID something is, use "IDNUM"
- No need to label names of devices (for example: "25 mm Carpentier-Edwards magna valve", "3.5mm by 32mm Taxus drug-eluting stent", Angioseal"). The "device id" is only for numbers that uniquely identify exactly **which** of a device was used, not the brand name.

QUESTIONS

If you have questions, please contact Amber Stubbs (astubbs@albany.edu)