Guidelines for Annotating Personal Identifiers in the Clinical Text Repository of the National Institutes of Health

Mehmet Kayaalp, MD, PhD
Pamela Sagan, RN
Allen C. Browne, MS
Clement J. McDonald, MD

June 28, 2016

Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
8600 Rockville Pike, Bethesda, Maryland 20894

Table of Contents

1 Overview 1

2 Personal Names 2

3 Personal Name Initials 4

4 Organization Names and Organizational Units 5

5 Occupation Types 9

6 Telecommunication Identifiers 12

7 Address Identifiers 13

8 Date Identifiers 16

9 Age Identifiers 18

10 Time Identifiers 21

11 Numeric and Alphanumeric Identifiers 22

12 Personally Identifying Contexts 25

13 Identified Person’s Kinship or Relations to the Patient 26

14 Non-identifying, Non-specific Text Parts 27

15 Non-Identifying Medical Terms 28

16 Discussion 31

17 Conclusions 31

18 Future Work 32

Acknowledgements 33

References 33

1 Overview

The annotation system we developed comprises 12 distinct identifier categories: **PersonalName** (personal name), **PNInit** (personal name initials), Organization, **Occupation**, **Telecom** (telecommunication identifier), Address, Date, Age, **Time**, Numeric and Alphanumeric identifiers, **PIC** (Personally Identifying Context) and **Role** and two non-identifier categories: Non-Identifying Text (**NPII**) and Non-Identifying Medical Terms (**Anatomy**, **Device**, **Diagnostics**, **Therapy** and **Eponym**). While some categories such as **PersonalName** are actual labels (noted in bold text), others such as Address are not labels themselves but are associated with a set of more granular labels, which altogether represent that category (see Table 1).

Table 1 Categories and their Labels

Category	Labels
Personal Name	**PersonalName**
Personal Name Initials	**PNInit**
Organization	**Organization, Unit**
Occupation	**Occupation**
Telecom	**Telecom**
Address	**Street, Building, Location, City, County, State, Zip, Country**
Date	**Year, Month, Day, DayOfWeek, SpecialDay**
Age	**AgePII, AgeFraction, AgeNPII, AgePast**
Time	**Time**
Numeric & Alphanumerical IDs	**MedicalRecordNo, HealthRecordID, ProtocolID, AlphanumericID**
Personally Identifying Context	**PIC**
Role	**Role**
Non-Identifying Text	**NPII**
Non-Identifying Medical Terms	**Anatomy, Device, Diagnostics, Therapy, Eponym**

Address is a category comprising eight labels, each of which denotes a part of an address: **Street** (e.g. *Rockville Pike*), **Building** (e.g., *Building 38A*), **Location** (e.g., *Room 7S714*), **City** (e.g., *Bethesda*), **County** (*e.g., Montgomery County*), **State** (e.g., *Maryland*), **Zip** (e.g., *20895*), or **Country** (e.g., the *United States*).

Date is a category comprised of the following five labels: **Year** (e.g., *2016*), **Month** (e.g., *February*), **Day** (e.g., *31*), **DayOfWeek** (e.g., *Monday*), and **SpecialDay** (e.g., *Cinco de Mayo*).

Age is a category comprised of the following four labels: **AgePII** (e.g., *95* year-old), **AgeNPII** (e.g., *70* year-old), **AgeFraction** (e.g., *2 month* old) and **AgePast** (e.g., when he was *5* years old).

Numeric and alphanumeric identifiers is the category that is comprised of the following four labels: **MedicalRecordNo**, **HealthRecordID**, **ProtocolID** and **AlphanumericID**, denoting medical record and account numbers, additional identifiers unique to the patient assigned by the health care organization, research protocol numbers and other numeric or alphanumeric identifiers, respectively.

Organization is a category that is comprised of two labels: **Organization** and **Unit**, which denote organization names (e.g., *Lister Hill National Center for Biomedical Communications*) and facility names within organizations (e.g., *Cognitive Science Branch*), respectively.

We associate almost every annotated identifier with a person type. Person types include **Patient**, the patient's **Relative** (which includes members of the patient's household as well as members of a formal union and/or relationship, and neighbors), the patient's **Employer** (which includes coworkers of the patient), health care **Provider** (which includes non-relative home caretakers, who are neither household members nor medical staff) and **Other**. For example, a phrase annotated with the identifier label **Street** may refer to the address of the patient (**Street**::**Patient**), the patient's relative (**Street**::**Relative**), the patient's employer (**Street**::**Employer**) or the patient's healthcare provider (**Street**::**Provider**). If the identifier is not related to the patient directly or indirectly, we do not specify a person type; e.g., the street name in "1600 *Pennsylvania Avenue*" would be labeled as **Street::Other**.

Identifiers and personhood are (mostly) orthogonal dimensions; e.g., a personal name, address, age, organization, occupation or alphanumeric identifier can be associated with any type of person. There are, however; a few exceptions. **MedicalRecordNo**, **HealthRecordID**, and **ProtocolID** are alphanumeric identifiers which are almost always associated with the patient and are only given the **Patient** personhood. Note that although dates and times found in the text are also almost always patient specific, we included all personhoods under these labels due to the discovery of non-patient dates and times particularly found in journal publications and familial references.

Non-Identifying Text, **NPII**, has no personhood association. **Eponym**, one of the labels under the category of Non-Identifying Medical Terms, has five unique subcategories (not personhoods) associated with it: **anatomy**, **device**, **diagnostic**, **therapy** and **other**.

2 Personal Names

We assign the label **PersonalName** to an individual's name that appeared in the document. We label the entire name, excluding professional titles, unless the title identifies a small circle of people comprising less than 20,000 individuals, like "*Governor*" or "*Secretary of State*". We consider non-professional suffix titles such as "*Jr.*" and "*Sr.*", if present, as the intrinsic component of the personal name. We distinguish actual individual names (e.g., *Bill Clinton,* Dr. *Jones*' lab) from other entities such as organizations named after individuals (e.g., *Bill Clinton Foundation, The Dr. Jones' Lab*). We use the label **PersonalName** to annotate the former examples but use the **Organization** label for the latter examples and use the message "epo#" when applicable to an eponymous term.

We use the label **PersonalName::Other** for personal names that denote persons who have no direct relations with the patient (e.g., "the patient was identifying himself as *Pope Francis*").

We use the person type label **Employer** broadly. For example, if the personal name denotes a coworker, we annotate it as **PersonalName::Employer**. A coworker may not be the employer of the patient, but under certain conditions, the patient might be re-identified through the identity of his/her coworker and their employment relationships; thus, **Employer** would be the most appropriate personhood category for coworkers.

Types of Personal Names

A. Personal Name::Patient

i. **<u>Joan R. Smith</u>** is a 30-yr old female patient
ii. **<u>Gov. Jones</u>** received treatment there
iii. The **<u>Smiths</u>** drove from New Jersey to be seen

B. Personal Name::Provider

i. Operated on by Dr. **<u>James Brown, Sr.</u>**
ii. Patient seen with nurse **<u>Joan S.</u>** who translated for her
iii. Specimen was sent to Dr. **<u>Smith</u>**'s lab for testing

C. Personal Name::Relative

i. Patient accompanied by his wife, **<u>Dianne</u>**
ii. Patient's partner, **<u>Cindy</u>**, and her daughter, **<u>Katie</u>** were also present

D. Personal Name::Employer

i. Patient accompanied by his supervisor, **<u>John Smith</u>**
ii. Patient and his co-worker, **<u>P. Jones</u>**, provided the medical history

E. Personal Name::Other

i. Patient cited **<u>President Obama</u>** as our president
ii. According to **<u>Greuhl</u>** and **<u>Pyle</u>** the bone age is 12yrs
iii. He heard voices from the **<u>Virgin Mary</u>** last night
iv. Cited from Journal of Dermatology; **<u>A. Smith</u>**, PhD, MD, pps 123-456
v. Pt was able to recall **<u>Presidents Bush</u>** and **<u>Clinton</u>**

F. Examples/Discussion

i. The patient brought her dog, ~~Maggie~~.
Maggie is not an individual person so it is not tagged. Personal Name only applies to humans.

3 Personal Name Initials

The **PNInit** label denotes the full set of personal name initials such as *JFK* (for John F. Kennedy). If only a part of the full name is initialized (e.g., J.K. in "J.K. Rowling"), we do *not* use **PNInit** but use the label **PersonalName** to annotate those initials.

The vast majority of **PNInit**s that we observe in our repository are those of the providers and transcriptionists signing the file. We do not annotate numbers associated with personal name initials.

We do not annotate stand-alone punctuations but if initials are accompanied by periods (e.g., *J.F.K.*), our tokenization method would yield tokens comprising both letters and periods; thus, the tokens labeled as **PNInit** could contain those periods as well.

Types of Personal Name Initials

A. PNInit::Provider

ii. Signed by **MLK** 3/12/14 07:45am
iii. Signed by: **MLK**1
iv. Signed by **mlk**/**tk**
v. Signed **MLK**:**TK**
vi. Thank you for the consult, **J.K.** occupational therapist

B. PNInit::Patient

i. Johnny, also known as "**JS**", is a 12yr old boy

C. Examples/Discussion

i. Endoscope performed by **ENT**
ENT = Occupation::Provider
Based on the context, ENT is not a person's initials but stands for Ear, Nose and Throat specialist.

ii. Clinical history: **MDS**/**BMT**
MDS = Diagnosis; BMT = Therapy
Based on the context, MDS/BMT are not initials but are acronyms for myelodysplastic syndrome and bone marrow transplant.

4 Organization Names and Organizational Units

We define Organization as a category that is comprised of two labels: **Organization** which denotes a specifically named or identified entity (e.g. *National Institutes of Health (NIH)*) and **Unit,** which denotes a facility type within the organization (e.g. "*Pathology Department*"). Some larger organizations (e.g., *Department of Health and Human Services* (*HHS*)) constitute a number of divisions with unique names. Many times, there is an organizational hierarchy as in the case of *HHS*, *NIH*, *U.S. National Library of Medicine (NLM)*, and *Lister Hill National Center for Biomedical Communications (LHNCBC)*, which are entities with unique names. However, LHNCBC's subdivisions, *Cognitive Science Branch (CgSB)* and *Computer Science Branch (CSB)*, are units with generic and descriptive names. Thus, while we would annotate *HHS*, *NIH*, *NLM* and *LHNCBC* with label **Organization**, we should use the label **Unit** to annotate *CgSB* and *CSB*.

The reason behind this differentiation may not be apparent in the above examples, but consider the following two examples: "Pt was working in *Acme Grocery Store*" versus "Pt was working in a *sales department* (of a chain store)." Certainly, the former contains more identifying information than the latter.

When two or more organization names are seen together we annotate them collectively if they refer to a single entity and separately if they represent distinct and different entities. For example, *NIH Clinical Center* would be annotated collectively as **Organization::Provider** but *Medical Floor Day Hospital* would be annotated separately with "*Medical Floor*" as **Unit::Provider** and "*Day Hospital*" as **Organization::Provider**. Similarly, if such a name or its acronym is mentioned within a URL, we do not split the URL; rather, we label the entire URL as a telecommunication identifier.

Patients and their relatives can be identified through their organizational associations. For example, if the patient's father is known to be a member of their local Rotary Club, this piece of information would narrow the circle of potential people to identify the patient. The same is true for the information when the person is a student of a particular school. However, being a customer of a local store is not as pertinent information for identification. So, we associate organizations with the patient (or patient's relative), if he or she is a member of the organization, but do not make the same association if the relationship is less established such as customership. For example, the name of the restaurant where the patient had dined or had an accident would be labeled with the label **Organization::Other**, as would be "*Weight Watchers*" in "patient had participated in *Weight Watchers.*"

If the patient's workplace (including volunteer sites and schools where the patient is a student) is mentioned by name, we label it as **Organization::Employer**. We do not annotate projected or future organizations (e.g. "Patient graduated college and will attend VCU in the fall"). For relatives, we do not make a distinction between the employment and other types of organizational memberships. So, if the mentioned organization is the workplace of the patient's relative, we label it as **Organization::Relative**.

Specifically named doctor's offices and hospitals where the patient is admitted, cared for, discharged from or referred to, are labeled as **Organization::Provider**. This also includes personal named references associated with those organizations (e.g., patient discharged from *John James Medical Center*) where we use the messaging text of "epo#" for "*John James*" under the label **Organization::Provider** (see examples 4.D.v and 4.E.iii .) **Organization::Provider** is also used to annotate initials that clearly represent the named branch and/or institute with which the provider is affiliated that are found at the end of many reports (see example 4.D.viii).

We do not use the label **Organization** to annotate organizational types or unnamed, non-specific organizations (e.g., Police, fire station). We do not annotate health care services as in the examples "the patient is doing well with, nutrition, rehab and wound care" and "specimen sent for cytological testing" as long as those terms do not explicitly denote by whom (see 5.C.i - **Occupation::Provider**) they are provided. We do not use the label **Organization** to annotate publication titles (e.g., "American Journal of Neuroradiology") or organizational references that only refer to or are derived from that organization (e.g., "NIHFA" which stands for NIH functional assessment).

We use the label of **Unit** to denote generically named facilities and departments within a larger organization, including clinics, departments, branches and sections of hospitals or buildings that were not specifically named or further defined alphanumerically. Note that we do not annotate general areas found in all buildings such as "unit", "floor" and "room" unless they are further specified by text ("*Intensive Care Unit*" would be **Unit::Provider**) or alphanumeric characters ("*Room 123-A*" would be **Location::Provider**, see Section 7 Address Identifiers).

It can be difficult to distinguish between **Unit::Provider** and **Occupation::Provider**. Our rule of thumb is to determine whether the entity in question is an actual physical place or entity (**Unit**) or an actual person(s) or individual(s) (**Occupation**).

We rely on the context and verbiage when deciding on the correct label to use, not only with organizational terms but all annotation. We consider to what the term(s) actually refer instead of what types of references they can potentially be. For example, "patient seen by *dermatology*", "specimen sent to *dermatology*" and "patient had many dermatology tests" would be tagged **Occupation::Provider**, **Unit::Provider** and left as **NPII** respectively.

Types of Organization Identifiers

A. Organization::Patient

i. The patient is a member of the **Rotary Club**.

ii. Pt resides in **The Chase.**

B. Organization::Relative

i. The patient’s mother is a math teacher at **Herbert Hoover** (use “epo#” to denote eponym) **Middle School**.

ii. The patient’s son attended **Wake Forest University** last fall.

iii. The patient’s son is a PhD at **NIH**

C. Organization::Employer

i. The patient studies law at **George Washington** (use “epo” to denote eponym)

ii. Patient was in the **Army**

iii. Patient is a genetics research student at **NIH**

iv. The patient is employed by **Great Works & Associates**

v. Patient works at the **DC Mayor’s office**

D. Organization::Provider

i. **Maryland Pediatrics Associates**, 123 Main Street, Anywhere USA

ii. Pt was seen at **Mid-Valley Urology Center**

iii. Parents staying at **The Lodge** at **NIH** during his hospital admission

iv. Pt was treated at **NIH at Frederick** in Frederick, MD

v. Pt was referred to **St. Joseph’s** (use “epo#” to denote eponym) **Hospital**

vi. Pt stopped meds 5 weeks prior to admission to **NIH**

vii. Pt referred from **Navy** for protocol

viii. Branch/Institute: **NCI**/**NIH**

ix. Patient was treated at **National Institutes of Health Clinical Center**

E. Organization::Other

i. Patient called the **LAM Foundation** to inquire about treatment.

ii. Received packages via **FedEx**.

iii. Pt attended a fundraiser for the **Bill Clinton** (“epo#” message) **Foundation**

F. Unit::Employer

i. Patient works in the **accounting department**

G. Unit::Provider

i. Specimen sent to the **cytology department**

ii. Specimen sent to **cytology**

iii. Patient was transferred to **critical care** for this issue

iv. Patient to follow up in the **RMD Clinic**

v. Patient was seen in **Special Procedures**

vi. Location: 4 West **Cardiac Rehab**

vii. Patient went to the **emergency room** on the night of January 2

viii. Patient was transferred from the floor to the **ICU**

ix. Specimen sent to Dr. Jones' **lab**

x. Results were called to the **OR** immediately

xi. Follow up in OP-12 **Lymphoma Clinic**

H. Unit::Other

i. Pt had an episode in the **newsroom**.

I. Examples/Discussion

i. Pt seen by the **NCI** ~~team~~

NCI = Organization::Provider

Team is a group of unspecified people from the organization National Cancer Institute and is not tagged.

ii. According to ~~CDC guidelines~~

CDC guidelines is one entity, the guidelines from the CDC. This does not fall into any of our categories of identifiers so nothing is tagged.

iii. Pt followed by Dr. **Jones**' ~~team~~

Jones = PersonalName::Provider

This is a group of unspecified people who work with Dr. Jones, so team is not tagged.

iv. Participated in the ~~pediatric oncology branch protocol~~

This is one entity, a named protocol, which does not fit our definition of protocol so it is not tagged.

v. Admitted to the **Sjogren** **Clinic**

Sjogren = Unit::Provider with "epo#" ; Clinic = Unit::Provider

In this example, Sjogren is referring to the unit, not the disease.

vi. Normal XR per ~~radiology~~ report

This is descriptive of the report, and does not refer to the specialist or the physical place of radiology.

vii. **NCI** **fellow**

NCI = Organization::Provider ; fellow = Occupation::Provider

The fellow is from the National Cancer Institute, and NCI does not further define the type of fellow.

viii. Patient lives in the ~~barracks~~, stationed at **Ft. Bliss**

Ft Bliss = Organization::Employer

Barracks are unspecific and only reveal that the patient is a military person and is not tagged; however, Ft. Bliss is an explicitly stated US Army post so is tagged as an organization and not as a city.

5 Occupation Types

Occupation is *not* one of the 18 personal identifiers (i.e., personally identifying information (PII)) specified in the HIPAA Privacy Rule and a de-identification system is not tasked to redact it. The reason behind our effort for annotating occupation is to study the value of occupation information in re-identifying the patient.

This label denotes an occupation that the person currently has or has had in the past including occupations mentioned as a potential culprit or environmental factor (e.g. "a *retired chemical engineer*" or "*disabled miner*"). We included descriptive qualifiers of an occupation only when those such terms helped to further define the occupation as "*attending physician*" and "*ICU nurse*" (see examples 5C.iii - 5C.xi below). We do not consider projected, hypothetical occupations ("the patient plans to open a car dealership") or general non-medical terms ("Patient knows the name of our president", "uses a computer at work", "was doing homework in school"). We exclude hobbies from occupation annotation, but not volunteer works. In other words, salary is not a prerequisite for occupation.

We annotate all provider occupations as **Occupation::Provider** including generalized health care professions such as *physician*, *health care provider* and *nurse* since we have removed the tag **Role::Provider** that was previously used for these terms. We annotate specified groups of providers which could include a variety of occupations such as "patient seen by the *ICU team*".

We consider studentship as **Occupation**, which may be expressed indirectly in terms of the grade level (e.g. "he is in *7th grade*" or "the patient is in *high school*"). We do not label a studentship that is not current nor do we label a grade level or studentship that is stated as a reference to denote a period in the past ("when the patient was a student…" or "the patient started using drugs while in high school"). If the patient is a student, the named title of his/her teacher would be labeled as **Occupation::Employer**, the name of his/her teacher would be labeled as **PersonalName::Employer** and the name of the school would be annotated as **Organization::Employer**.

Occupation (e.g. a *cook*) does not specify the employer, like where the person works does (e.g., "… at *Acme Restaurant*"), but sometimes, they are very closely knit together. For example, "he is an *Army Master Sergeant*," where *Army* is the **Organization::Employer** and *Master Sergeant* is the **Occupation::Patient**.

As previously mentioned, it can be difficult to distinguish between **Occupation::Provider** and **Unit::Provider**. Our rule of thumb is to determine whether the entity in question is an individual person/specialist (**Occupation**) or a physical place (**Unit**).

We do not annotate terms that are so general that when found alone do not imply an occupation such as “staff” and “team”.

Types of Occupation

A. Occupation::Patient

i. Pt **volunteers** part-time once a week

ii. He **studies computer science**.

iii. Mary is a summer **student**.

iv. Patient works in the **fellowship program** at NIH

v. Patient has a **PhD in creative writing**.

vi. Patient is an **active duty Marine**.

vii. He is an Army **Master Sergeant**.

viii. John G. Smith is a 82-yr old **retired** male

ix. The patient is currently **unemployed**.

x. Patient was performing an **internship** at NIH

xi. Patient is a **retired anesthesiologist**

xii. Patient is on **disability** and stays at home.

xiii. Patient attends the **2nd grade**

xiv. Patient just finished his **junior year in high school**

B. Occupation::Relative

i. The patient’s husband is in the **military**.

ii. His father is a **math teacher** at …

iii. Patient’s daughter is a **PhD** at NIH

iv. Patient’s brother is in **college**

C. Occupation::Provider

i. Patient received consults from **physical therapy** and **occupational therapy**

ii. Evaluation by the **urology service** showed an infection

iii. The patient’s **primary care doctor** was also notified.

iv. The patient was referred to Dr. Smith, **chief of oncology** at NIH

v. **Cytopathology fellow**: Paul Jones, MD

vi. Attending **surgeon**::Paul Jones, MD

vii. Patient to follow up with his **pediatric neurologist** in 2 weeks

viii. Patient was seen by the on-call **NIH fellow** for this episode

ix. The **PCMD** is present as well (abbreviation for primary care medical doctor)

x. Patient has private **dentist** as well

xi. Signed by GHK, **Cytopathology Clinical Fellow**

xii. Her **physician** told her to take baby aspirin.

D. Occupation::Other

i. Patient cited Obama as our **president**.

ii. The patient’s daughter sees a **psychologist** for these issues.

iii. Staging of the cancer is determined by a **pathologist** after review of all material

E. Occupation::Employer

i. His **first grade teacher** witnessed the seizure at school.

F. Examples/Discussion

i. primary **surgeon**

surgeon = Occupation::Provider

Surgeon is the occupation, and primary does not further define what type of surgeon so it is not tagged.

ii. results discussed with the **OR team**

OR team = Occupation::Provider

Team implies a group of OR people, a specific group of health care providers and thus is tagged.

iii. **RMD physician**

RMD physician = Occupation::Provider

The physician is the provider who specializes in the practice of Rehabilitation Medicine (RMD). RMD, while a distinct unit elsewhere, is identifying the occupation to a specific field, thus is included within the occupation tag and not as a separate unit.

iv. Will ~~graduate high school~~ next fall

This is a future occupation and by definition is not tagged as it might not happen, circumstances could prevent the student from graduating at that time.

v. Pt seen by **cardiology service**

cardiology service = Occupation::Provider

When referenced as seeing the patient in this manner, the cardiology service is a specific group of cardiology trained health care providers, so is tagged under Occupation instead of Unit.

vi. **ICU nurse**

ICU nurse = Occupation::Provider

The nurse is the occupation that is further defined by the fact that she specializes in ICU care.

vii. **Pediatric oncology branch fellow**

Pediatric oncology branch fellow= Occupation::Provider

Fellow is the occupation that is further defined by pediatric oncology branch; it refers to a pediatric oncologist, not the physical location.

viii. Seen by Dr. **Jones**’ ~~team~~

Jones = PersonalName::Provider

Team is an unspecified group of people working with Dr. Jones which could include a variety of different occupations. Team is not tagged.

ix. **OP13** **nurse**

nurse = Occupation::Provider ; OP13 = Location::Provider

Nurse is the occupation. OP13 does not further define what type of nurse, only the location.

x. Patient is ~~disabled~~

This does not imply the patient is on disability, it could be descriptive, as in a physical handicap.

xi. Patient is **deployed** to Northern Asia

Deployed = Occupation::Patient

Deployed implies that the patient is in the military, thus is his occupation.

6 Telecommunication Identifiers

The telecommunication identifier tag **Telecom** denotes identifiers like telephone, pager, beeper, and facsimile numbers as well as email accounts, URLs and hashtags. Distinct telecommunication identifiers are separated into individual **Telecom** labels if they are separated by spaces; otherwise, the entire phrase is annotated as one **Telecom** label.

Types of Telecommunication Identifiers

A. Telecom::Patient

i. Email results to patient at **mehmet.kayaalp@nih.gov**

ii. The patient’s website is **http://mehmet.kayaalp.us**.

B. Telecom::Provider

i. Call MD at Bp **#101-234**

ii. Please call clinic at **(123) 456-7890 ext.123**

iii. Call Dr. Jones for appointment at **(999)-999-9999x123**

C. Telecom::Other

i. The patient called **911** for help.

ii. The article is available at **http://lhncbc.nlm.nih.gov/publication/pub8978**.

7 Address Identifiers

Every token in a full address does not have the same value of information to identify a person. This is also recognized by the Privacy Rule, allowing the state information to be present in an otherwise fully de-identified clinical document. Within the realm of limited data set provision, where the document is partially de-identified, the Privacy Rule allows all address information to remain intact, except street names and numbers, which must be de-identified. If a de-identification system misses a city name, it would not be as detrimental as missing a street name. So, we annotate each type of address information separately in order to evaluate the performance of a de-identification system in a sensible manner. The labels we use are **Street**, **Building**, **Location**, **City**, **County**, **State**, **Zip**, and **Country**.

Street denotes the street name. **Building** denotes the building name (e.g., *The Dakota*) and/or number (e.g., *Building 38A*). We use **Location** to annotate numerically defined parts of an address, which includes P.O. Box numbers, house or street numbers, apartments, suite and office numbers as well as floor and room numbers inside office buildings or clinics. We included words such as *Building* and *Suite* that further specify location information (*Suite #15*, *Bldg 101*) to differentiate these labels from other alphanumeric labels (**Zip**, **AlphanumericID**).

The **City** label denotes cities, towns, and villages. We do not presume that a mentioned city is the place where the patient resides (e.g., while the patient was visiting Louvre Museum in *Paris*...) and annotate it as **City::Other**. We assume *New York* when standing alone in the text implies *New York City*. We also annotate city acronyms (e.g., *NYC*) and city nicknames (e.g., *Big Apple*) with the label **City**.

We use the label **County** for officially designated US counties only; whereas, we use the **State** label more liberally. **State** may denote not only any of the US states and territories (e.g., *Guam*, *Virgin Islands*, *Washington DC*, *and DC*) but also any equivalent unit from other countries (e.g., *Alberta*, *Okinawa*). We also use **State** to annotate regions covering areas of multiple counties (e.g., *Ohio River Valley*, *Western Pennsylvania*, *Midwest* and *East Coast*). We annotate descriptive qualifiers as part of the address. For example, we annotate *Downtown Dallas* as **City** and *upstate New York* as **State**.

We use the label **Country** liberally. It may denote the country of residence or the country of origin (e.g. "He was a 40-yr old *Ethiopian* man"). We do not annotate areas covering larger than a country (e.g. "he is from the West Indies"; "she is a young Asian female"). We do not use **Country** for religious associations (e.g. Jewish, Muslim, etc.) or ethnicities not associated with a country (e.g., Caucasian, Latino) but we do label a token as a **Country** if the country of residence or country of origin is explicit (e.g., *Mexican-American*, *Jewish-American* and *African-American*). If the country is spelled out along with the ethnicity or religion in the same phrase but in different tokens (e.g., “she is an *American* Jew”), we only annotate the country portion. We do not use **Country** for descriptive references such as “Patient takes a Canadian pain med” and “Japanese-speaking patient”.

We annotate five- or nine-digit US ZIP codes and foreign postal equivalents as **Zip**.

We use the personhood **Other**, when we cannot associate the address information with the patient, relative, employer or the provider.

As mentioned before, distinct address identifiers are separated into individual address labels if they are separated by spaces; otherwise, the entire phrase is annotated as one address label (see example 7D.ii and 7D.iii).

We annotate eponymous streets and other locations such as buildings, suites and departments, usually associated with a donor’s personal name, by messaging “epo#” along with the appropriate label. We do not specify eponymous cities, states, countries or geographical entities such as islands and peninsulas.

Types of Address Identifiers

A. Street::Patient

i. Patient lives in Apt 123 at 15 **Central Park West**.

B. Building::Provider

i. CC: Dr. Jones, **Woodward** (use “epo# to denote eponym) **Building**, NIH

C. Location::Patient

i. Patient lives in **Apt 123** at **15** Central Park West.

D. Location::Provider

i. Dictator address: **123** Main St. **Suite #301**

ii. CC: Dr. Jones, Woodward Bldg., **5th Floor** **Room 31A**

iii. Report filed in **Station 10-Room 33-A**

iv. Patient taken to **OP-5** where he was prepped for surgery

v. Pt transferred from the ICU to the **4th Floor** for care

vi. Location: **4 West** Cardiac Rehab

vii. **Room #123**

E. City::Patient

i. Patient from **Newark**, NJ

ii. Patient resides in **Harper’s Ferry**.

F. City::Relative

i. Patient was in **Boston** caring for her father

G. City::Provider

i. Patient treated in **Dallas** with radiation

ii. Pt’s doctor is in **Los Angeles**

H. City::Other

i. Patient attended the **New Orleans** Jazz Festival

ii. While traveling to **Paris**, the patient had a stroke

I. City::Employer

i. Patient is a journalist for CNN in **Atlanta**

J. County::Patient

i. Patient is from **Montgomery County**

K. State::Patient

i. Patient lived in Charlotte, **NC**

ii. Patient resides in **upstate New York**

iii. Send to patient at 1 Main St., **Alberta**, Canada 111111

L. State::Provider

i. He received chemo in Bethesda, **Maryland**

ii. Patient was seen in **Washington, DC** for his seizure that night

M. Country::Patient

i. Patient lives in **Monaco**.

ii. She is an **American** Jew.

iii. Patient is of **Polish** extraction

iv. Patient is a 40 yr old **Chinese** male

v. Patient is a 40-yr old **AA** female

N. Country::Relative

i. Father is of **Scottish** background

O. Country::Other

i. Patient had a heart attack while she was visiting **Vatican City**.

ii. Patient flying to **Kuwait** for work-related reasons.

P. Examples/Discussion

i. Pt is a 40 yr old **AA**F

AA = Country::Patient

This is an abbreviation for African-American female so we tag the country per definition. Gender/sex are excluded from the country tag even though there is no space.

ii. **OP13 nurse**
OP13 = **Location::Provider; nurse = Occupation::Provider**
OP13 is the place where the nurse works and does not further define the occupation of nurse.

iii. Patient enrolled in the ~~Chinese~~ root study
Chinese root study is one entity, a study, which we do not have a tag for and Chinese is the descriptive reference to being from China, so it is not tagged.

iv. Patient lives on the **Cape** in Massachusetts
Cape = County::Patient
Cape refers to Cape Cod, a geographic region that makes up Barnstable County, Massachusetts.

v. Report filed in **Heart Station** **12/2A123**
Heart Station = Unit::Other ; **12/2A123 = Location::Other**
These are physical places where a written report is located and labeled as such.

vi. Patient is from **Caroline County**, **Virginia**.
Caroline County = County::Patient ; Virginia = State::Patient
We do not consider eponymous counties or states, per the guidelines, even though "Caroline" and "Virginia" are personal names. They are annotated only with their applicable tags of "county" and "state".

8 Date Identifiers

Date is an annotation category comprising 5 identifier labels: **Year** (e.g., *2001*), **Month** (e.g., *September*), **Day** (e.g., *11th*), **DayOfWeek** (e.g., *Tuesday* but not Tuesdays) and **SpecialDay** (e.g., *9/11*, *Hurricane Sandy*, *Katrina*, *Cinco de Mayo*, *New Year's*).

We no longer annotate descriptive qualifiers associated with dates (the weekend of the *25th*, early *2001*) as they describe a general period of time, not a specific date, and do not provide identifiable information when seen alone. However, when the year is explicitly stated as in *Cinco de Mayo 2000*, we annotate *Cinco de Mayo* as **SpecialDay** and annotate *2000* as **Year**, because in this example, the date term refers to a full date, May 5, 2000. We use **SpecialDay::Patient** when the noted special day is linked together with something about the patient, otherwise we use **SpecialDay::Other** for those special days found standing alone.

We annotate not only those special days that are fixed in history such as *Pearl Harbor*, *2008 Market Crash* but also those special days that occur every year such as *New Year*, whose exact dates can be construed when combined with year information, which HIPAA Privacy Rule does not consider as personally identified information (PII). We also label personal special days such as birthdays or *Bar Mitzvah*, not only due to potential privacy concerns as such linkable information may be available from external sources, but also due to their potential importance in reference to other events in the narrative text.

If a date is described in terms of an interval or a range with clear and explicit begin and end date identifiers (examples 8.A.ii and 8.A.iii below), we separately annotate begin and end points with the appropriate date label. Note that we include the hyphen or other text found within this range if the date is unclear when standing alone (example 8.A.i). We do not label dates expressed in general terms where an exact date cannot be defined (“in the 1990s” and “during the winter months”).

We do not annotate cyclical day references: every day, every other day, every couple of days, never, MWF, Mondays, every other Tuesdays, first day (or Monday) of every month, every Christmas, every 4 months, which are seen frequently with medication or routine treatments.

Types of Dates

A. Year::Patient

i. Patient took meds from **2004** to **2005**

ii. Patient took meds from **2004**-**2005**

iii. This was used from **1991** to **1994** by the patient

iv. Pt was treated in early **1987** for 1 ½ years

v. After a miscarriage in **1990**, the patient was hospitalized

B. Year::Relative

i. Brother diagnosed in **1990**

ii. Brother diagnosed in Feb. **‘90**

C. Year::Other

i. Publication: American Journal of Cardiology, Vol 2, June **2001**

D. Month::Patient

i. Pt was admitted **06**/12/1945

ii. Patient was born **Nov.** 11, 1990

iii. Pt was seen in the beginning of **May**, 2007.

iv. Pt diagnosed since at least **June** 2010

v. Pt was seen in mid-**April**

vi. Patient was admitted **Nov.** 13th through **Nov.** 14th.

E. Day::Patient

i. Pt was admitted June **12th**, 1945

ii. Had chemo on the **second** of May

iii. Discharged the weekend of May **1** and **2**

F. DayOfWeek::Patient

i. He had surgery on **Monday**, June 1, 2011

ii. He felt intermittent chest pain last **Sunday**.

iii. Pt is to follow up between next **Tuesday** and **Friday**

G. SpecialDay::Patient

i. Patient had a stroke on **Christmas** last year.

ii. Patient was admitted on **Labor Day**.

iii. Patient was seen in the ER on **Independence Day**

iv. As the patient was celebrating his **59th birthday** during the **US Presidential Election Day** in 2008**…**

H. Special Day::Other

i. Patient stated that he started Atkinson diet on **the day of US Presidential Election** in 2008.

ii. The patient spent **Thanksgiving** with his friends.

I. Examples/Discussion

i. Pt attended a **July** **4th** party

July= Month::Other; 4th= Day::Other

Although this date is also known as a holiday, "Independence Day", it is written here as month and day and labeled with the corresponding tags. In this example, the date is *not* associated with any patient information, ie: patient had a stroke on July 4th, and is tagged with the personhood of Other.

ii. Patient took medication from **2004-5**

2004-5 = Year::Patient

The date 2004-5 implies the range of 2004 – 2005 and must be tagged as one entity for clarity.

9 Age Identifiers

We annotate chronological age using four different labels: **AgePII**, **AgeNPII**, **AgeFraction** and **AgePast**. Following HIPAA Privacy Rule, we annotate any chronological age above 89 as **AgePII** and any current age less than 90 expressed in whole numbers as **AgeNPII**. This holds true for stated ages of deceased patients at the time of death, regardless of a pronunciation of date of death, as such an age is a terminal point in his/her lifetime and not an arbitrary period in it. If the patient's current age is above 89 years or the narrative report provides indirect reference to the patient's age such that re-identification can be done through a simple arithmetic (e.g., "Twenty years ago, at the age of *75*, he had an ischemic attack"), we would annotate the numeric age reference (i.e., "*75* "in the example above) as **AgePII** as well.

We do not label descriptive terms since they do not further define a specifically stated age (e.g. "approximately *24* years old"). We do not label ages stated as periods of time ("she spent the third year of her life in the hospital").

If the patient's age is less than 89, but the age is mentioned in the report as a fraction of years (e.g., "Patient is a *4 and ½ month* old boy", "he will be *11 months* old in *two days*" or "patient is *36 months* old") we label such ages as **AgeFraction**. As noted in the above example, we annotate age references in the future. For fractional ages, we do not label tokens that do not convey age information (words *"*in*"* and "old*"* above) but do include the token that connects two parts of the age fraction together (the word "*and*" above).

The **AgePast** label is used when annotating ages in the past seen frequently with medical history. We use the definitions noted above when determining which tokens to label for past ages. For example, if a past age is expressed in terms of a whole number less than 90 years old, the whole number is annotated as **AgePast** ("when the patient was *84* years old"). Past ages expressed as fractions would include the terms that convey fractional age information (e.g. "when she was *22 months* old she was diagnosed") and be labeled as **AgePast**. As noted above, the exception here is if a past age is expressed in such a way that the patient's current age can be re-identified using simple arithmetic then we no longer use the **AgePast** label, but instead the appropriate **AgePII** or **AgeNPII** label (e.g. "In 2010 at the age of *88*, the patient was diagnosed" we would label "*88*" as **AgePII**).

The date of the medical file should be used as the reference point for the current age, given that relabeling an individual that would have turned 90 or older without knowledge that the individual lived to that age or not would make too many assumptions.

If past ages are expressed in terms of an interval or range with explicit past ages, we annotate the past ages individually with the appropriate labels.

We do not annotate bone age unless it is stated as identical to the chronological age. We do not annotate gestational age.

Age Types

A. AgeNPII::Patient

i. Patient is **<u>45</u>** years old
ii. He and his **<u>43</u>** yo twin sister
iii. In 1998 at age **<u>14</u>** the patient was first diagnosed
iv. Three years ago, in June 2009, at age **<u>40</u>**, the patient was diagnosed

v. Patient is an almost **4**-yr old boy who lives locally

vi. Patient was a **89** yr old woman who died from cancer

vii. When she expired in June 2010 she was **88** years old.

B. AgeNPII::Relative

i. Patient has **80**-yr old mother

ii. Patient's father, deceased at age **75**, had hypertension

iii. Patient is accompanied by her **four** year old twin sons

C. AgeNPII::Other

i. Patient works with **6**-yr olds

ii. Mammograms recommended at age **40**

D. AgePII::Patient

i. Patient is a **90**-yr old male

ii. Patient was **ninety-seven** when diagnosed

iii. Patient suffered from age 85 to present, **97** years old

iv. When she was **70**, back in 1990, the patient had a recurrence

E. AgePII::Relative

i. Significant for patient's mother being diagnosed in 1990 at age **80** from the disease

ii. Patient's grandfather who died when he was **95** was a carrier of the gene

F. AgeFraction::Patient

i. The patient's chronological age is **12 year and 6 month**.

ii. Patient is a **4 and 11/12 month** old boy.

G. AgeFraction::Relative

i. Her son is **3 weeks** old

ii. Here with her **13yr and 5mo** old sister

iii. Patient's nephew, **36 months** old, is a carrier of the disease as well

iv. Patient's brother died at **nine days** of life

H. AgePast::Patient

i. When the patient was **12** yrs old, she fractured her skull

ii. At the age of **16weeks** old patient was diagnosed

iii. Patient had surgery when he was almost **7** years old

iv. Patient received therapy from **12months** to **18 months** old

v. PMH: GERD from age **0** – **3**

vi. Menarche at age **11**

vii. Patient suffered from age **85**-97 years old

viii. During childhood (**2**-**3** years old) the patient noticed a change

ix. When she was **20 ½ years** old, the patient started therapy.

x. Pt was diagnosed at **day 6** of life

I. AgePast::Relative

i. Patient's niece had a stroke at age **17** and has been doing well

ii. Patient's son was diagnosed at **14 weeks** old

J. Examples/Discussion

i. Bone age is **1yr 4 months**. This is identical to the patient's age of **1 yr** **4 months**.

1yr 4 months = AgeFx::Patient

Both ages are explicated stated as being the exact same so both are tagged.

ii. Bone age is 1 yr 4 months. The patient's chronological age is **1 yr 4 months**.

1 yr 4 months = AgeFx::Patient

There is nothing stating that these ages are identical so you only tag the stated chronological age.

iii. Patient had surgery at age **3 and 6/12**

3 and 6/12 = AgePast::Patient

Based on context and verbiage one can assume this was a single surgery at age 3 ½ years old and not two surgeries at two different past ages of 3yrs and 6/12 years old.

iv. He has two sons, ages **4** years and **6 months**, both healthy

4= AgeNPII::Relative; **6 months = AgeFx::Relative**

Based on context and verbiage these are two different current ages of the patient's children.

10 Time Identifiers

The time label marks times as whole to include descriptors such as "*AM*" or "*PM*" and "*hours*" and "*o'clock*" when associated with actual times. Most times noted in the patient's chart are associated with the patient. We do not annotate generalized cyclical or noncyclical references to a time period within a day: every hour, early morning, mornings, every morning, afternoon, evening, night; however, we annotate specifically named time references (every day at *4:00pm*) as well as *noon* and *midnight* as **Time**, since they usually refer to 12PM and 12AM, respectively.

Types of Time

A. Time::Patient

i. Patient arrived at **4:30 in the morning**

ii. Transcribed by JL 9/1/13 **7:45 P**

iii. Takes pills every morning at **8:00a.m.**

iv. Episodes occur between **12:00** and **4:00pm**

v. Takes meds from **11am**-**1pm** daily

vi. Patient was discharged at **2100 hours**

vii. Patient’s ectopy was recorded at **02:13:00-2**

B. Time::Relative

i. Patient’s mom is a RN and works the **3-11pm** shift.

C. Examples/Discussion

i. Breast biopsy (~~6 o’clock)~~ is pending
By context this is the position of the biopsy site and not an actual time that the biopsy occurred and thus is not tagged.

11 Numeric and Alphanumeric Identifiers

We annotate numeric and alphanumeric identifiers using the following 4 distinct labels: **MedicalRecordNo**, **ProtocolID, HealthRecordID**, and **AlphanumericID**. The first three are almost exclusively associated with the patient.

When considering alphanumeric identifiers, we annotate every consecutive non-space character in a token including non-letter, non-digit characters such as “#”, “*”, “/”, “\”, and “:”. We use **MedicalRecordNo** to annotate any patient-specific identification number related to that hospital course or visit, which includes medical record numbers, account numbers and chart numbers.

We define **ProtocolID** as the alphanumeric characters of clinical protocols. We exclude generic named non-clinical protocols such as “research protocol” seen frequently in the files. We distinguish protocol identifiers from other alphanumeric identifiers due to their high information value for NIH clinical studies. NLM Scrubber tries to preserve protocol identifiers but redacts other common alphanumeric identifiers. Protocol identifiers are usually treatment specific information associated with a cohort and cannot be linked to the patient by anyone other than a small number of protocol administrators. Protocols used in NIH clinical trials have a distinct format (11-AA-1111 or 11-A-1111) and should be annotated as **ProtocolID** when recognized. We do not label trial (enrolled in the A12345 double-blind placebo trial) or study names (the Chinese Root study) at this time.

We use **HealthRecordID** to label other numeric patient identifiers that are issued by the provider and are uniquely linked to the patient, such as order numbers, accession numbers, specimen numbers and other hospital assigned numeric or alphanumeric identifiers.

We use the label **AlphanumericID::Patient** to annotate all numeric and alphanumeric identifiers specific to the patient that are not issued by the provider (e.g., social security numbers) and are not telecommunication identifiers.

We use **AlphanumericID::Other** when we cannot distinguish the nature of the number or what it denotes. This includes generic billing codes (ICD-9, SNOMED codes, etc.) and hospital issued procedure codes that although are not patient-specific they could possibly be linked back to the patient. Radiology report series numbers and image numbers that are not patient-specific but still could be linked back to the patient would only be accessible to those who access to do so. These generic alphanumeric terms, ie: *image #40*, describe clinical information that is pertinent to the researchers who have access to particular databases and are not identifiable to the patient.

We have encountered a vast number of other alphanumeric tokens that are not defined by the labels noted above. These include, but are not limited to, microbiology terms, pathology terms and genetic markers. We currently have partially annotated files containing such alphanumeric tokens when they can be defined by one of our current labels (e.g., cells are positive for *CD163* would be labeled as **Diagnostic** and *QRS4500 gene splicer* would be labeled as **Device**). We do not annotate generic alphanumeric tokens that cannot be linked back to the patient (e.g., refer to cassette #1a).

Types of Alphanumeric Identifiers

A. MedicalRecordNo::Patient

i. Medical record number **123-34-22**

ii. MR **#123-34-22**

iii. Patient Jane Smith, Account# **123345**

iv. Patient ID: **1234567**

v. Chart number: **123ABC-456efg**

B. HealthRecordID::Patient

i. Specimen# **S10-1235-A** sent to lab

ii. NM **#123-56-1234**

iii. Labs noted with ID # **123456789**

iv. Occupational Therapy Order Number: **00123VNV33**

v. Accession No.: **987654**

vi. DNA **#B123**

vii. **DNA#B123**

viii. DNA # **B123**

ix. Cassettes **#123-ABC-456** #2A-C

C. ProtocolID::Patient

i. Participated in protocol # **13-AA-6789**

ii. This is visit #6 for **#95-BB-1234** protocol

iii. Accepted into protocol: **#123-A-1234**: Jones-Smith disease in the elderly on Alpha-Beta Drug infusions

iv. Patient was screened for **12-AB-1234** study

D. AlphanumericID::Patient

i. John Smith SS# **123-45-6789**

E. AlphanumericID::Provider

i. Thank you for this consult **1234**/**54321** RS/MD

ii. Dictated by **123-45**, 7/1/12 0800AM

F. AlphanumericID::Other

i. **1234**/**5678** (i.e., numbers with no context)

ii. Saffra Lodge Conf# **R123456**

iii. US abdomen (**123**)

iv. Dx: **123.0**, **456.7**

G. Examples/Discussion

i. Patient takes ~~MS-275~~ per protocol

Based on the context, this is a medicine, therefore not tagged as an alphanumeric identifier.

ii. Pt enrolled in NIH protocol**#123-A-1234**

#123-A-1234 = ProtocolID::Patient

Although the word "protocol" is not separated by white space we know that it is not part of the protocol ID but instead a typographically error.

iii. Patient wears ~~SPF-30~~ when outside

Based on the context, this is sunscreen and not an alphanumeric identifier.

iv. Patient is stage **T1-M1-M0**

T1-M1-M0 = Diagnostic

Based on the context, this is a cancer stage and labeled with a medical tag.

v. Stains positive for **CD4**, **AFB** and **HMB-45**

CD4 = Diagnostics; AFB=Diagnostics; HMB-45= Diagnostics

We currently annotate stains, genes, primers and similar tokens seen frequently in pathology reports as diagnostics.

vi. Lymph nodes are removed, ~~22/24~~ negative for cancer

Based on context this means 22 out of 24 of the lymph nodes and is not tagged.

vii. Patient receives **BL22 therapy** twice a month

BL22 therapy = Therapy

BL22 is an immunotoxin used for a specific disease, this phrase is a type of therapy that uses it.

viii. Patient is **G1P0A0L1**

G1P0A0L1 = Diagnostic

G1P0A0L1 is an alphanumeric way to define the patient's obstetric history and labeled with a medical tag.

ix. Refer to slides ~~#1-A-D~~

This generic slide number is not identifying to the patient. The slide number does contain clinical information that can be important to the clinicians who have access to the patient's data.

12 Personally Identifying Contexts

In these guidelines, we discussed how we label entities that were mentioned in the HIPAA Privacy Rule along with a few other, closely related entities, some of which can be PII in certain contexts. We are aware of the fact that due to intricacies of natural languages, it is possible to specify a context in which the person could be identified indirectly without using the set of labels that we discussed so far. In those cases, we label the tokens with **PIC**, denoting Personally Identifying Context.

In the hypothetical example, "*received his injuries* while he was reporting from *Tahrir Square*", we would annotate *reporting* with label **Occupation::Patient** and *received his injuries* and *Tahrir Square* with **PIC::Patient**, since the latter would provide context so specific that along with the occupation information would probably identify the person directly.

Types of PIC

A. PIC::Patient

i. The examination of **his injury** that he endured **during his US championship match** today

B. PIC::Relative

i. Her sister is the **first recipient of the Nobel Peace Prize** in this decade.

C. PIC::Employer

i. He was the CEO of **the largest US contractor during the Operation Iraqi Liberation**.

D. Examples/Discussion

i. Patient is a **locally ranked runner, 2nd in his age group** in the **Baltimore** area

locally ranked runner, 2nd in his age group = PIC::Patient

Baltimore = City::Patient
We do not annotate hobbies such as running as occupation so this information cannot be labeled with other tags, it is PIC because along with the city information it can identify the patient.

13 Identified Person's Kinship or Relations to the Patient

A **Role** is a reference to a person such as *mother*, *father*, *daughter* and *boyfriend*. We annotate such relationship references with **Role** when they are associated with an entity or description in the nearby text. The **Role** is the subject of the text in question.

The aim of **Role** annotation is to substantiate any non-patient personhood reference with respect to the patient. For example, if the word *mom* and a telephone number are mentioned in the same (or subsequent) sentences and we annotate the phone number with label **Telecom::Relative**, and *mom* with the **Role::Relative** informing us to which relative that alphanumeric identifier belongs. We exclude pronouns such as he, she, him, hers, their, themselves, etc. We use the label **Role** only if no other label is suitable for that annotation. If the reference specifies a personally identifying context, instead of using the label **Role,** we would annotate it as **PIC**, because **PIC** denotes PII; whereas, **Role** only informs us about the relationship between the patient and the person whose identifier is mentioned in the text.

We do not use the label **Role::Patient** for "patient" based on practicality due to the amount of time it takes to annotate these tokens due to the numerous appearances in a file. We no longer use **Role::Provider** as these occupations are defined and labeled as **Occupation::Provider**.

Types of Roles

A. Role::Patient

Patient is the **brother** of another patient admitted here

B. Role::Relative

i. Patient's **maternal uncle** died from the disease in 1999
ii. Patient's **twin sister** is healthy
iii. He underwent a **sibling** donor transplant
iv. Both the patient and her **partner** are at increased risk
v. The patient is the brother of another **patient** admitted here

C. Role::Employer

i. His **supervisor** was forcing him to do things that he was not willing to do

D. Role::Other

i. The patient lives with his **girlfriend**, Jane.

ii. Pt was discovered by a **woman** visiting her husband in the hospital

iii. Her **friend** noticed she was unresponsive and called 911.

E. Examples/Discussion

i. Patient is a healthy **volunteer**

volunteer = Role::Patient

Based on context, volunteer is his role in the study and not the occupation, i.e.: patient is volunteering at this place.

ii. The patient's **mother** is the **translator**

Mother = Role::Relative ; translator = Occupation::Relative

Mother is the role that further substantiates the occupation of translator.

iii. ~~Patient~~ is accompanied by his ~~mother~~

We do not annotate patient as the subject of a sentence per guidelines and mother has no entity or description associated with it, so we do not tag either.

iv. There is a history of stroke on the ~~maternal~~ side of the family.

This does not reference a specific person (could be the mother, the mother's aunt, etc) and is not tagged.

v. The patient's **maternal** height is 167 cm.

Maternal = Role::Relative

In this example, maternal refers to a specific person's (patient's mother) height and is tagged as role.

14 Non-identifying, Non-specific Text Parts

Although most labels that we discussed in these guidelines so far denote PII, a few of them (e.g., **PersonalName::Provider**, **PersonalName::Other**, **AgeNPII::Patient**, **AlphanumericID::Other**) denote identifiers that are not considered PII, which we call non-PII. If any text portion has not been annotated with any labels, we presume that are not PII and provide no significant aid to the de-identification process and to the evaluation of de-identification performance. We use the label **NPII** to label all remaining tokens that we do not annotate with the label set introduced in this document. As previously mentioned, **NPII** labels are not associated with personhood.

15 Non-Identifying Medical Terms

Medical terms found in our annotation process are not considered PII but do provide important information to the researchers. We created the labels **Anatomy**, **Device**, **Diagnostics**, **Therapy** and **Eponym** for these non-identifying medical terms. The first four labels are used for medical terms and are not affiliated with any particular personhood (patient, relative, employer, provider and other) as seen with the identifier labels. The **Eponym** label captures any proper name associated with these medical tags and has five subcategories associated with it (**Anatomy**, **Device**, **Diagnostic**, **Therapy** and **Other**).

We used the label **Anatomy** to tag anatomical locations of the body. We included "left" and "right" as they further describe the particular anatomical location (e.g., *left leg*) but not positional descriptive terms that do not differentiate between an anatomical location (e.g., posterior *leg*). We did not include microscopic anatomy (e.g., red blood cells) even though anatomy applies to the cellular level of the body as this would create a difficult task when annotating pathology reports and other microbiology files: where do you stop tagging-- white blood cells, leucocytes, neutrophils, DNA? We included specific fluids and specimens associated with a specific body part/location (e.g., *cerebral spinal fluid*) but not general terms like blood. We tagged cavities and spaces as **Anatomy** as well. We did not tag stand-alone anatomical-like words with multiple meanings such as "band".

We used the tag **Device** to label tools or objects used for medical purposes. We included descriptors that helped to identify the device (e.g., *spine biter*). We included the manufacturer as **Device** only when the manufacturer's name was interchangeable with the device or named along with the device (see 15.B.iv - 15.B.v below). We did not tag general terms that had multiple meanings outside of the medial realm such as "tube". We did not tag lab reagents.

We used the tag **Diagnostics** to label stated diagnoses, differential diagnoses, conditions that indicate or imply a diagnosis (e.g., *relapse*), clinical signs observed by the health care provider (e.g., *drop in hemoglobin*), factual non-subjective symptoms observed or stated by the patient (e.g., *blood in stool*) and general tests, procedures and scoring methods performed as tools to obtain a diagnosis (e.g., *laparoscopy*). We did not tag negative, normal or hypothetical findings as **Diagnostics**. We included descriptors of these tests (e.g., *skin biopsy*) as well as terms of severity or range (e.g., *mild ataxia*) that further define these tests; however, we used **Anatomy** for additional anatomical descriptions (e.g., *left leg* skin biopsy). We tagged diagnoses stated as microscopic findings only if they were previously stated as a diagnosis. We did not tag patient-specific laboratory values that were too specific to one patient to be considered a diagnosis (e.g., BP= 140/80). We included genetic markers and other pathology terms as **Diagnostics** as well.

We used the label **Therapy** to tag procedures and tests performed directly on the patient's body to improve his health status (e.g., *excision*) not used as a diagnostic tool. We included noun and noun phrases that described these therapies which included surgeries (e.g., *nephrectomy*), vaccinations (e.g., *shingles vaccine*), noun phrases that described these procedures (e.g., *extraction of wisdom teeth*) and medicinal therapies specifically stated as such (e.g., *Campath Therapy*). We did not include medicine names or procedures done microscopically to specimens of the body.

We created the tag **Eponym** to capture any proper name associated with the previous medical labels. This allows us to differentiate between eponymous medical terms and other personal names. The subcategories reflect which medical term is described (**Eponym::Anatomy**, **Eponym::Device**, **Eponym::Diagnostics**, **and Eponym::Therapy**). We use **Eponym::Other** when the proper name in question did not fit into the other subcategories or any other pre-existing identifying tags, such as eponymous pathology stains. We tagged eponymous diagnosis terms even when the diagnosis was negative (see 15.G.ii below).

Types of Non-Identifying Medical Tags

A. Anatomy

i. **Skin**, **right clavicular**
ii. **Colon**, distal ring
iii. Dorsal aspect of **spiny processes of the T5 level**
iv. **Bone marrow**
v. **Syrinx cavity**
vi. Lower **left arm**
vii. **Left lower leg**

B. Device

i. **Catheter line**
ii. **sutures**
iii. **three pin head holder**
iv. **3100 Genetic Analyzer** (ABI)
v. **ABI 3100 Genetic Analyzer**
vi. **Ultrasound gel**
vii. **Spatula**
viii. **Biopsy bag**

C. Diagnostics

i. Dx: **adult T-cell lymphoma**
ii. Dx: **cultured cells**, no lymphoma
iii. Patient is **MRSA+**
iv. **Bone marrow aspirate**
v. **HPV DNA, low risk** detected
vi. **Lumbar puncture**
vii. **Liver biopsy** tissue
viii. Ishak **fibrosis score**=1
ix. **Atypical cells** were present
x. Dx: kidney, left **kidney tumor** (**biopsy**), **renal cell carcinoma clear cell type**
xi. Findings include **dense connective tissue** with **dense fibrous tissue**
xii. **High white blood cell count**, WBC=34.7
xiii. **Enlarging ovarian mass**
xiv. Positive for **CD36** and **Her-2**
xv. **CD79a** stains are observed

D. Therapy

i. **Stem cell transplant**
ii. **Hormonal therapy** : estrogen
iii. **Estrogen therapy**
iv. Inferior T5 **laminectomy**
v. **Blood pressure test**
vi. **Excision**
vii. **Create a site** for **drainage**

E. Eponym::Anatomy

i. Veins of **Batson**'s plexus
ii. **Eustachian** tube
iii. Circle of **Willis**
iv. Tumor located in the loop of **Henle**

F. Eponym::Device

i. **Kerrison** ronguer
ii. **Jackson-Pratt** drain
iii. **Midas Rex** drill

G. Eponym::Diagnostics

i. **Ishak** fibrosis score of 12
ii. Dx: no evidence of **Parkinson**'s disease
iii. **Downs** Syndrome

iv. **Burkitt**’s lymphoma

H. Eponym::Therapy

i. **Heimlich** maneuver

ii. The **Valsalva** was attempted

iii. The **Trendelenburg** position

iv. **Whipple** procedure

16 Discussion

Although in this document, we provided some rationale for these guidelines, the main purpose of this document is to serve as a guide to de-identification system designers, human annotators, and users of de-identified data with a number of concrete examples. We described the rationale behind these guidelines in our freely available article (Kayaalp et al. 2015), which with these guidelines are complimentary to each other. In order to fully understand and most effectively apply these guidelines, the user may need to use both documents in parallel.

Unlike any scholarly publication, this set of guidelines is dynamic in nature. As we learn from new experience, we adjust and alter the guidelines so that they best suit our needs. Therefore, it is inevitable that these guidelines will deviate from what we have stated in our publications in the past.

Actually, even in this current (as of June 2016) version of these guidelines, we can mention a few changes that are different from our aforementioned article (Kayaalp et al. 2015). We no longer annotate organizational units as organizations. The label **Organization** denotes organization names only; whereas, **Unit** denotes a subdivision of a larger organization, unless it has been addressed with a unique name. We no longer use **Role::Provider** for general health care occupations, which are now labeled as **Occupation::Provider**.

We also have been reconsidering the use of label **Period**, which we discussed in our paper but excluded from our current guidelines. Since our annotators had difficulties in distinguishing periods in the past medical history, we decided to refine the concept. In these guidelines, we have introduced the label **AgePast** to denote a particular age in the medical history of the patient.

17 Conclusions

In this document, we compiled a comprehensive list of well-defined labels used for de-identification purposes. This document has been developed by the NLM Scrubber team through several years of experience on clinical text de-identification. Although our main goal is to provide the prerequisite guidelines to the NLM Scrubber team, we also hope that others who are interested in annotating clinical text would find these guidelines as useful suggestions.

In our paper (Kayaalp et al. 2015), we discussed challenges of annotating clinical text for de-identification purposes. The discussions both in our paper and in these guidelines ought to be taken by inexperienced annotators as cautionary remarks, since applying HIPAA Privacy Rules may not be as easy as it may appear at first glance. However, we hope and believe that these guidelines accompanied with our paper would provide them the necessary aid in their annotation tasks. It is important to note that even if clinical text is de-identified through automatic means such as NLM Scrubber, clinicians, who are guardians of the protected health information, are the ultimate responsible parties who have to verify and confirm that the de-identified clinical text is free from personal identifiers. In order to do that properly and effectively, those clinicians need to be well versed on the issues discussed in our paper and these guidelines.

The first and foremost target of these guidelines has been the members of the NLM Scrubber annotation team, who need a set of references upon which they all agree which parts of clinical text need to be annotated and which labels should be used to annotate them.

These guidelines have also been useful to design NLM Scrubber and to choose NLM Scrubber labels that replace redacted PII in clinical text. We maintain a transformation map between these two sets of (human assigned vs. NLM Scrubber assigned) labels in order to correctly evaluate the de-identification performance of NLM Scrubber.

By making these guidelines widely available, we also intend to provide some guidance to users of NLM Scrubber on how to interpret labels in the de-identified text. Although these labels are self-explanatory in most cases, at certain times (e.g., labels replacing identifiers of neighbors or co-workers of the patient) they may not be as straightforward without these guidelines.

Other target audiences of these guidelines can be parties who are interested in designing de-identification systems, annotating clinical text for building a training dataset for a machine learning system or for building an evaluation dataset to measure the performance of their de-identification system.

18 Future Work

There are a number of small pilot projects on which we are currently working. We are testing various ideas, checking their feasibilities and their added values to the de-identification process. One of these ideas is annotating rare diseases. Some argue that rare diseases may help re-identify patients since they are rare, but we are *not* convinced fully that records linking rare diseases to patients or vice versa are readily available. Those who have such linkage information have almost always access rights to the corresponding protected health information. Having said that, this issue is still open in our project.

Another outstanding issue is annotating professional titles and degrees. It may seem contradictory not to annotate them since we annotate occupations. We had to make that decision because annotating provider titles (e.g., MD, RN, etc.) imposes a great burden to our annotation group due to their sheer number of occurrences. We are currently work on a pilot to annotate them automatically through NLM Scrubber and let our annotation team verify the correctness of their annotation. We will make a decision on this issue after we conclude the pilot.

Acknowledgements

We are grateful to our summer student interns Tyne McGee, Justin Jones, Huong Tran, Kayla Saadeh, Kecia Ellick and Marissa Beldon for their diligent annotation works and the feedback they provided us. We also thank the other contributors to this work, who are Zeyno Dodd and Ming Chen, who provided us software support, and Vojtech Huser, who influenced us to introduce the label **Role**.

References

Kayaalp, M., Browne, A.C., Dodd, Z.A., Sagan, P., McGee, T., McDonald, C.J. (2015). Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation. *Proceedings of the Annual American Medical Informatics Association Fall Symposium.*