How to De-identify Data

Xulei Shirley Liu
Department of Biostatistics
Vanderbilt University
03/07/2008

Outline

- The problem
- Brief history
- The solutions
- Examples with SAS and R code

Background

- The adoption of electronic medical records (EMRs) is growing.
- Researchers are increasingly turning to EMRs as a source of clinically relevant patient data.
- There are calls for the use of EMRs in observational studies, such as epidemiologic and health services research, and clinical studies, such as clinical trials.
- On the other hand, a majority of patients, and the public in general, are concerned about unauthorized disclosure and use of their personal health information in an era of the EMRs.
- Furthermore, rates of medical identity theft have been increasing, and the risks are exacerbated with the use of EMRs.

HIPAA Privacy Rule

The Department of Health and Human Services (HHS) issued the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule in December 2000 and it took effect on April 14, 2003.

Source: http://privacyruleandresearch.nih.gov/pr_04.asp

Commonly Used Terms

- Protected Health Information - PHI is individually identifiable health information transmitted by electronic media, maintained in electronic media, or transmitted or maintained in any other form or medium.
- Covered Entity - A health plan, a health care clearinghouse, or a health care provider who transmits health information in electronic form in connection with a transaction for which HHS has adopted a standard.

Source: http://privacyruleandresearch.nih.gov/pr_04.asp

Purposes of the HIPAA Privacy Rule

- HIPAA Privacy Rule establishes minimum Federal standards for protecting the privacy of individually identifiable health information.
- The Privacy Rule establishes conditions under which covered entities can provide researchers access to and use of protected health information (PHI) when necessary to conduct research. The Rule is not intended to impede research.

Source: http://privacyruleandresearch.nih.gov/pr_04.asp

Compliance with HIPAA Privacy Rule

The Privacy Rule permits covered entities to use and disclose data that have been de-identified without obtaining an authorization and without further restrictions on use or disclosure because ***de-identified data*** are no longer PHI and, therefore, are not subject to the Privacy Rule.

Source: http://privacyruleandresearch.nih.gov/healthservicesprivacy.asp

What's de-identified data?

The term de-identified data refers to patient data from which all information that could reasonably be used to identify the patient has been removed/replaced (eg., removing name, address, SSN, etc…).

Methods of Data De-identification

- Statistical methods

  It uses statistical and scientific principles and methods proven to render information not individually identifiable. – not used that often.

- Heuristic methods

  It consists of rules about which variables to generalize and which variables to exclude from a data set when it is disclosed.

Heuristic Methods of De-identification

- Creation of de-identified data sets (safe-harbor method)
- Creation of limited data sets (partially de-identified)
- Generation of variables

Creation of de-identified data sets

- Removes 18 specified identifiers from a dataset. This requirement includes data elements related to the individual or relatives, employers, or household members of the individual.

  1. Names
  2. All geographic subdivisions smaller than a state, including:
     * Street Addresses
     * City
     * County
     * Precinct
     * Zip Code

Creation of de-identified data sets (continued)

3. All elements of dates (except year) or dates directly related to an individual, including:

   * Birth date

   * Admission date

   * Discharge date

   * Date of death

   * All ages over 89 and all elements of dates indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.

Creation of de-identified data sets (continued)

4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical Record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web universal resource locators (URLs)
15. Internet Protocol (IP) address numbers

Creation of de-identified data sets (continued)

16. Biometric identifiers, including finger and voice prints
17. Full-face photographic images and any comparable images

Creation of de-identified data sets (continued)

18. Any other unique identifying number, characteristic, or code, except as permitted by the re-identification rules.

Creation of limited data sets

Removes 16 specified identifiers from a dataset. This requirement includes data elements related to the individual or relatives, employers, or household members of the individual.

Creation of limited data sets (continued)

Importantly, unlike de-identified data, PHI in limited data sets may include the following:

1. Addresses other than street name or street address or post office boxes, such as town/city and/or 5-digit zip codes.
2. All elements of dates (such as admission and discharge dates)
3. Unique codes or identifiers not listed as direct identifiers.

Creation of limited data sets (continued)

- Before disclosing a limited data set to a researcher, a covered entity must enter into a data use agreement with the researcher.
- Data use agreement establishes the terms and conditions in which the covered entity will allow the use and disclosure of a limited data set to the data recipient.

Generation of Variables

Generation of unidentifiable variables based on specified identifiers not only ensures patient privacy but also keeps all the useful information for studies.

Data Re-identification

A researcher authorized to view the PHI may assign a code or other means of record identification to allow de-identified information to be re-identified, provided that:

1. Derivation: The code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual.
2. Security: The authorized researcher does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.

Methods of Generating Variables

- Generate study identification numbers and replace study subjects' names or social security numbers or medical record numbers, etc.
- Shift dates by using date of birth or study start date, etc.

Generation of Study Identification (ID) Number

- Dataset - PSA

SSN	TESTDATE	PSA	DOB	DOD
638621207	11/7/2002	2.92	7/25/1923	
222441045	4/10/2003	1.1	3/29/1917	
123456789	5/29/2001	0.3	12/2/1915	3/13/2003
111223333	11/12/1999	2.7	5/5/1915	3/11/2001
222441045	7/29/2004	1	3/29/1917	
343551104	8/9/2000	2.97	8/7/1920	4/7/2004
111223333	9/12/2000	2.2	5/5/1915	3/11/2001
343551104	9/17/2001	2.69	8/7/1920	4/7/2004
638621207	2/11/2002	33.8	7/25/1923	
343551104	11/6/2002	2.83	8/7/1920	4/7/2004
289761125	6/7/2000	8	9/21/1921	
746921198	9/26/2000	0.7	5/5/1923	
638621207	10/12/2000	1.84	7/25/1923	
343551104	11/5/2003	3.51	8/7/1920	4/7/2004
123456789	4/18/2000	0.2	12/2/1915	3/13/2003

Generate sequential study ID numbers

SAS code

```
proc sort data=psa;by ssn;run;
data psa_temp1;
        set psa;
        by ssn;
        retain STUDY_ID 0;
        if first.ssn then STUDY_ID = STUDY_ID + 1;
run;
```

R code

```
psa_temp1 <- psa[order(psa$SSN),]
psa_temp1$STUDY_ID <- as.integer(unclass(factor(psa_temp1$SSN)))
```

Generation of Study Identification (ID) Number

- Dataset - PSA (sorted by SSN)

SSN	TESTDATE	PSA	DOB	DOD
111223333	11/12/1999	2.7	5/5/1915	3/11/2001
111223333	9/12/2000	2.2	5/5/1915	3/11/2001
123456789	5/29/2001	0.3	12/2/1915	3/13/2003
123456789	4/18/2000	0.2	12/2/1915	3/13/2003
222441045	4/10/2003	1.1	3/29/1917	
222441045	7/29/2004	1	3/29/1917	
289761125	6/7/2000	8	9/21/1921	
343551104	8/9/2000	2.97	8/7/1920	4/7/2004
343551104	9/17/2001	2.69	8/7/1920	4/7/2004
343551104	11/6/2002	2.83	8/7/1920	4/7/2004
343551104	11/5/2003	3.51	8/7/1920	4/7/2004
638621207	11/7/2002	2.92	7/25/1923	
638621207	2/11/2002	33.8	7/25/1923	
638621207	10/12/2000	1.84	7/25/1923	
746921198	9/26/2000	0.7	5/5/1923	

Generate sequential study ID numbers (continued)

- Dataset - PSA_temp1 (add study IDs)

STUDY_ID	SSN	TESTDATE	PSA	DOB	DOD
1	111223333	11/12/1999	2.7	5/5/1915	3/11/2001
1	111223333	9/12/2000	2.2	5/5/1915	3/11/2001
2	123456789	5/29/2001	0.3	12/2/1915	3/13/2003
2	123456789	4/18/2000	0.2	12/2/1915	3/13/2003
3	222441045	4/10/2003	1.1	3/29/1917	
3	222441045	7/29/2004	1	3/29/1917	
4	289761125	6/7/2000	8	9/21/1921	
5	343551104	8/9/2000	2.97	8/7/1920	4/7/2004
5	343551104	9/17/2001	2.69	8/7/1920	4/7/2004
5	343551104	11/6/2002	2.83	8/7/1920	4/7/2004
5	343551104	11/5/2003	3.51	8/7/1920	4/7/2004
6	638621207	11/7/2002	2.92	7/25/1923	
6	638621207	2/11/2002	33.8	7/25/1923	
6	638621207	10/12/2000	1.84	7/25/1923	
7	746921198	9/26/2000	0.7	5/5/1923	

Generate random study ID numbers

SAS code

```
proc sort data=psa nodupkey out=psa_uni;by ssn;run;
data ran;
        set psa_uni;
        ran=ranuni(23423); /* generate the unique random number */
run;

proc sort data=ran;by ran; run;
data study_id(keep=ssn study_id);
        set ran;
        by ran;
        retain study_id 0;
        if first.ran then study_id=study_id+1;
run;

proc sort data=study_id;by SSN;run;
data psa_temp1_ran;
        merge study_id psa;
        by ssn;
run;
```

Generate random study ID numbers (continued)

R code

```
set.seed(1)
STUDY_ID<-data.frame(STUDY_ID=sample(1:length(unique(psa_temp1$SSN)),
                replace = FALSE), SSN = unique(psa_temp1$SSN))
psa_temp1_ran <- merge(STUDY_ID, psa_temp1, by = "SSN", all = TRUE)
```

Generate random study ID numbers (continued)

- Dataset - PSA_temp1.1 (add study IDs)

STUDY_ID	SSN	TESTDATE	PSA	DOB	DOD
5	111223333	11/12/1999	2.7	5/5/1915	3/11/2001
5	111223333	9/12/2000	2.2	5/5/1915	3/11/2001
2	123456789	5/29/2001	0.3	12/2/1915	3/13/2003
2	123456789	4/18/2000	0.2	12/2/1915	3/13/2003
7	222441045	4/10/2003	1.1	3/29/1917	
7	222441045	7/29/2004	1	3/29/1917	
1	289761125	6/7/2000	8	9/21/1921	
3	343551104	8/9/2000	2.97	8/7/1920	4/7/2004
3	343551104	9/17/2001	2.69	8/7/1920	4/7/2004
3	343551104	11/6/2002	2.83	8/7/1920	4/7/2004
3	343551104	11/5/2003	3.51	8/7/1920	4/7/2004
4	638621207	11/7/2002	2.92	7/25/1923	
4	638621207	2/11/2002	33.8	7/25/1923	
4	638621207	10/12/2000	1.84	7/25/1923	
6	746921198	9/26/2000	0.7	5/5/1923	

Shift dates

Choice of date - date of birth, research subject enrolment date, etc.

Shift dates by date of birth (DOB)

SAS Code

```
data psa_temp2;
        set psa_temp1;
        Centered_TestDate=TestDate – DOB;
        Centered_DOD=DOD-DOB;
run;
```

R code

```
psa_temp2 <- transform(psa_temp1, Centered_TESTDATE=TESTDATE-
DOB, Centered_DOD=DOD-DOB)
```

Shift dates by date of birth (DOB) (continued)

Dataset – PSA_temp2

STUDY_ID	SSN	TESTDATE	PSA	DOB	DOD	Centered_TESTDATE	Centered_DOD
1	111223333	11/12/1999	2.7	5/5/1915	3/11/2001	30872	31357
1	111223333	9/12/2000	2.2	5/5/1915	3/11/2001	31177	31357
2	123456789	5/29/2001	0.3	12/2/1915	3/13/2003	31225	31878
2	123456789	4/18/2000	0.2	12/2/1915	3/13/2003	30819	31878
3	222441045	4/10/2003	1.1	3/29/1917		31423	
3	222441045	7/29/2004	1	3/29/1917		31899	
4	289761125	6/7/2000	8	9/21/1921		28749	
5	343551104	8/9/2000	2.97	8/7/1920	4/7/2004	29222	30559
5	343551104	9/17/2001	2.69	8/7/1920	4/7/2004	29626	30559
5	343551104	11/6/2002	2.83	8/7/1920	4/7/2004	30041	30559
5	343551104	11/5/2003	3.51	8/7/1920	4/7/2004	30405	30559
6	638621207	11/7/2002	2.92	7/25/1923		28960	
6	638621207	2/11/2002	33.8	7/25/1923		28691	
6	638621207	10/12/2000	1.84	7/25/1923		28204	
7	746921198	9/26/2000	0.7	5/5/1923		28269	

Shift dates by date of birth (DOB) (continued)

Dataset – PSA_temp2

STUDY_ID	SSN	TESTDATE	PSA	DOB	DOD	Age_TESTDATE	Age_DOD
1	111223333	11/12/1999	2.7	5/5/1915	3/11/2001	85	86
1	111223333	9/12/2000	2.2	5/5/1915	3/11/2001	85	86
2	123456789	5/29/2001	0.3	12/2/1915	3/13/2003	85	87
2	123456789	4/18/2000	0.2	12/2/1915	3/13/2003	84	87
3	222441045	4/10/2003	1.1	3/29/1917		86	
3	222441045	7/29/2004	1	3/29/1917		87	
4	289761125	6/7/2000	8	9/21/1921		79	
5	343551104	8/9/2000	2.97	8/7/1920	4/7/2004	80	84
5	343551104	9/17/2001	2.69	8/7/1920	4/7/2004	81	84
5	343551104	11/6/2002	2.83	8/7/1920	4/7/2004	82	84
5	343551104	11/5/2003	3.51	8/7/1920	4/7/2004	83	84
6	638621207	11/7/2002	2.92	7/25/1923		79	
6	638621207	2/11/2002	33.8	7/25/1923		79	
6	638621207	10/12/2000	1.84	7/25/1923		77	
7	746921198	9/26/2000	0.7	5/5/1923		77	

De-identified data set

SAS Code

```
data De_ID(keep=STUDY_ID PSA Age_TESTDATE
                        Age_DOD);
    set psa_temp2;
run;
```

R code

```
De_ID <- subset(psa_temp2, select=c(STUDY_ID, PSA,
            Age_TESTDATE, Age_DOD))
```

De-identified data set (continued)

Dataset – De_ID

STUDY_ID	PSA	Centered_TESTDATE	Centered_DOD
1	2.7	30872	31357
1	2.2	31177	31357
2	0.3		31878
2	0.2		31878
3	1.1	31423	
3	1	31899	
4	8	28749	
5	2.97	29222	30559
5	2.69	29626	30559
5	2.83	30041	30559
5	3.51	30405	30559
6	2.92	28960	
6	33.8	28691	
6	1.84	28204	
7	0.7	28269	

De-identified data set (continued)

Dataset – De_ID

STUDY_ID	PSA	Age_TESTDATE	Age_DOD
1	2.7	85	86
1	2.2	85	86
2	0.3	85	87
2	0.2	84	87
3	1.1	86	
3	1	87	
4	8	79	
5	2.97	80	84
5	2.69	81	84
5	2.83	82	84
5	3.51	83	84
6	2.92	79	
6	33.8	79	
6	1.84	77	
7	0.7	77	

Re-identification File

- Re-identification file is used to re-identify de-identified data set.
- Re-identification file is the file that has both specified identifiers and the corresponding generated variables.
- Re-identification file must be kept secure.

Re-identification File (continued)

- Create unique re-identification file

SAS code

```
data link(keep=STUDY_ID SSN DOB);
      set psa_temp2;
      by ssn;
      if first.ssn;
run;
```

R code

```
link=unique(subset(psa_temp2, select = c(STUDY_ID, SSN, DOB)))
```

Re-identification file (continued)

STUDY_ID	SSN	DOB
1	111223333	5/5/1915
2	123456789	12/2/1915
3	222441045	3/29/1917
4	289761125	9/21/1921
5	343551104	8/7/1920
6	638621207	7/25/1923
7	746921198	5/5/1923

Use re-identification file to re-identify data

- Merge re-identification file into de-identified data set.
- Shift dates back.

SAS code

```
data psa_temp3;
      merge link De_ID;
      by STUDY_ID;
      TESTDATE = Centered_TESTDATE + DOB;
      DOD = Centered_DOD + DOB;
      format TESTDATE DOD mmddyy10.;
run;
```

R code

```
psa_temp3 <- transform(merge(link, De_ID, by="STUDY_ID"),
                       TESTDATE=DOB+Centered_TESTDATE,
                       DOD=DOB+Centered_DOD)
```

Re-identification file

STUDY_ID	SSN	DOB
1	111223333	5/5/1915
2	123456789	12/2/1915
3	222441045	3/29/1917
4	289761125	9/21/1921
5	343551104	8/7/1920
6	638621207	7/25/1923
7	746921198	5/5/1923

De-identified data set

STUDY_ID	PSA	Centered_TESTDATE	Centered_DOD
1	2.7	30872	31357
1	2.2	31177	31357
2	0.3		31878
2	0.2		31878
3	1.1	31423	
.	.	.	.
.	.	.	.
.	.	.	.
6	33.8	28691	
6	1.84	28204	
7	0.7	28269	

Re-identified data set

STUDY_ID	SSN	TESTDATE	PSA	DOB	DOD	Centered_TESTDATE	Centered_DOD
1	111223333	11/12/1999	2.7	5/5/1915	3/11/2001	30872	31357
1	111223333	9/12/2000	2.2	5/5/1915	3/11/2001	31177	31357
2	123456789	5/29/2001	0.3	12/2/1915	3/13/2003	31225	31878
2	123456789	4/18/2000	0.2	12/2/1915	3/13/2003	30819	31878
3	222441045	4/10/2003	1.1	3/29/1917		31423	
3	222441045	7/29/2004	1	3/29/1917		31899	
4	289761125	6/7/2000	8	9/21/1921		28749	
5	343551104	8/9/2000	2.97	8/7/1920	4/7/2004	29222	30559
5	343551104	9/17/2001	2.69	8/7/1920	4/7/2004	29626	30559
5	343551104	11/6/2002	2.83	8/7/1920	4/7/2004	30041	30559
5	343551104	11/5/2003	3.51	8/7/1920	4/7/2004	30405	30559
6	638621207	11/7/2002	2.92	7/25/1923		28960	
6	638621207	2/11/2002	33.8	7/25/1923		28691	
6	638621207	10/12/2000	1.84	7/25/1923		28204	
7	746921198	9/26/2000	0.7	5/5/1923		28269	

Summary

- With the growing use of EMRs in observational and clinical research, the issue of safeguarding personal health information was raised.
- HIPAA Privacy Rule is the first comprehensive Federal protection for the privacy of personal health information.
- Heuristic data de-identification is a frequent means to comply with HIPAA Privacy Rule

Summary (continued)

- Heuristic methods of data de-identification
  a. Deletion of 18 specified identifiers from a data set
     – de-identified data set.
  b. Deletion of 16 specified identifiers from a data set
     – limited data set.
  c. Generation of variables to replace identifiers.
- Re-identification

http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/XuleiLiu

Acknowledgements

- Robert Greevy, PhD
- Theresa Scott, MS

http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/XuleiLiu