I got a fake ID!
A macro to de-identify user-specified variables in a dataset

Jennifer R. Popovic, DVM, MA
Harvard Medical School
Harvard Pilgrim Health Care Institute

Outline

- Scope of presentation*
- What is de-identification?
  - Definition and purpose
  - Healthcare de-identification regulations and standards
    - Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule Safe Harbor standard
  - Example implementation approaches
    - Which is best?
- SAS® implementation of one de-identification approach

Scope of presentation

- De-identification methods discussed here primarily rely on Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule's **Safe Harbor standard as framework**
  - ***Generally applicable*** to clinical, administrative and/or survey ***healthcare*** data
  - Based on ***U.S. framework***
- Other data streams (i.e., genomic data) raise different set of issues
- Other countries may operate under different privacy definitions, frameworks and regulations

De-identification definition and purpose

- **Removing or obscuring personally identifiable** information from individual records in a way that ***minimizes*** **the risk of unintended disclosure of the identity of individuals** and information about them
- **Facilitate data sharing** and dissemination
  - **Without violating** Federal or other **regulations** protecting individual privacy
  - **Maintaining analytic value**

De-identification regulations and standards

- Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, a.k.a “Privacy Rule”
  - Safe Harbor
  - Statistical or expert determination methods
- Federal Policy for the Protection of Human Subjects, a.k.a. “Common Rule”

HIPAA Safe Harbor

- Requires a specific set of 18 data elements be **removed or generalized** in a dataset in order for it to be considered de-identified

HIPAA Safe Harbor data elements

- Names
- Social security numbers
- All elements of dates, except year
- Telephone numbers
- FAX numbers
- Email addresses
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Biometric identifiers, including finger and voice prints
- Device identifiers and serial numbers
- Geographic subdivisions smaller than a state
- Vehicle identifiers and serial numbers, including license plate numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) addresses
- Full-face photographs and any comparable images
- Any other unique identifying number, characteristic, or code, except as permitted by other specific guidance (not to be covered here)

Example de-identification implementation approaches

Method	Explanation/Example
Variable/field redaction	Delete columns
Record or cell suppression/redaction	Delete rows or cells
Randomization	Retains a field that represents the direct identifier but the original values have been replaced with randomly-generated values. E.g.: 123-456-7890→SSN1
Value scrambling, masking, truncation or encoding	JONES→NSEOJ JONES→*O*E* JONES→JO JONES→%$&*#
Blurring, aggregating, generalizing	Take exact age and aggregating into age groups. 54→50-60

Which de-identification implementation method is best?

IT DEPENDS!

- Degree to which data should be altered or obscured depends on **how they will be used/disseminated and the rules/regulations that govern them**

SAS implementation of one de-identification approach: Randomization

Program purpose

- Creates a de-identified dataset by **assigning a randomly generated *caseid*** to one or more user-specified variables in an existing dataset.
- **Creates crosswalk files** for each variable replaced with a random *caseid*.
- Has utility in studies where **individual-level data are needed** but **actual values of person-level identifiers are required to be masked**.

Program parameters

- **INFILE**: Name of the dataset containing variables to be masked
- **VARLIST**: Name(s) of variable(s) that requires replacement with a randomly-generated caseid
- **OUTFILE**: Name of the output dataset
- **XWALKLIB**: Libref to which crosswalk files will be saved

Program steps

1. Error trapping: Check for missing values in any field listed in VARLIST parameter (CMISS function) and abort program and issue custom WARNING to log if missing values found
2. Create dataset of unique ids and a random number for each unique VARLIST variable
3. Create crosswalks containing new 'caseid' for each variable in VARLIST
4. Create final de-identified dataset: Merge 'caseid' variables and drop real identifiers; perform in a loop over all variables names in VARLIST

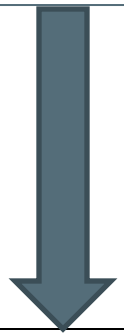
Show-and-tell: Example input dataset

- Want to replace **patid** and **ndc** fields with randomized caseids.

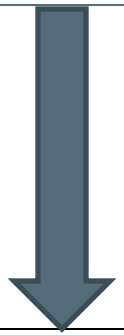
patid	rxdate	ndc	rxsup	rxamt
111-11-1111	02/01/2007	00002032902	30	30
222-22-2222	05/01/2007	00047093930	60	30
333-33-3333	11/30/2007	00182190601	90	30

Show-and-tell: Create dataset of unique IDs and a random number

patid	rxdate	ndc	rxsup	rxamt
111-11-1111	02/01/2007	00002032902	30	30
222-22-2222	05/01/2007	00047093930	60	30
333-33-3333	11/30/2007	00182190601	90	30



patid	random
222-22-2222	0.07833
333-33-3333	0.19938
111-11-1111	0.87697



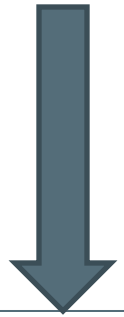
ndc	random
00002032902	0.08995
00182190601	0.35924
00047093930	0.74836

Show-and-tell: Create crosswalks

patid	random
222-22-2222	0.07833
333-33-3333	0.19938
111-11-1111	0.87697

patid	patid_caseid
222-22-2222	patid1
333-33-3333	patid2
111-11-1111	patid3

ndc	random
00002032902	0.08995
00182190601	0.35924
00047093930	0.74836



ndc	ndc_caseid
00182190601	ndc1
00002032902	ndc2
00047093930	ndc3

Show-and-tell: Example outputs

patid	rxdate	ndc	rxsup	rxamt
111-11-1111	02/01/2007	00002032902	30	30
222-22-2222	05/01/2007	00047093930	60	30
333-33-3333	11/30/2007	00182190601	90	30

- Example output dataset

patid_caseid	ndc_caseid	rxdate	rxsup	rxamt
patid3	ndc2	02/01/2007	30	30
patid1	ndc3	05/01/2007	60	30
patid2	ndc1	11/30/2007	90	30

- Example crosswalks

patid_caseid	patid
patid1	222-22-2222
patid2	333-33-3333
patid3	111-11-1111

ndc_caseid	ndc
ndc1	00182190601
ndc2	00002032902
ndc3	00047093930

SAS® version limitation

- **Code can only run on version 9.2 and higher of SAS®**, as it makes use of the **CMISS function** that was unavailable in prior software versions
- CMISS counts missing values across observations; used in the error trapping portion of the program

```
newvariable=cmiss(argument1, argument2,...);
```

A character expression is counted as missing if it evaluates to a string that contains all blanks or has a length of zero.

A numeric expression is counted as missing if it evaluates to a numeric missing value: ., ._, .A, ... , .Z.

Opportunities for future enhancements

- Redact variables
- “Relative-ize” dates
- Redact cells

Acknowledgments

- Colleagues at the Sentinel Operations Center (SOC), Harvard Medical School / Harvard Pilgrim Health Care Institute
  - Sentinel Analytic Development and Programming Team
  - Robert Rosofsky

References

- Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington (DC): National Academies Press (US); 2015 Apr 20. Appendix B, Concepts and Methods for De-identifying Clinical Trial Data. Available from: http://www.ncbi.nlm.nih.gov/books/NBK285994/
- Nelson GS. 2015. "Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification". Proceedings of the 2015 SAS Global Forum, Dallas, TX. http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf
- Shostak JS. 2006. "De-Identification of Clinical Trials Data Demystified". Proceedings of the 2006 PharmaSUG Conference, Bonita Springs, FL. http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf
- U.S. Department of Health and Human Services, Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. 2012 Nov 26. Available from: http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

Questions?

Jennifer R. Popovic, DVM, MA
Harvard Medical School / Harvard Pilgrim Health Care Institute
jennifer_popovic@harvardpilgrim.org
http://www.mini-sentinel.org/