DH02
Data de-identification made simple

Jørgen Mangor Iversen, LEO Pharma, 2016
Introduction

Inspiration

• PhUSE: De-Identification Standard for CDISC SDTM 3.2
  ◦ PhUSE de-identification working group
  ◦ 20-May-2015
  ◦ PHUSE_STDM_redaction.xls
  ◦ Available for members on the PhUSE web site
  ◦ Small set of well defined rules
  ◦ Exhaustive list of rules per variable
Introduction

Basic idea

- Metadata extract from definition spreadsheet
- Metadata extract from trial data
- Match trial and definition metadata
- Determine order of operations
- One operation => one macro
- Execute and report
De-Identification Standard

Rules

- Recode subject ID (key across domains)
- Recode ID variable (independent)
- Offset (dates as if all begins on the same date)
- Remove (variables)
- Keep (no action)
- No further de-identification (no action)
- Elevate to continent (simple mapping)
- Derive Age (handle few very old subjects)
- Aggregate Age (manual assessment)
- Review and only redact values with personal information (manual)
- Remove dataset (always CO, maybe SU and others)

- Remove observation (not in standard)
De-Identification Standard
SDTM Implementation Guide definitions

- Small excerpt of interesting columns

<table>
<thead>
<tr>
<th>Domain</th>
<th>Variable_Name</th>
<th>DI_Primary_Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STUDYID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DOMAIN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>USUBJID</td>
<td>Recode subject ID</td>
</tr>
<tr>
<td></td>
<td>POOLID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SPDEVID</td>
<td>Remove</td>
</tr>
<tr>
<td></td>
<td>--SEQ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>--GRPID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>--REFID</td>
<td>Recode ID variable</td>
</tr>
<tr>
<td></td>
<td>--SPID</td>
<td>Recode ID variable</td>
</tr>
<tr>
<td></td>
<td>--TRT</td>
<td>Remove</td>
</tr>
<tr>
<td></td>
<td>--TERM</td>
<td>Remove</td>
</tr>
<tr>
<td>DM</td>
<td>INVID</td>
<td>Remove</td>
</tr>
<tr>
<td>DM</td>
<td>INVNAM</td>
<td>Remove</td>
</tr>
<tr>
<td>DM</td>
<td>BRTHDTC</td>
<td>Remove</td>
</tr>
<tr>
<td>DM</td>
<td>AGE</td>
<td>Derive Age</td>
</tr>
<tr>
<td>DM</td>
<td>AGEU</td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td>SEX</td>
<td>Keep</td>
</tr>
<tr>
<td>DM</td>
<td>RACE</td>
<td>Keep</td>
</tr>
<tr>
<td>DM</td>
<td>ETHNIC</td>
<td>Keep</td>
</tr>
<tr>
<td>DM</td>
<td>ARMCD</td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td>ARM</td>
<td></td>
</tr>
</tbody>
</table>
proc format;
  value $rules
    'Remove dataset'                                          = '%di_remove_dsn'
    'Derive Age'                                              = '%di_derive_age'
    'Offset'                                                  = '%di_offset'
    'Elevate to continent'                                    = '%di_continent'
    'Recode subject ID'                                       = '%di_recode_subject'
    'Recode ID variable'                                      = '%di_recode_id'
    'Remove'                                                  = '%di_remove_var'
    'Keep'                                                    = '%di_keep'
    'No further de-identification'                            = '%di_no_further'
    'Review and only redact values with personal information' = '%di_manual'
    other                                                      = '';
proc format;
  invalue rules
  '%di_remove_dsn'    = 1
  '%di_derive_age'    = 2
  '%di_offset'        = 3
  '%di_continent'     = 4
  '%di_recode_subject' = 5
  '%di_recode_id'     = 6
  '%di_remove_var'    = 7
  '%di_keep'          = 9
  '%di_no_further'    = 8
  '%di_manual'        = 10
  other               = 99;
Metadata
Mapping countries to continents

• Obtaining metadata apparently difficult
• SAS maps data always at hand
  ◦ maps.names, maps.metamaps
• Merge by name and isoname
  ◦ Slightly different results. Use both and merge
• Russia and Mexico(?) also in Europe!
  ◦ Data needs to be fixed; duplicates removed
• Once per new SAS version/maps update
Metadata

Study metadata

• Strictly SDTM domains and variables
• Global variables across domains
  ◦ Without -- prefix; USUBJID, VISITNUM
• Domain specific variables
  ◦ Without -- prefix; AETERM, EXROUTE
• List of domains from study datasets
  ◦ Excluding SUPP-- domains
• Other variables
  ◦ With -- prefix, replaced with domain code
• All variables
  ◦ Merge and remove duplicates
Metadata

Study metadata principles

• Merge enriched specification variable list with variables in study data
• Add rules to each variable
• Leave out all variables not in specification
• Assign priority to each variable and sort
Operations
Preparing operations

• Calculate offset per subject
  ◦ Earliest of
    ─ RFSTDTC (first reference)
    ─ SVSTDTC (visit 1)
    ─ DSSTDTC (date of informed consent)
  ◦ …as study start date
  ◦ Each subject have individual offset relative to study start date

• Create USUBJID look-up table
  ◦ USUBJID must be consistent across domains
Operations macros
Common structure

• Validate arguments
  ◦ Derive any support values
• Register operation
  ◦ Categorize operation by importance
• Perform operation
  ◦ Update study data

• Debugging comments
  ◦ Stand alone test data
Operations

1. Remove dataset

• Avoid superfluous work
  ◦ Update metadata, remove as well!
Operations

2. Derive age

• Age is expected in DM
  ◦ May be empty; calculate
  ◦ May need redaction
    ─ If any subject > 89 years old

• Age unit may vary
  ◦ Days, hours, weeks, months, years

• Create plots of age distribution
Operations

3. Offset dates

• Simple subtraction of offset from date
• Handling of partial dates
  ◦ Impute date to mid month/year as needed
  ◦ Subtract offset from imputed date
  ◦ Remove imputed date parts
Operations

4. Elevate to continent

- Convert continent dataset to format
- Create variable REGIONDI containing continent
- Ensure REGIONDI has sufficient length
Operations

5. Recode subject identifier

• All identifiers are recoded using the same algorithm
• Recoding USUBJID requires re-sorting of the dataset
6. Recoding

- Calculate max recoded value
- Obfuscate variable (by un-sorting)
  - Sort by some semi-random sequence
- Calculate new monotonic value from next multiple of 10 of max recoded value
  - 215, 101, 147… becomes 1001, 1002, 1003…
- Update values in situ
Operations

7. Remove variable

- Delay dropping variables until after dependent values are calculated
  - Age depends on date of birth
Operations
8 & 9. Do nothing

- Keep variable
- No further action
Operations

10. Manual handling

• Review and only redact values with personal information
Operations
Addendum to specification

- Remove observation
  - Qval in SUPP-- may contain names and dates etc.
    - Subject initials
    - Investigators names and/or initials

- Country should be removed after elevation to continent
Reporting

Messages collected

• HTML output
• Resorting messages
  ◦ Operation 10, manual processing, is highlighted at the top of the list
• Age distribution is plotted
## Reporting

### Example of messages

**Issues handled during DE-identification**

<table>
<thead>
<tr>
<th>Category</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Variable AEACNOTH in dataset AE must be handled manually</td>
</tr>
<tr>
<td></td>
<td>Variable CMDOSFRQ in dataset CM must be handled manually</td>
</tr>
<tr>
<td></td>
<td>Variable CMDOSU in dataset CM must be handled manually</td>
</tr>
</tbody>
</table>

**Issues handled during DE-identification**

<table>
<thead>
<tr>
<th>Category</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Removing observations in dataset SUPPDM where SUBSTR(QVAL,1,5)=&quot;EARLY&quot;</td>
</tr>
</tbody>
</table>

**Issues handled during DE-identification**

<table>
<thead>
<tr>
<th>Category</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Dataset AE re-sorted by USUBJID</td>
</tr>
<tr>
<td></td>
<td>Dataset CM re-sorted by USUBJID</td>
</tr>
<tr>
<td></td>
<td>Dataset DM re-sorted by USUBJID</td>
</tr>
<tr>
<td></td>
<td>Dataset DS re-sorted by USUBJID</td>
</tr>
<tr>
<td></td>
<td>Dataset EX re-sorted by USUBJID</td>
</tr>
</tbody>
</table>
Reporting
Example of age distribution

Issues handled during DE-identification

Category=3

Action taken

Deriving variable AGE in dataset DM

Distribution of age summarized

Distribution of age in detail
Process
Always several iterations

• Setup and run the program suite
• Inspect the report for manual operations and age distribution
• Inspect all domains for dates, names, initials etc. not mentioned
• Insert study specific operations
• Run again
• Repeat
Summary
Solution statistics

• 16 programs and macros including metadata build
• 1062 Lines of code
  ◦ 15% comments
  ◦ < 900 lines of code doing something
• 10 Days of work
• 5 Days of testing (15 remaining)
Summary
Reference and contact


• Contact details
  ◦ Jørgen Mangor Iversen
  ◦ LEO Pharma A/S
  ◦ jmidk@leo-pharma.com
Questions

...and hopefully answers