DH01
De-Identification Standards for CDISC Data Models

PhUSE AC, Vienna
13. October 2015
Jean-Marc Ferran (Qualiance)
Khaled El Emam (Privacy Analytics)
Sarah Nolan (Liverpool University)
Nick De Donder (Business & Decision Life Sciences)
Boris Grimm (Boehringer Ingelheim)
<table>
<thead>
<tr>
<th>Participants</th>
<th>Working Group</th>
<th>Company/Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinitha Arumugam &amp; Patricia Coyle (GSK)</td>
<td>Jean-Marc Ferran</td>
<td>Nancy Freidland (IBM)</td>
</tr>
<tr>
<td>Per-Arne Stahl (AstraZeneca)</td>
<td>Nick De Donder (BDL)</td>
<td>Gene Lightfoot (SAS)</td>
</tr>
<tr>
<td>Sherry Meeh (Johnson &amp; Johnson)</td>
<td>Cathal Gallagher (d-Wise)</td>
<td>Jacques Lanoue &amp; Benoit Vernay (Novartis)</td>
</tr>
<tr>
<td>Kim Musgrave (Amgen)</td>
<td>Nate Freimark (Theorem &amp; CDISC)</td>
<td>Joanna Koft (Biogen Idec)</td>
</tr>
<tr>
<td>Gary Chen (Shire)</td>
<td>Khaled El Emam (Privacy Analytics)</td>
<td>Jennifer Chin (EISAI)</td>
</tr>
<tr>
<td>Carl Herremans (Merck)</td>
<td>Beate Hientzsch &amp; Sven Greiner (Accovion)</td>
<td>Kishore Papineni, Thijs van den Hoven &amp; Bharat Jaswani (Astellas)</td>
</tr>
<tr>
<td>Kelly Mewes (Roche)</td>
<td>Kristin Kelly (Accenture)</td>
<td>Sarah Nolan (Liverpool University &amp; Cochrane)</td>
</tr>
<tr>
<td>Boris Grimm (Boehringer Ingelheim)</td>
<td>Shafi Chowdury (Shafi Consultancy)</td>
<td></td>
</tr>
</tbody>
</table>
Agenda

• Overview of Data Sharing

• Data De-identification Standards for SDTM 3.2

• Current EMA CSR anonymization guidance
Overview of Data Sharing

1. Introduction and purpose

The aim of the European Medicines Agency ("the Agency") is to protect and foster public health. Transparency is a key consideration for the Agency in delivering its service to patients and society. Although the Agency since its creation has launched several initiatives to increase transparency of information on medicinal products, there is growing demand from stakeholders for additional transparency, not only about the Agency's deliberations and actions, but also about the clinical data on which regulatory decisions are based. The Agency is committed to continuously extend its approach to transparency and has, therefore, taken the initiative to develop a policy on publication of clinical data, in accordance with article 89 of Regulation (EC) No 726/2004.1 Consultations with a broad range of stakeholders and European Union (EU) bodies have taken place in drafting this policy. It should be noted that this policy is without prejudice to Regulation (EC) No 1094/2001 and, therefore, it does not replace the existing policy on access to documents (created by medicinal products for human and veterinary use) (POLICY/0034 CMAR/110195/2005), which came into effect in December 2010. However, the provisions of this policy are not intended in any manner to limit the application or the rights given by Regulation (EC) No. 1094/2001. Any natural or legal person may continue to submit a request for access to documents to the Agency independently of the proactive publication mechanisms established by this policy.


About

Medien & Visual

The Project Data Sharing database allows researchers affiliated with life science companies, hospitals, and institutions, as well as independent researchers to share, integrate, and analyze patient-level, comparative arm phase II cancer data, which provides unprecedented deep understanding, protocols, data descriptors, and case report forms. The database is also provided to enable users to tap into the value of the data. Additionally, the platform provides a Community where Authorized Users can collaborate around all aspects of data analysis and cancer research.
Research proposals requesting access to patient level data (number of proposals)

<table>
<thead>
<tr>
<th>Number of Research Proposals submitted up to 31 August 2015</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements check</td>
<td></td>
</tr>
<tr>
<td>In process</td>
<td>12</td>
</tr>
<tr>
<td>Withdrawn by the requestor</td>
<td>16</td>
</tr>
<tr>
<td>Did not meet requirements (further details)</td>
<td>11</td>
</tr>
<tr>
<td>Potential conflict of interest or an actual or potential competitive risk</td>
<td>0</td>
</tr>
<tr>
<td>Met requirements</td>
<td>119</td>
</tr>
<tr>
<td>IRP review</td>
<td></td>
</tr>
<tr>
<td>In process</td>
<td>2</td>
</tr>
<tr>
<td>Withdrawn by the requestor</td>
<td>1</td>
</tr>
<tr>
<td>Rejected or advised to re-submit</td>
<td>11</td>
</tr>
<tr>
<td>Approved or approved with conditions</td>
<td>105</td>
</tr>
<tr>
<td>Data Sharing Agreement</td>
<td></td>
</tr>
<tr>
<td>In process</td>
<td>26</td>
</tr>
<tr>
<td>Withdrawn by the requestor</td>
<td>2</td>
</tr>
<tr>
<td>Not agreed (not signed)</td>
<td>0</td>
</tr>
<tr>
<td>Agreed (signed) View details of these research proposals</td>
<td>77</td>
</tr>
<tr>
<td>Data preparation</td>
<td></td>
</tr>
<tr>
<td>In process</td>
<td>7</td>
</tr>
<tr>
<td>Withdrawn by the requestor</td>
<td>1</td>
</tr>
<tr>
<td>Complete (data available)</td>
<td>69</td>
</tr>
<tr>
<td>Research project</td>
<td></td>
</tr>
<tr>
<td>In process</td>
<td>69</td>
</tr>
<tr>
<td>Withdrawn by the requestor</td>
<td>0</td>
</tr>
<tr>
<td>Not published</td>
<td>0</td>
</tr>
<tr>
<td>Published</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: clinicalstudydatarequest.com/Metrics.aspx
11OCT2015
Data De-Identification Guidelines

Quasi/Direct Identifiers Assessment

Processes

Rules

Residual Risk

Data De-Identification

#PhUSE
Disclaimer

De-Identification Standards for CDISC SDTM 3.2

• The views in the deliverable represent the consensus of the Working Group

• The rules described do not guarantee an acceptable or very small residual risk of re-identification
  – “It is generally recommended if certain conditions are met, that after the application of the rules described in this document, a second pass examining low frequency should be performed to confirm that there are no risks from low frequencies.”
Key Principles

Direct & Quasi Identifiers are identified

- **Direct identifiers**: One or more direct identifiers can be used to uniquely identify an individual. E.g. Subject ID, Social Security Number, Telephone number, Exact address, etc. It is compulsory to remove or pseudonymize any direct identifier.

- **Quasi identifiers**: Quasi identifiers are background information that can be used in connection with other information to identify an individual with a high probability. E.g. Age at baseline, Race, Sex, Events, Specific Findings, etc.

Primary & Alternative Rules for De-Identification are assigned

- **Primary rule**: Pro-active data de-identification maximizing data utility
- **Alternative rule**: Reactive data de-identification and special cases
- **Impact on data utility** is evaluated qualitatively
- **Implementation guidance** for each rule is provided
- **Rules address different scenarios** rather than different implementation possibilities

Comments are added to guide the reader

- To explain further the **rational of a given assessment**
- To warn users for **exceptions or special considerations**
MH dates and patients > 89

Dates

Core

Offset
Partial Dates
DoB/DoD
Adaptive Design
Extension Studies
No further de-identification

Relative
**Dates Offset**

**Recommended Algorithm**

See Appendix 1
### Issue with Partial Dates

**Ex: Delta applied of -14 days**

<table>
<thead>
<tr>
<th>Visit/Event</th>
<th>Date (Source)</th>
<th>Imputed Date</th>
<th>Offset Date</th>
<th>Offset Partial Date (Final)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit 0</td>
<td>10JAN2013</td>
<td>10JAN2013</td>
<td>27DEC2012</td>
<td>27DEC2012</td>
</tr>
<tr>
<td>Visit 1</td>
<td>10FEB2013</td>
<td>10FEB2013</td>
<td>27JAN2013</td>
<td>27JAN2013</td>
</tr>
<tr>
<td>Visit 2</td>
<td>08MAR2013</td>
<td>08MAR2013</td>
<td>22FEB2013</td>
<td>22FEB2013</td>
</tr>
<tr>
<td>Event X</td>
<td>MAR2013</td>
<td>15MAR2013</td>
<td>01MAR2013</td>
<td>MAR2013</td>
</tr>
<tr>
<td>Visit 3</td>
<td>12APR2013</td>
<td>12APR2013</td>
<td>29MAR2013</td>
<td>29MAR2013</td>
</tr>
</tbody>
</table>
Free-text

- Dictionary Coded Variable Required/Expected
  - Remove if not Important to Analysis (or recoded)

- Dictionary Coded Variable Permissible
  - Review and redact PII

- Remove
Review and Redact PII in Free-Text

```
“Dr James diagnosed broken right arm of B.K”
```

- Redact or Remove
- Review and redact PII

```
“--redacted--” or Remove variable
```

```
“--redacted-- diagnosed broken right arm of --redacted--”
```

Not dictionary coded and important to analysis
Geographical Location

- CMTRT
- Geographical Location
- PII of Third-Party
- Country
- Race
- Site/Investigator ID/Name
- Elevate to Continent
- Keep if important to Analysis
- Remove
- Recode ID if important to Analysis
Deliverable
De-Identification Standards for CDISC SDTM 3.2

<table>
<thead>
<tr>
<th>Observation Class</th>
<th>Domain Prefix</th>
<th>Variable Name</th>
<th>Variable Label</th>
<th>Type</th>
<th>Direct_Quasi_Ientifier</th>
<th>Di_Primary_Rule</th>
<th>Di_Alternative_Rule</th>
<th>Di_Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>RPENDTC</td>
<td>Date/Time of End of Participation</td>
<td>Char</td>
<td>Quasi Level 2</td>
<td>Offset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>DTHDTC</td>
<td>Date/Time of Death</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Offset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>DTHFL</td>
<td>Subject Death Flag</td>
<td>Char</td>
<td>Quasi Level 2</td>
<td>Keep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>SITEID</td>
<td>Study Site Identifier</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Remove</td>
<td>Recode ID variable</td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>INID</td>
<td>Investigator Identifier</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Remove</td>
<td>Recode ID variable</td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>INVNNAM</td>
<td>Investigator Name</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Remove</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>BIRTHDTC</td>
<td>Date/Time of Birth</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Remove</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>AGEI</td>
<td>Age</td>
<td>Num</td>
<td>Quasi Level 1</td>
<td>Derive Age</td>
<td>Aggregate Age</td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>AGEU</td>
<td>Age Units</td>
<td>Char</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>SEX</td>
<td>Sex</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Keep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>RACE</td>
<td>Race</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Keep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>ETHNIC</td>
<td>Ethnicity</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Keep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>ARMCD</td>
<td>Planned Arm Code</td>
<td>Char</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>ARM</td>
<td>Description of Planned Arm</td>
<td>Char</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>ACTARM</td>
<td>Actual Arm Code</td>
<td>Char</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>ACTARM</td>
<td>Description of Actual Arm</td>
<td>Char</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>COUNTRY</td>
<td>Country</td>
<td>Char</td>
<td>Quasi Level 1</td>
<td>Elevate to continent</td>
<td>Keep</td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>DMOTC</td>
<td>Date/Time of Collection</td>
<td>Char</td>
<td>Quasi Level 2</td>
<td>Offset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special-Purpose</td>
<td>DM</td>
<td>DMDY</td>
<td>Study Day of Collection</td>
<td>Num</td>
<td>Quasi Level 2</td>
<td>No further de-identification</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+1300 variables

Dates
Low frequency & rare events
Recoding of unique identifiers
Handling of free-text variables
Extensible code lists
Geographical location
Sensitive data
Quasi identifiers to keep
PII of third-party

#PhUSE
Residual Risk Assessment
PhUSE Approach & Criteria

• “Because identifying a specific set of variables that need to be modified as per the general Safe Harbor approach does not guarantee that the risk of re-identification is always sufficiently small, a second step of residual risk analysis is generally recommended if any of the conditions below are met. There may be residual re-identification risk under certain conditions, such as:
  – the data is not being released through a secure portal with adequate privacy and security controls,
  – the data recipients do not sign a data sharing agreement that has sufficient limitations on what the recipients can and cannot do,
  – the trial is for a rare disease,
  – there are extreme values in the data set,
  – there are observable or knowable serious adverse events in the trial (e.g., deaths and suicides),
  – the data set has extensive demographic and socioeconomic information about the participants, or
  – the data set includes detailed medical histories of the participants.

• The sponsor can decide whether any of these conditions are met in making the determination about whether this additional residual risk assessment is required”
Residual Risk Assessment Methodology

- The evaluation of residual risk is a quantitative exercise and involves 4 general steps:
  1. Assessing the context of the data sharing.
  2. Setting an acceptable threshold for anonymizing the data.
  3. Measuring the actual probability of re-identification in the de-identified data.
  4. Adjusting data de-identification if necessary.
Residual Risk Assessment Example

Population:
- Study population
- Similar Clinical Studies population
- Geographical population

Risk Metrics:
- Average risk for controlled disclosure
- Maximum risk for public disclosure

<table>
<thead>
<tr>
<th>Gender</th>
<th>Year of Birth (10 years)</th>
<th>Population Group Size</th>
<th>Probability of Re-identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1970-1979</td>
<td>200</td>
<td>0.005</td>
</tr>
<tr>
<td>Male</td>
<td>1980-1989</td>
<td>110</td>
<td>0.009</td>
</tr>
<tr>
<td>Male</td>
<td>1970-1979</td>
<td>200</td>
<td>0.005</td>
</tr>
<tr>
<td>Female</td>
<td>1990-1999</td>
<td>80</td>
<td>0.0125</td>
</tr>
<tr>
<td>Female</td>
<td>1980-1989</td>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>Male</td>
<td>1990-1999</td>
<td>50</td>
<td>0.02</td>
</tr>
<tr>
<td>Male</td>
<td>1990-1999</td>
<td>50</td>
<td>0.02</td>
</tr>
<tr>
<td>Female</td>
<td>1980-1989</td>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>Male</td>
<td>1970-1979</td>
<td>200</td>
<td>0.005</td>
</tr>
<tr>
<td>Female</td>
<td>1990-1999</td>
<td>80</td>
<td>0.0125</td>
</tr>
<tr>
<td>Male</td>
<td>1980-1989</td>
<td>110</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Residual Risk Assessment

Benefits

• Applying rules does not guarantee that
  – Risk is small or
  – May lead to too much data de-identification

• Provides documentation and claim.

• Allow the release of highly granular data.
Guidance on the anonymisation of clinical reports for the purpose of publication in accordance with policy 0070

Industry stakeholder follow-up meeting, 23 June 2015
Agenda topic 6

Presented by Monica Dias
Policy Officer
Article 29 Working Party Opinion on anonymisation techniques

- Article 29 Opinion on anonymisation provides **two options** to establish if a dataset is anonymised:

  1. Demonstrate that after anonymisation it is no longer possible to:
     - *Singling out:* possibility to isolate some records of an individual in the dataset*;
     - *Linkability:* ability to link, at least, two records concerning the same data subject or a group of data subjects (in the same database or in two different databases);
     - *Inference:* the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes

   **OR**

  2. Perform an analysis of re-identification risk.

* In the context of phase 1 of policy 0070, dataset are the set of clinical reports published by the Agency
Legal framework and available standards

- EU data protection legislation
- Article 29 Data Protection Working Party opinion of anonymisation techniques (Opinion 05/2014)
- Information Commissioner’s Office (ICO) Code of Practice. Anonymisation: managing data protection risk
- Sharing clinical trial data: Maximizing benefits, minimizing risk. Institute of Medicine (IOM)
- Pharmaceutical Users Software Exchange (PhUSE) de-identification standards for CDISC SDTM 3.2
- Transcelerate BioPharma Inc., Clinical Study Reports Approach to Protection of Personal Data and Data De-identification and Anonymisation of Individual Patient Data in Clinical Studies – A Model Approach
Available for Download
Published on 15. May 2015

Pharmaceutical Users Software Exchange

Data Transparency Download

DISCLAIMER

This set of de-identification rules defined for CDISC SDTM 3.2 is written with the goal of both facilitating the assessment of direct and quasi-identifiers in SDTM datasets and ensuring consistency in anonymized data shared across sponsors.

The definitions of direct and quasi-identifiers and the decisions and concepts described in this deliverable represent the consensus of the working group rather than an endorsement of the companies represented in the working group.

However, the rules described here do not guarantee an acceptable or very small residual risk of re-identification in the data and it is the responsibility of the sponsors to define and measure what the residual risk is and define an acceptable risk threshold.

SDTM being also a normalized data model, not all direct nor quasi-identifiers may be captured in this deliverable and it is the responsibility of the sponsor to ensure that such assessment is conducted and reviewed according to defined internal procedures.

To download the documents, please enter your details below. By entering your details, PHUSE will keep you informed of future updates to these documents. In line with PHUSE’s Data Protection guidelines, PHUSE will not sell rent or lease to others your personally identifiable information.

Forename*: Jean-Marc
Surname*: Ferran
Email Address*: jmf@qualliance.dk
Company*: jmf@qualliance.dk

* Required Information

Please contact office@phuse.eu should you need any assistance.

#PhUSE
Questions?

Jean-Marc Ferran
Special Projects Director, PhUSE
E: jean-marcl.ferran@phuse.eu
The premier community for people working in the biometric area

@PhUSETwitter  PhUSE  www.phusewiki.org