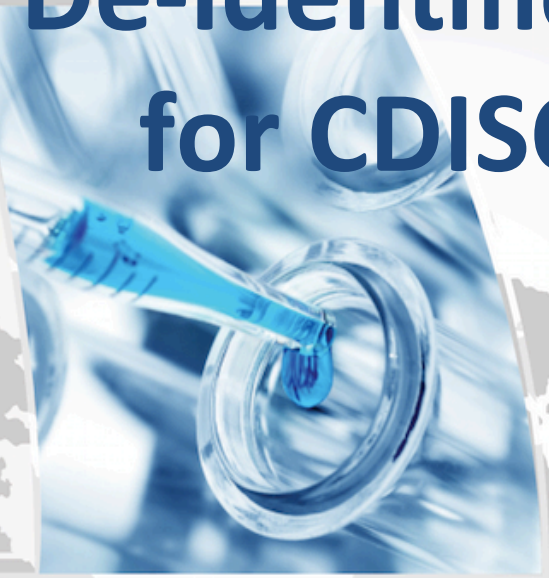
DH01
De-Identification Standards for CDISC Data Models



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)

PhUSE Data De-Identification Working Group Participants

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

Agenda

- Overview of Data Sharing
- Data De-identification Standards for SDTM 3.2
- Current EMA CSR anonymization guidance

Overview of Data Sharing

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

2 October 2014
EMA/240810/2013

European Medicines Agency policy on publication of clinical data for medicinal products for human use

POLICY/0070
Status: Adopted
Effective date: 1 January 2015
Review date: No later than June 2016
Supersedes: Not applicable

1. Introduction and purpose

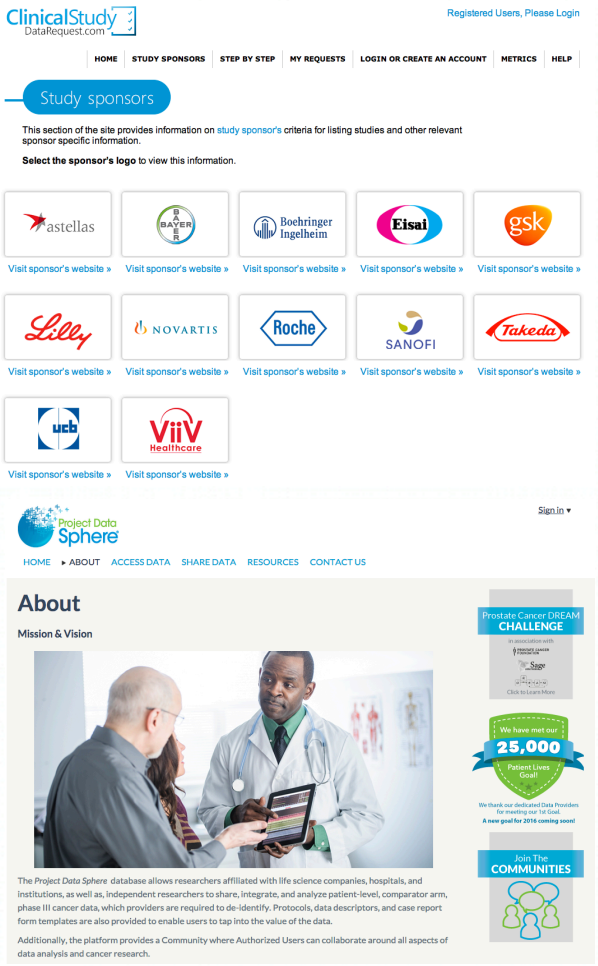
The aim of the European Medicines Agency ('the Agency') is to protect and foster public health. Transparency is a key consideration for the Agency in delivering its service to patients and society.

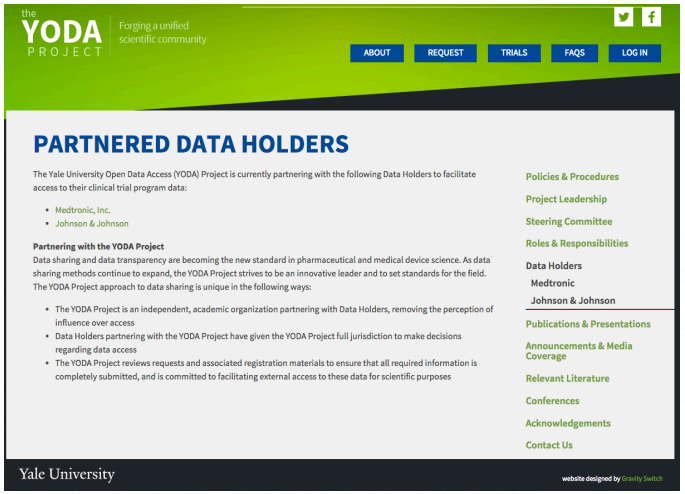
Although the Agency since its creation has launched several initiatives to increase transparency of information on medicinal products, there is growing demand from stakeholders for additional transparency, not only about the Agency's deliberations and actions, but also about the clinical data on which 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 80 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 1049/2001[2], and, therefore, it does not replace the existing 'Policy on access to documents (related to medicinal products for human and veterinary use)' (POLICY/0043) (EMA/110196/2006), which came into effect in December 2010. Moreover, the provisions of this policy are not intended in any manner to limit the application or the rights given by Regulation (EC) No. 1049/2001. 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.

[1] Regulation (EC) No 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down community procedures for the authorisation and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency.
[2] Regulation (EC) No 1049/2001 of the European Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents.

30 Churchill Place • Canary Wharf • London E14 5EU • United Kingdom
Telephone +44 (0)20 3660 6000 Facsimile +44 (0)20 3660 5555
Send a question via our website www.ema.europa.eu/contact

An agency of the European Union





#PhUSE

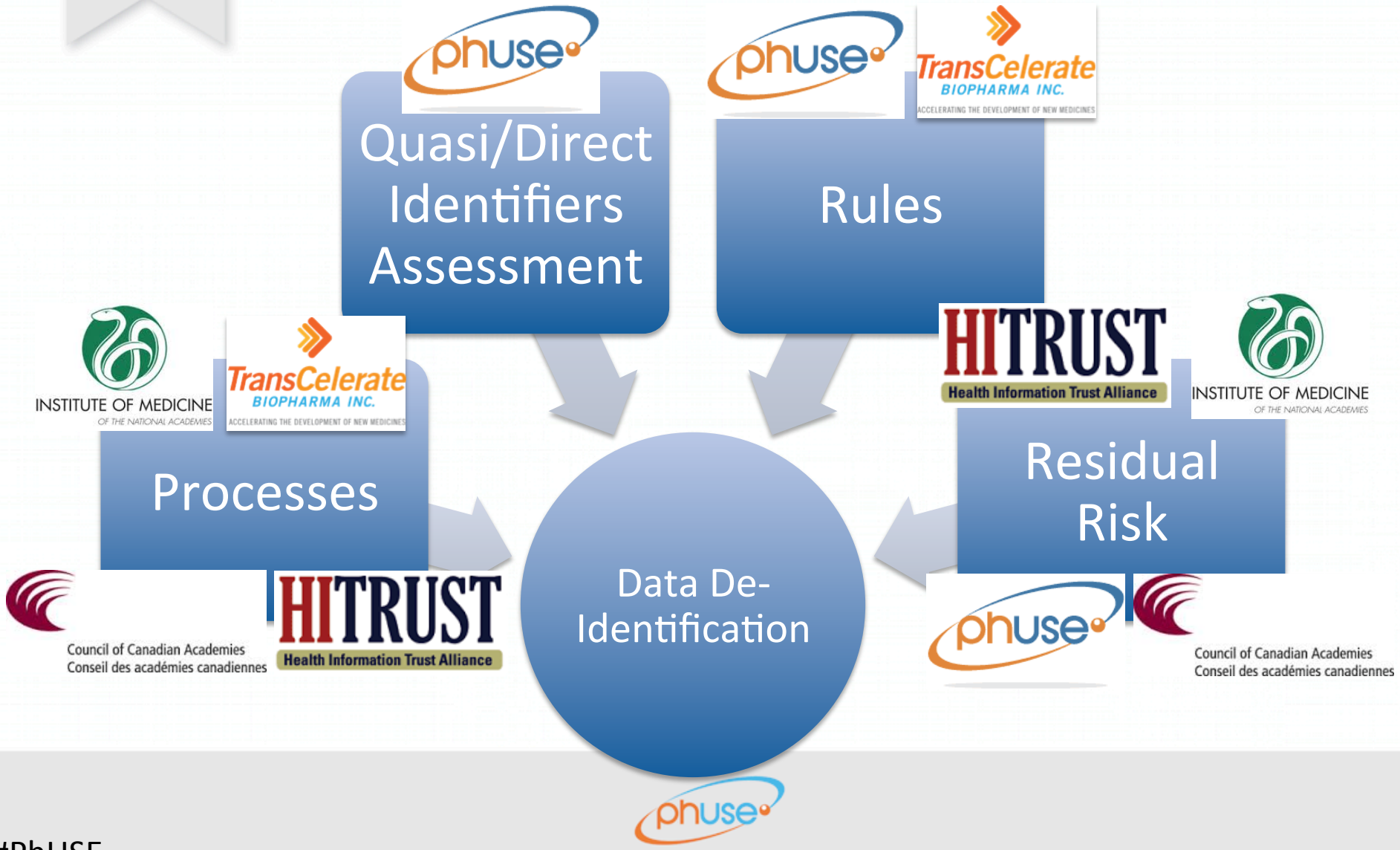
Metrics from ClinicalStudy DataRequest.com

Research proposals requesting access to patient level data (number of proposals)

Number of Research Proposals submitted up to 31 August 2015		160
Requirements check	In process	12
	Withdrawn by the requestor	18
	Did not meet requirements (further details)	11
	Potential conflict of interest or an actual or potential competitive risk	0
	Met requirements	119
IRP review	In process	2
	Withdrawn by the requestor	1
	Rejected or advised to re-submit	11
	Approved or approved with conditions	105
Data Sharing Agreement	In process	26
	Withdrawn by the requestor	2
	Not agreed (not signed)	0
	Agreed (signed) View details of these research proposals	77
Data preparation	In process	7
	Withdrawn by the requestor	1
	Complete (data available)	69
Research project	In process	69
	Withdrawn by the requestor	0
	Not published	0
	Published	0

Source:
clinicalstudydatarequest.com/Metrics.aspx
11OCT2015

Data De-Identification Guidelines



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.

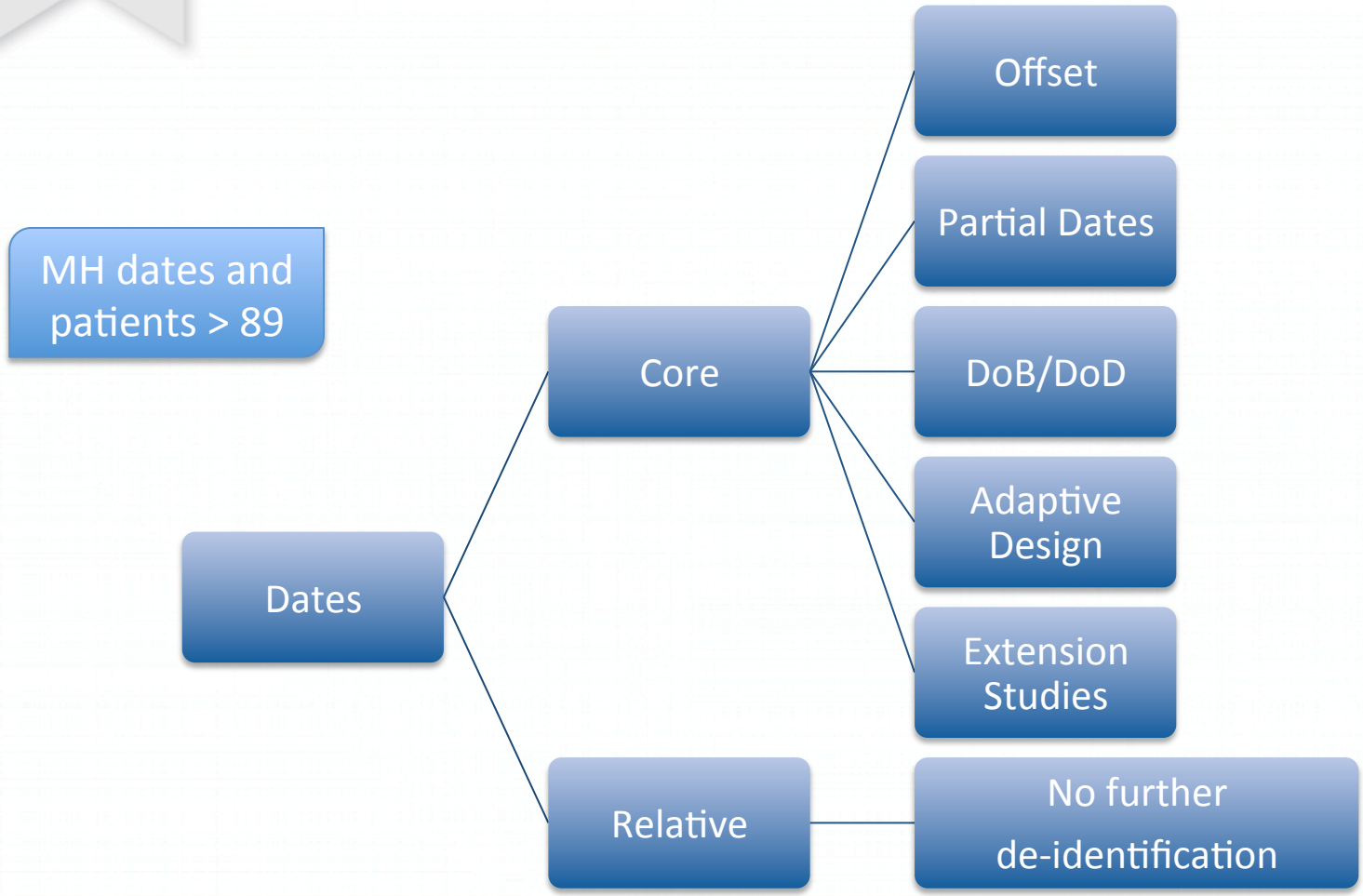
Primary & Alternative Rules for De-Identification are assigned

- **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

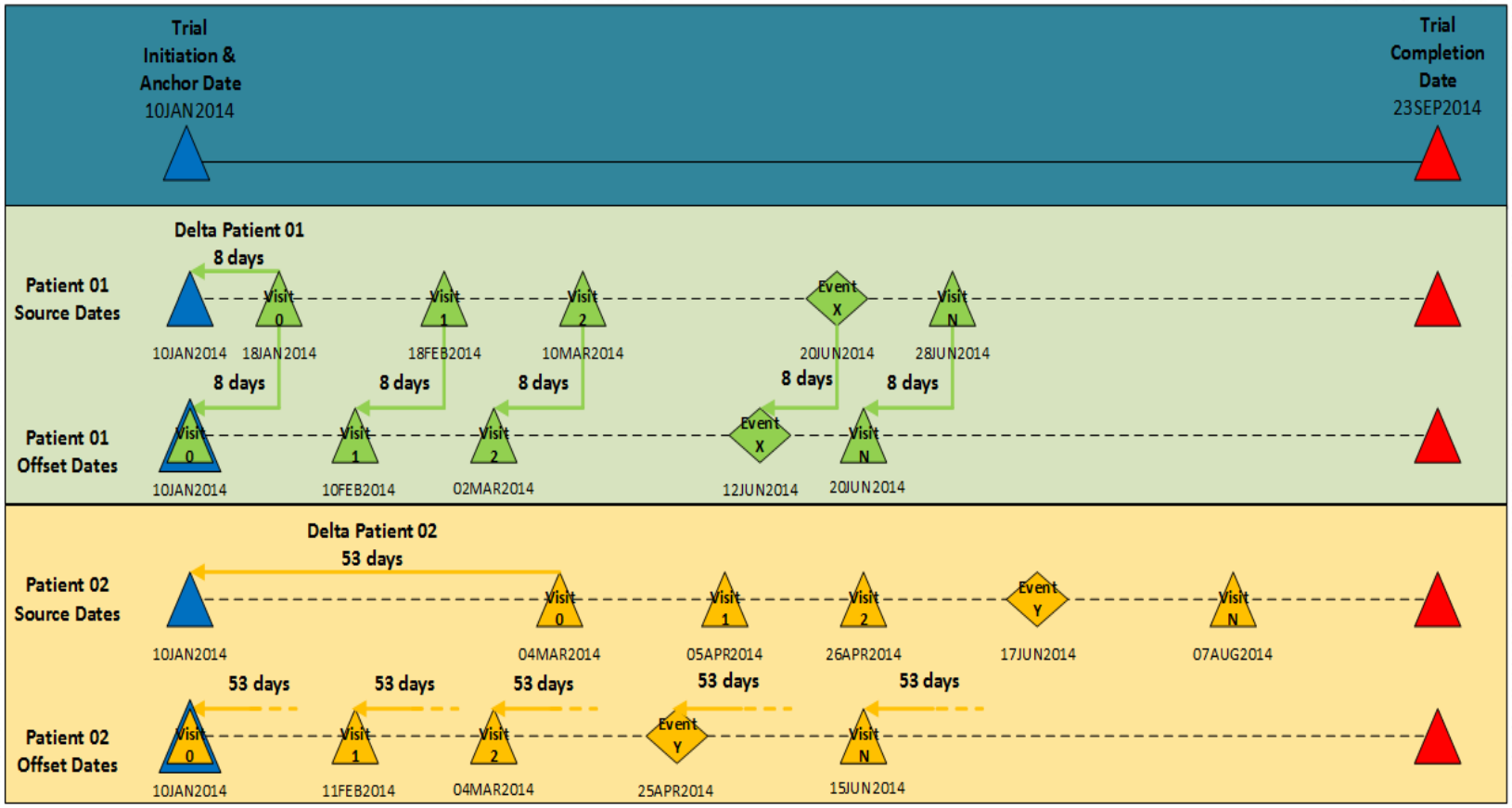
- To explain further the **rational of a given assessment**
- To warn users for **exceptions or special considerations**

Dates


MH dates and patients > 89
Dates
Core
Offset
Partial Dates
DoB/DoD
Adaptive Design
Extension Studies
Relative
No further de-identification

Dates Offset Recommended Algorithm

See Appendix 1

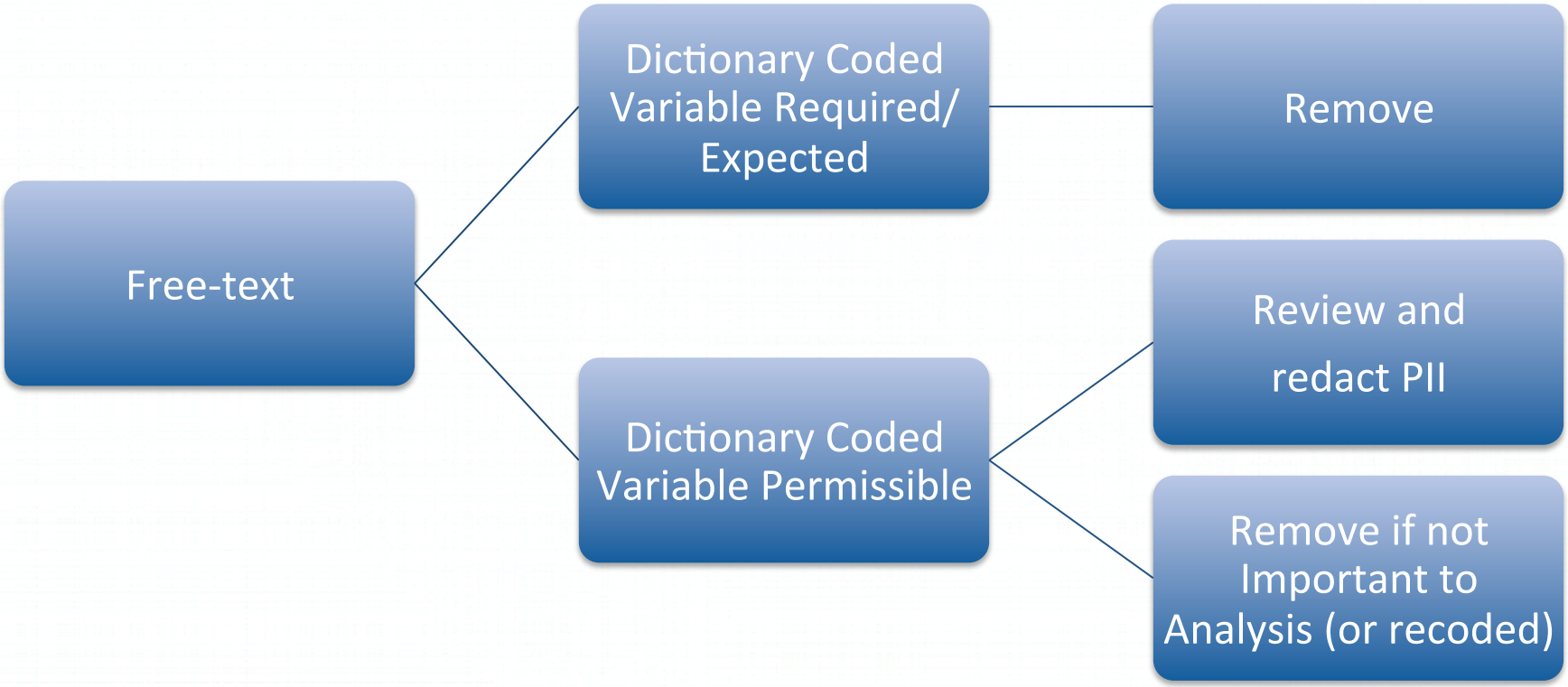


Issue with Partial Dates

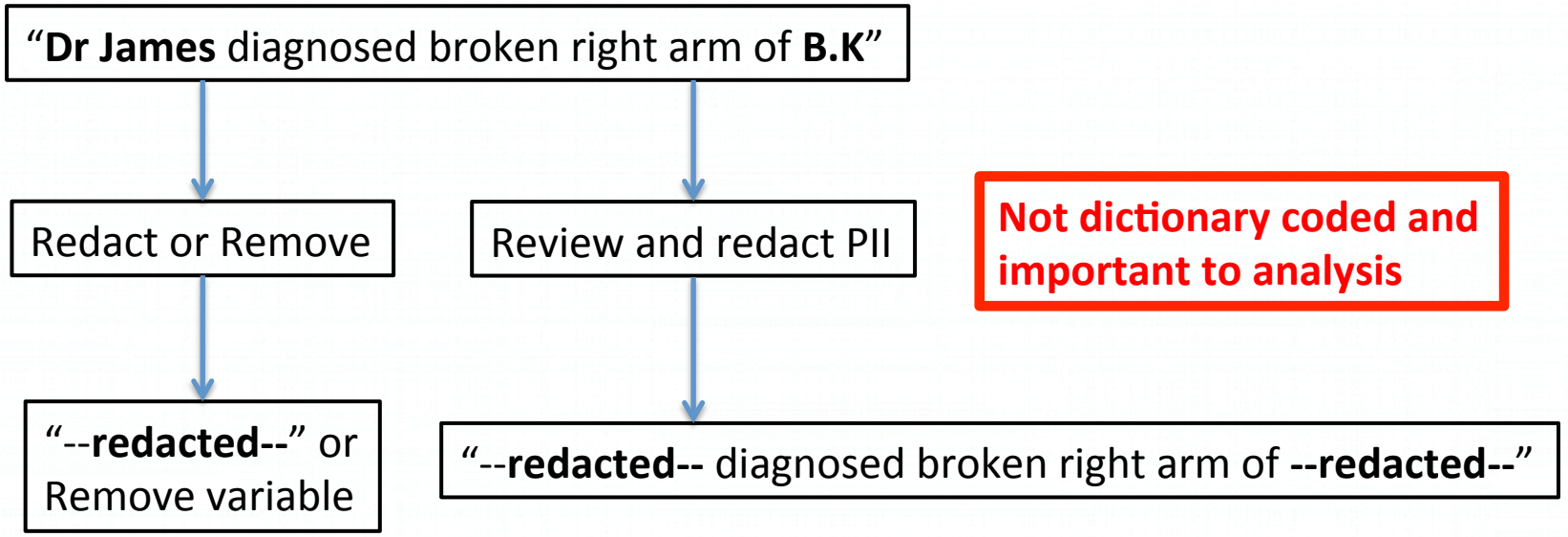
Ex: Delta applied of -14 days

Visit/Event	Date (Source)	Imputed Date	Offset Date	Offset Partial Date (Final)
Visit 0	10JAN2013	10JAN2013	27DEC2012	27DEC2012
Visit 1	10FEB2013	10FEB2013	27JAN2013	27JAN2013
Visit 2	08MAR2013	08MAR2013	22FEB2013	22FEB2013
Event X	MAR2013	15MAR2013	01MAR2013	MAR2013
Visit 3	12APR2013	12APR2013	29MAR2013	29MAR2013

Free-text

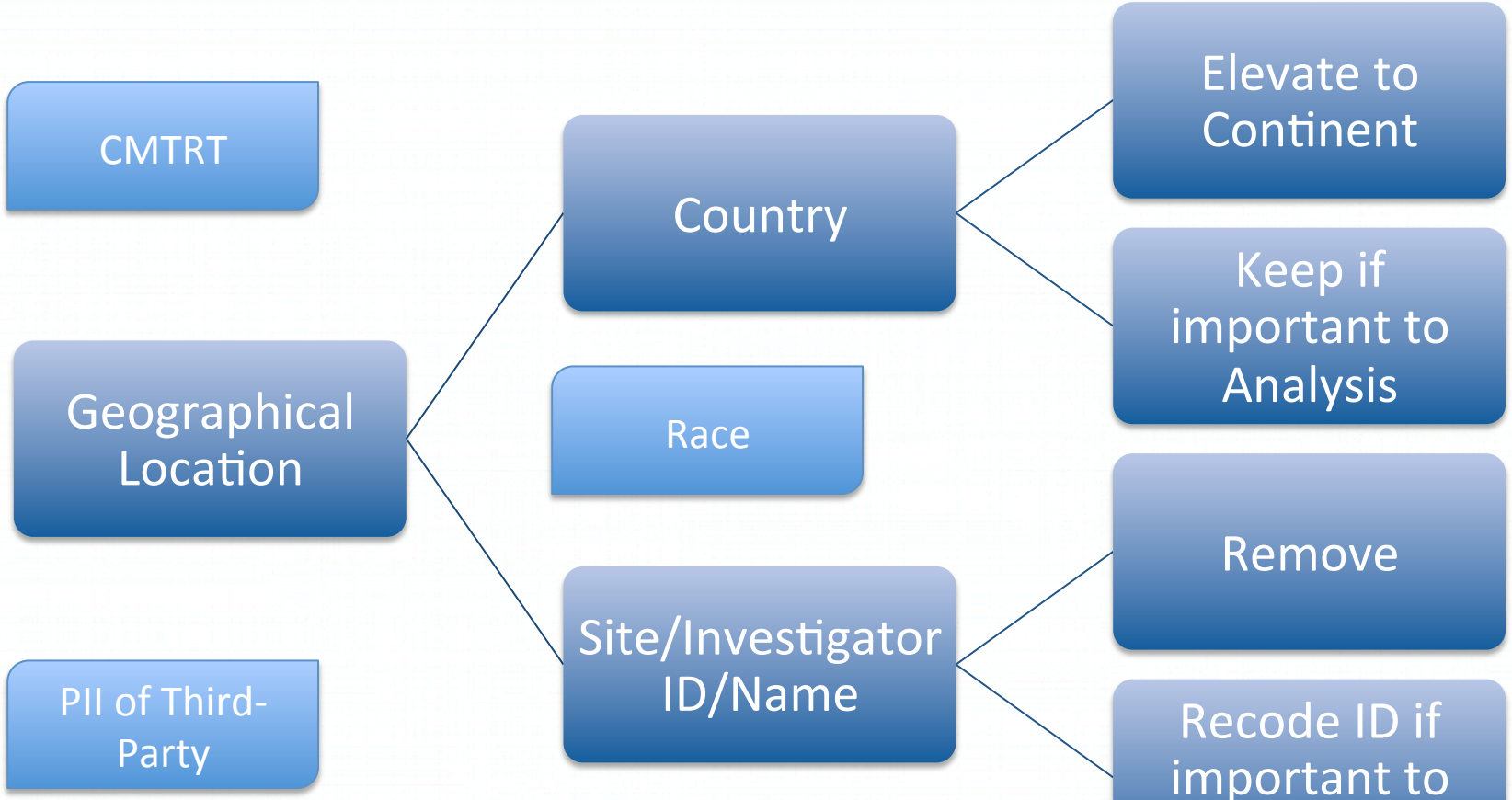

Free-text
Dictionary Coded Variable Required/ Expected
Remove
Dictionary Coded Variable Permissible
Review and redact PII
Remove if not Important to Analysis (or recoded)

Review and Redact PII in Free-Text



Not dictionary coded and important to analysis

Geographical Location



Deliverable

De-Identification Standards for CDISC SDTM 3.2

Observation_Class	Domain_Prefix	Variable_Name	Variable_Label	Type	Direct_Quasi_Identifier (Direct/Quasi	DI_Primary_Rul	DI_Alternative_Rule	DI_Comment
Special-Purpose	DM	RFPENDTC	Date/Time of End of Participation	Char	Quasi Level 2	Offset		
Special-Purpose	DM	DTHDTC	Date/Time of Death	Char	Quasi Level 1	Offset		In case of Fatal event, this may be considered for further de-identification for low-frequency of dead patients. This is the responsibility of the sponsor to conduct such assessment considering among other occurence of such death for the concerned subjects in the general population.
Special-Purpose	DM	DTHFL	Subject Death Flag	Char	Quasi Level 2	Keep		In case of Fatal event, this may be considered for further de-identification for low-frequency of dead patients. This is the responsibility of the sponsor to conduct such assessment considering among other occurence of such death for the concerned subjects in the general population.
Special-Purpose	DM	SITEID	Study Site Identifier	Char	Quasi Level 1	Remove	Recode ID variable	If SITEID is required and is recoded as per the alternative rule, it must be considered within the risk assessment.
Special-Purpose	DM	INVID	Investigator Identifier	Char	Quasi Level 1	Remove	Recode ID variable	If INVID is required and is recoded as per the alternative rule, it must be considered within the risk assessment.
Special-Purpose	DM	INVNAM	Investigator Name	Char	Quasi Level 1	Remove		Such information is related to other individuals than the patients and can also reveal geographic location of site. In addition, it holds little data utility.
Special-Purpose	DM	BRTHDTC	Date/Time of Birth	Char	Quasi Level 1	Remove		
Special-Purpose	DM	AGE	Age	Num	Quasi Level 1	Derive Age	Aggregate Age	
Special-Purpose	DM	AGEU	Age Units	Char				
Special-Purpose	DM	SEX	Sex	Char	Quasi Level 1	Keep		
Special-Purpose	DM	RACE	Race	Char	Quasi Level 1	Keep		If necessary remap to CDISC code lists and consider races with low frequency into a category "OTHERDI".
Special-Purpose	DM	ETHNIC	Ethnicity	Char	Quasi Level 1	Keep		
Special-Purpose	DM	ARMCD	Planned Arm Code	Char				
Special-Purpose	DM	ARM	Description of Planned Arm	Char				
Special-Purpose	DM	ACTARMCD	Actual Arm Code	Char				
Special-Purpose	DM	ACTARM	Description of Actual Arm	Char				
Special-Purpose	DM	COUNTRY	Country	Char	Quasi Level 1	Elevate to continent	Keep	If country is critical to the analysis (e.g. required to reproduce a result), it may be kept and it is the responsibility of the sponsor to assess whether the residual risk is acceptable and take further actions on other variables if necessary. Countries with less than 10 patients must be grouped in country OTHERDI.
Special-Purpose	DM	DMDTC	Date/Time of Collection	Char	Quasi Level 2	Offset		
Special-Purpose	DM	DMDY	Study Day of Collection	Num	Quasi Level 2	No further de-identification		

- 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

+1300 variables

#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

Gender	Year of Birth (10 years)	Population Group Size	Probability of Re-identification
Male	1970-1979	200	0.005
Male	1980-1989	110	0.009
Male	1970-1979	200	0.005
Female	1990-1999	80	0.0125
Female	1980-1989	100	0.01
Male	1990-1999	50	0.02
Male	1990-1999	50	0.02
Female	1980-1989	100	0.01
Male	1970-1979	200	0.005
Female	1990-1999	80	0.0125
Male	1980-1989	110	0.009

Population:

- Study population
- Similar Clinical Studies population
- Geographical population

Risk Metrics:

- Average risk for controlled disclosure
- Maximum risk for public disclosure

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**.

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

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

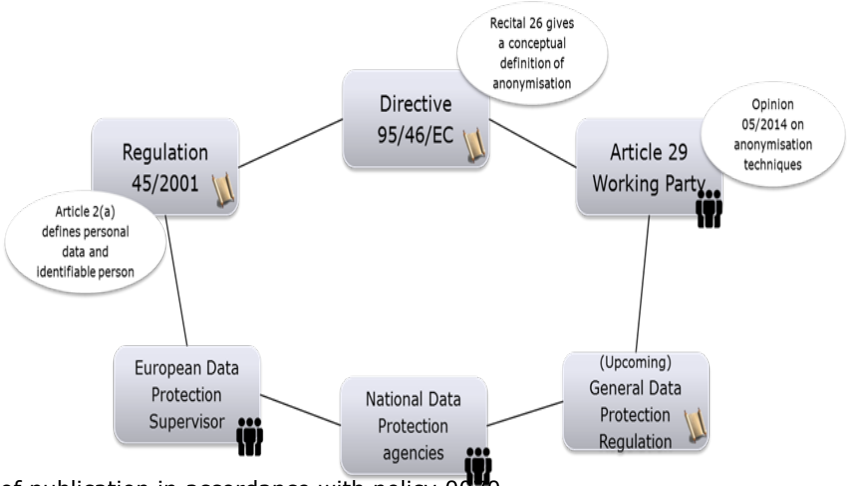
   OR

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

300+ downloads

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@qualiance.dk
Company*: jmf@qualiance.dk
Send

* Required information

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

Questions?

Jean-Marc Ferran

Special Projects Director, PhUSE

E: jean-marc.ferran@phuse.eu

W: http://www.phuse.eu/Data_Transparency.aspx


CELEBRATING 10 YEARS
phuse
SHARE – CONTRIBUTE – ADVANCE
2005 – 2014
phuse
The premier community for people working in the biometric area
@PhUSETwitta
PhUSE
www.phusewiki.org