

De-Identification Standards for CDISC Data Models

PhUSE Data Transparency Working Group

Kelly Mewes, Roche Products Limited

Jean-Marc Ferran, Qualiance

The PhUSE Data Transparency Working Group & Data Sharing

In this era of Data Transparency and sharing data with researchers, companies are defining their processes and de-identification guidance in order to comply with data privacy regulations. In particular, it is possible for researchers to request access to data across sponsors and both the difference of data models and de-identification techniques may make the analyses cumbersome and error-prone. As the CDISC data models are now adopted in the industry, PhUSE launched in July 2014 a dedicated Working Group to define de-identification standards for CDISC data models starting with SDTM. Participants from Pharmaceuticals, CROs, Software Vendors, CDISC specialists, Data Privacy specialists and Academia have joined forces to define a set of rules against SDTM to provide the industry with a consistent approach to data de-identification and increase consistency across anonymized datasets. Each domain and variables holding potentially Personally Identifying Information (PII) have been rated in terms of impact on data privacy. Based on that rating the variables are allocated standard rules of de-identification, along with the rationale and the impact on data utility being documented.

PhUSE De-Identification Standard for SDTM

Every domain and variable defined in SDTMIG 3.2 is assessed and variables that hold PII were evaluated in terms of data privacy and what rules to apply. Data privacy is defined across 3 levels, Direct Identifiers, Level 1 quasi Identifiers and Level 2 quasi Identifiers of decreasing impact on data privacy and risk over patient re-identification. For these variables, rules to apply are recommended. In some cases, it is recommended to keep them 'as-is' as they hold critical data for analysis (e.g. Sex) or an alternative rule is suggested to address different cases (e.g. a device number may need to be recorded in a medical device study while not holding any data utility in a non-device study, where it can be removed).

Please visit http://www.phuse.eu/Data_Transparency.aspx to download the latest version.

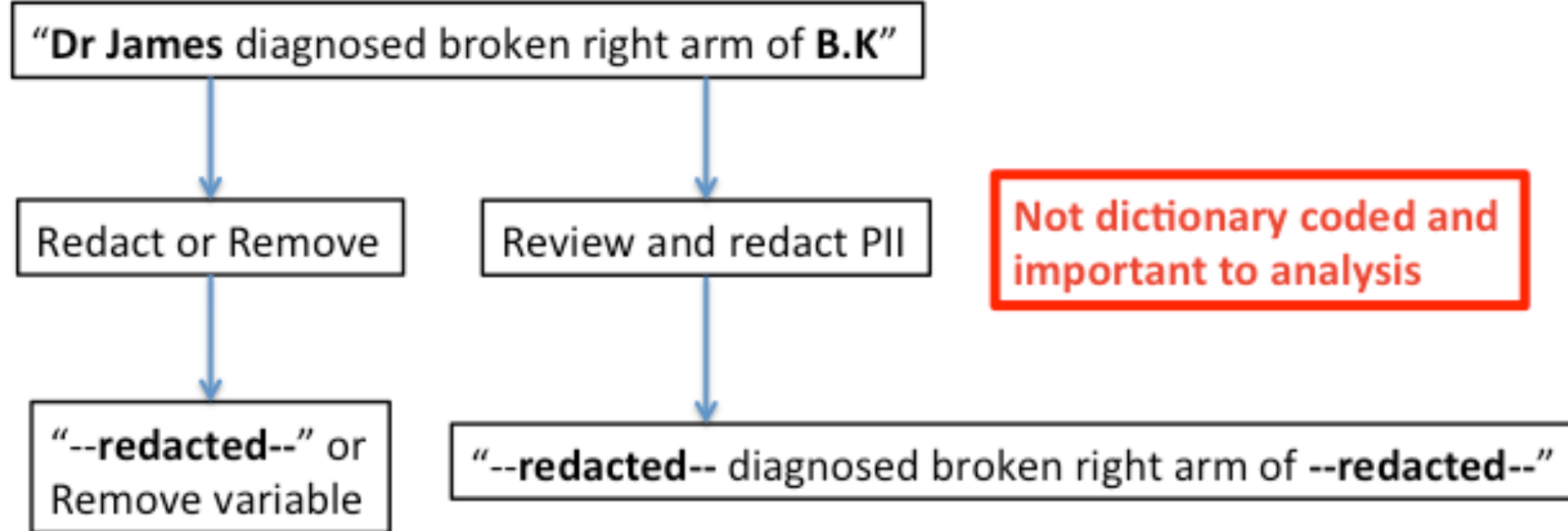
Observation Class	Domain Prefix	Variable Name	Variable Label	Type	Direct/Quasi Identifier (Direct/Quasi)	DI Primary Rule	DI Alternative Rule	DI Comment
Special-Purpose	DM	RFPENDTC	Date/Time of End of Participation	Char	Quasi Level 2	Offset		
Special-Purpose	DM	DTHTDC	Date/Time of Death	Char	Quasi Level 1	Offset		
Special-Purpose	D				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 occurrence of such death for the concerned subjects in the general population.
Special-Purpose	D				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	D				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	D				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 low frequency into a category 'OTHERDI'.
Special-Purpose	DM	ETHNIC	Ethnicity	Char	Quasi Level 1	Keep		
Special-Purpose	DM	ARMCD	Planned Arm Code	Char	Quasi Level 1	Keep		
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		

+1300 variables

400+ downloads!

Handling Free-Text

In general, free-text data must be deleted as it is considered to be a direct Identifier, as it can hold any data including PII. If there is no associated coded information available in the dataset that can be used instead, free-text data must be considered for review and redaction of values with PII. However such a measure needs to be balanced with the criticality of such variables for future research and as an alternative rule, a non-recoded free-text variable can be removed from the dataset.



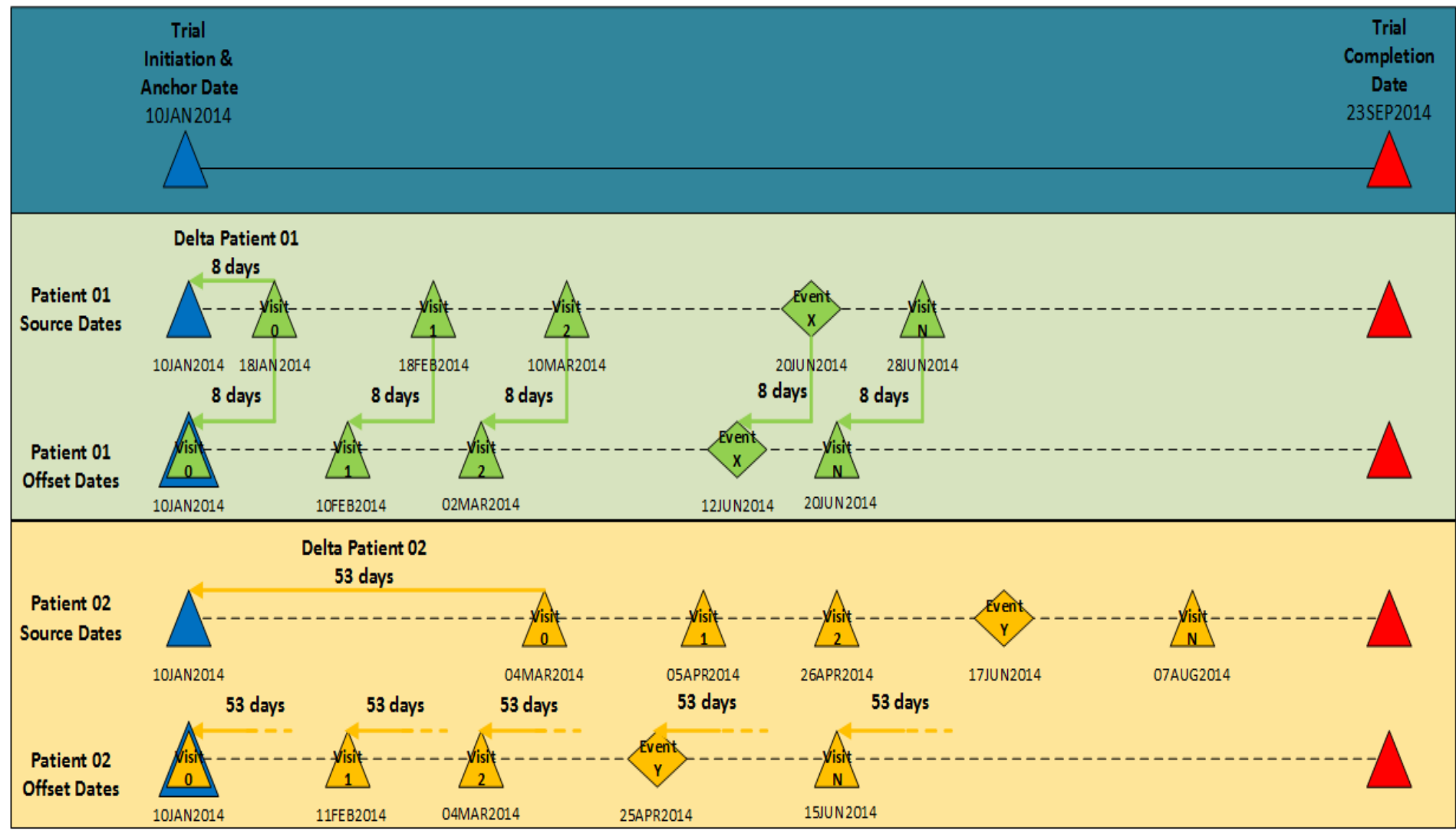
Data De-Identification

The PhUSE de-identification standards aim to:

- Facilitate the identification of direct and quasi identifiers
- Provide rules to apply together with technical guidelines.
- Ensure consistency in data de-identification across sponsors.
- Provide guidance in estimating residual risk in de-id data.

Offset Dates

All time-related information is important in clinical research, in particular dates, and they present themselves in two forms, date and relative days. Conforming with CDISC standards (both dates and relative days are present), it is preferable from a data utility perspective to keep both types of dates, but to offset actual dates, and keep study day and other relative dates 'as-is'. Full and partial dates must be offset (guidance is provided in Appendix 1 of deliverable [3]) while relative dates such Study Day can be kept 'as-is' in the datasets. The figure below describes a date offset algorithm that enables the offset dates to be kept within the range of possible study dates.



It was also decided that date of Death must be offset as in any other date handling, due to its importance in clinical research. Whilst date of Birth must be derived into "Age at baseline" and patients over 89 years old must be aggregated into one category. It is also possible as an alternative rule to derive into age folds (10-15, 15-20, 20-25 etc., 18-20, 20-22, 22-24, etc.) to be defined by the sponsor.

Other Areas

- Recoding of Unique Identifiers

- Geographical Location

- Low Frequency & Rare Events

- Sensitive Data

- Quasi Identifiers to Keep (e.g. Sex)

- PII of Third Party

Residual Risk Assessment

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 following conditions are met:

- Data is not released through a secure portal
- Researchers do not sign a Data Sharing Agreement
- Rare disease trial
- Extreme or known/observable values or serious events
- The data contains extensive demographic, socioeconomic or detailed medical histories.

The evaluation of residual risk is a quantitative exercise and involves 4 general steps:

- Assessing the context of the data sharing.
- Setting an acceptable threshold for anonymization.
- Measuring the actual probability of re-identification.
- Adjusting data de-identification if necessary.

Population to consider:

- Study population
- Similar Clinical Studies population
- Geographical population

Risk Metrics to consider:

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

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

Extension to CDISC ADaM

The PhUSE De-Identification Standard for SDTMIG 3.2 was released on 15 May 2015 and a number of companies are currently implementing it or conducting pilots. Based on the initial feedback that will be provided through this process, a version 1.1 of the standard will be developed to further clarify the different concepts and rules.

A natural step for the working group would be to develop a similar de-identification standard for analysis datasets and CDISC ADaM in particular. While most concepts developed for de-identifying SDTM datasets can be reused and most SDTM data is available in ADaM, a number of challenges may be posed:

- Ensuring consistency between de-identified SDTM and ADaM datasets

- Develop mechanism to evaluate and de-identify derived information that may contain data at risk whilst preserving data utility

- Preserve traceability between SDTM and ADaM data layers

Summary

This set of rules defined for CDISC SDTMIG 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 represent the consensus of 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.