Paper DH01

PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models

Jean-Marc Ferran, Qualiance & PhUSE
Khaled El Emam, Privacy Analytics & University of Ottawa
Sarah Nolan, University of Liverpool & The Cochrane Collaboration
Boris Grimm, Boehringer Ingelheim
Nick De Donder, Business & Decision Life Sciences

ABSTRACT

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.

While 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 domains 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, the rationale and the impact on data utility is documented.

This presentation will elaborate on the Working Group main findings, the current deliverables and the perspective to take this first initiative to the next stage.

Table of Content

INTRODUCTION

There are current efforts by regulators such as the EMA to examine how to make Individual Patient Data (IPD) from clinical trials shared more widely, such as for researchers. Sponsors have started sharing data based on request proposals from researchers. However, a number of challenges remain:

- Data is presented in different data models.
- Each company seems to be defining their own high-level guidelines for data de-identification (See "Study Sponsors" tab of [1]).
- It is uncertain whether these de-identification guidelines sufficient to protect privacy.
- It is sometimes necessary to meet the analysis objective to request data from different companies within same research proposal.
- It is uncertain whether the application of these de-identification standards will result in sufficient data utility to meet the analytic needs of researchers.
- The level of resources needed to de-identify data sets is uncertain.

Since its inception, PhUSE has been an advocate of sharing and exchanging knowledge and has been supporting the data transparency initiatives. A working group was set-up back in July 2014 with the goal of defining data de-identification standards for CDISC data models, and address some of the challenges noted above. The first deliverable is focusing on CDISC SDTM 3.2 [2] and was finalized in May 2015.

The PhUSE de-identification standards aim to:

1. Facilitate the identification of direct and quasi identifiers (See "Definitions" tab of [3]).
2. Provide rules to apply together with technical guidelines.
3. Ensure consistency in data de-identification across sponsors.
4. Provide guidance in estimating residual risk in de-identified data.

This paper elaborates on the approach, the key principles and important areas for data de-identification that are addressed in the PhUSE De-Identification Standards for CDISC SDTMIG 3.2 [3].

OVERVIEW OF DATA SHARING

The launch of the first data sharing platform, led by GlaxoSmithKline (GSK) in May 2013 [4], marked the beginning of a new era of data transparency within the Pharmaceutical Industry. Thirteen study sponsors have now committed to the sharing of participant level data from their clinical trials via multi-sponsor platform <ClinicalStudyDataRequest.com> (CSDR) [1]. CSDR provides a structured format for requesting data, including a step-by-step diagram, user guide, supporting guidance videos and the opportunity to communicate with the sponsor throughout the process. Functionalities of the platform include the ability to:

- select available studies from a list provided by the sponsor
- submit a research proposal for the studies required for the research, including statistical analysis plan for review by an independent review panel
- signing data sharing agreements by the researcher and sponsor for approved proposals
- remote access to requested de-identified datasets and all related documentation is provided via a secured SAS analytic environment

To September 2015, out of 165 research proposals submitted to CSDR and processed, 126 (76%) met initial requirements and 113 (68%) were approved by the independent review panel. Data sharing agreements have been signed for 83 proposals and de-identified datasets provided for 72 proposals. Sixteen multi-sponsor proposals have been submitted since January 2014 (11% of proposals over the time frame).

Other sponsors have opted for a single sponsor environment in contrast to the multi-sponsor format of CSDR. For example, Johnson & Johnson (J&J) announced an agreement with Yale University School of Medicine's Open Data Access (YODA) project in January 2014 in which YODA acts as an independent review panel for research proposals requesting access to J&J datasets [5] in a similar format to CSDR. By October 2015, all 18 fully reviewed research proposals for 76 datasets were approved by YODA for J&J datasets.

Both CSDR and YODA allow “enquiries” for datasets of studies which are not listed on the websites. In the spirit of transparency, full reasons are provided where enquires (and research proposals) result in a negative outcome. The range of studies which can be made available on request is at the discretion of the sponsor and is not standardized by the CSDR multi-sponsor platform.

To date, no successful research proposal to CSDR or YODA has resulted in a publication. This may reflect the originality and complexity of the research hypotheses proposed. Titles of approved CSDR and YODA research proposals range from individual participant data meta-analyses, development of prognostic, pharmacokinetic and genetic models, the development of novel statistical methodology, to investigation of adverse drug events and the design of new randomized controlled trials. While the provision of access to original datasets may have previously been associated with re-analysis to confirm validity of trial results [6], the actual focus of the approved proposals seems to be original research which can take several years to reach final publication stage. Potential restrictions relating to the remote analysis of highly de-identified data compared to planned analyses specified in the research proposal and the impact on subsequent publications are unknown.

Project Data Sphere (PDS) [7], launched in April 2014, and provides an alternative format to clinical trial data sharing. PDS allows researchers, whether independent or affiliated to industry, hospitals or academic institutions, access to historical, participant level, comparator arm phase III trial datasets and accompanying documentation with the aim of development and improvement of trial design and methodology, as well as acceleration of future research hypotheses. Datasets have been provided a by industry and CROs (Amgen AstraZeneca, Bayer, Celgene, Janssen Research and Development, Pfizer, Quintiles and Sanofi) and other clinical organizations (Memorial Sloan Kettering Cancer Center and the Alliance for Clinical Trials in Oncology). Unlike CSDR and YODA, access to de-identified datasets is granted to all approved researchers, without the need for a formal research proposal. While this format has the advantage of immediate access to datasets, PDS only has the control arm of trials rather than the full trial data set.

It must also be noted that CSDR and YODA initiatives fall under controlled data releases while PDS falls under semi-public data release, which are 2 different contexts. The PhUSE De-Identification Standard for CDISC SDTM 3.2 address differences of the risk profiles of these 2 contexts and provides guidance on how to measure the risk including thresholds to apply in Appendix 2 of the standard [3].

APPROACH

This PhUSE De-Identification Standards for CDISC SDTMIG 3.2 [3] has been written mainly in the context of the data transparency initiative in the pharmaceutical industry and assumes the following main requirements are also being met as part of the data disclosure process:

1. De-identified data is shared with researchers through a secure portal where the download and upload of data is controlled and is the responsibility of the sponsor.
2. Data sharing agreements are signed between sponsors and researchers, and these commit the researchers not to attempt to re-identify patients, download data or carry out analyses outside the approved research request, attempt to contact any of the participants or presumed participants, link the data with other data sets that they may have access to, and provide access to the portal to someone else.
3. The researchers have privacy practices in place at their institution.

In case of public or semi-public data sharing, stricter measures must be applied in terms of data de-identification. This is discussed in paragraph "Public Data" in Appendix 2 [3].

Approaches to data preparation, in addition to approaches of data-de identification, are variable across sponsors. For example, certain sponsors prefer to prepare and de-identify datasets when a dataset is requested according to the research proposal, which could potentially result in multiple forms of the same dataset, while other sponsors prefer to prepare analysis ready datasets and de-identify before listing as an available study for request. Currently all CSDR sponsors list their own varied de-identification standards, despite the possibility of a multi-sponsor request within a single research proposal. Of course, multi-sponsor data requests could create challenges pooling data or doing comparative analysis across trials if the de-identification methods used are not uniform.

The approach taken to data preparation and de-identification is likely to impact not only on the time taken from initial data request to provide access to the dataset, but also to the content of the data provided and hence the research objectives which can be addressed. Such issues become even more challenging in multi-sponsor requests and could potentially impede the synthesis of datasets in research such as meta-analysis. While generalized approaches to de-identification (i.e. the removal of all free text, offsetting of all dates) may impact upon the utility of the data provided, de-identification performed manually on a variable level (and in some cases a participant level) would require a large

commitment of time, resources and potentially cost for sponsors.

Therefore the PhUSE approach to the development of de-identification standards aims to ensure consistency of de-identification across sponsors, minimization of time and resource use for sponsors in order to perform de-identification and maximization of data utility for researchers while minimizing the risk of re-identification. Such a standardized approach within the SDTM model would introduce consistency across sponsor datasets, potentially allows for automated de-identification rule assignment of common variable names (prefix and suffix) and minimize the requirement for manual variable inspection. The de-identification standards have been developed for the CDISC SDTMIG 3.2 [2], a data model which is commonly adopted across industry, and considers both proactive data de-identification outside the context of an approved research request and reactive data de-identification within the context of a particular approved research request. For each variable identified as a direct or quasi identifier, both a primary and alternative rules are suggested. The primary rule was defined mostly in the context of pro-active data de-identification maximizing data utility. While the alternative rule was defined to address special cases and reactive data de-identification. It is assumed that sponsors verify the data utility is adequate in general for proactive data de-identification or for a given research request for reactive data de-identification.

It must also be noted that documents (e.g. CSR) disclosed together with the de-identified data must be redacted in a manner consistent with the data.

The following other aspects of data de-identification and data sharing are not directly addressed in the deliverable and are the responsibility of the sponsors to define. These items are important and need to be implemented by the sponsor in conjunction with this standard:

1. The processes that support data transparency implementation (Use of Independent Review Boards, deletion of the keys, provision of full database or subset fitted to the request, etc.). Note that TransCelerate [8] provides guidance in these areas.
2. The specific process to assess residual risk in de-identified data. One methodology is described in more detail in the IOM report [9] and some guidance is provided in section "Low frequency" in "Decisions" tab and Appendix 2 of the deliverable [3].
3. The actual documentation of the de-identification that was applied.
4. The definition and approach to the de-identification of sensitive data.

The definitions, decisions and assessments in the deliverable represent the consensus of the working group.

PHUSE DE-IDENTIFICATION STANDARDS FOR SDTMIG 3.2

DELIVERABLE STRUCTURE

The deliverable consists of an MS Excel spreadsheet [3] with different tabs:

- Cover tab: Document information.
- Intro tab: Introduction including background, important considerations, out of scope, disclaimer and approach.
- Definitions tab: List of important terms with their definitions and examples when applicable.
- Decisions tab: Important areas with rationale for decisions.
- Rules tab: The different rules to be applied together with technical guidance.
- SDTMIG tab: The SDTMIG 3.2 variables (1300+) together with their assessment for direct/quasi identifiers, Primary rules, Secondary rules and Comment for De-Identification. See Figure 1.
- References tab: Sources used for the elaboration of the deliverable.
- Appendices tab: Appendices for guidance on “Dates Offset” and “Low Frequency”
- Change log tab: Different versions of the deliverable together with list of changes.

The large number of variables to assess justified the choice of MS Excel as support media, also in the perspective of using the deliverable to automatize the data de-identification using software.

Every domain and variable defined in SDTMIG 3.2 (See SDTMIG tab of [3]) 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 (see the Definitions tab of [3]) of decreasing impact

on data privacy and risk over patient re-identification. For these variables, rules to apply are recommended (see the Rules tab of [3] for details). 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 recoded in a medical device study while not holding any data utility in a non-device study where it can be removed). Note that columns A to L of the SDTMIG tab of [3] come from CDISC SDTM 3.2 Implementation Guide [2]. The topics that the working group addressed and the associated decisions are included in the "Decisions" tab.

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

A	B	C	D	E	F	G	H	I	J	K	L	N	O	P	Q
Seq_Fo r	Observa tion_Cla ss	Domai n_Pref ix	Variable_Name	Variable_Label	Type	Length	Controlled_T erms_or_For mat	Role	CDISC_Notes	Core	References	Direct_Quasi_I dentifier (Direct/Qu	DI_Primary_R	DI_Alternative_Ru	DI_Comment
970	All classes		--EVINTX	Evaluation Interval Text	Char	20		Timing	Evaluation interval associated with an observation, where the interval is not able to be represented in ISO 8601 format. Examples: LIFETIME, LAST NIGHT, RECENTLY, OVER THE LAST FEW WEEKS.			Quasi Level 2	No further de-identification		Low frequencies in a very special (not standardised) interval could lead to a higher probability of identification
1	Special-Purpose	DM	STUDYID	Study Identifier	Char	20		Identifier	Unique identifier for a study.	Req					
2	Special-Purpose	DM	DOMAIN	Domain Abbreviation	Char	2	(DOMAIN)	Identifier	Two-character abbreviation for the domain.	Req					
3	Special-Purpose	DM	USUBJID	Unique Subject Identifier	Char	20		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product. This must be a unique number, and could be a compound identifier formed by concatenating	Req		Direct	Recode subject ID		
4	Special-Purpose	DM	SUBJID	Subject Identifier for the Study	Char	20		Topic	Subject identifier, which must be unique within the study. Often the ID of	Req		Direct	Recode subject ID		
5	Special-Purpose	DM	RFSTDTC	Subject Reference Start Date/Time	Char	20	ISO 8601	Record Qualifier	Reference Start Date/time for the subject in ISO 8601 character format. Usually equivalent to date/time when subject was first exposed to study treatment. Required for all randomized subjects; null for all subjects who did not meet the milestone the date requires, such as screen failures or unassigned subjects.	Exp		Quasi Level 2	Offset		
6	Special-Purpose	DM	RFENDTC	Subject Reference End Date/Time	Char	20	ISO 8601	Record Qualifier	Reference End Date/time for the subject in ISO 8601 character format. Usually equivalent to the date/time when subject was determined to have ended the trial, and often equivalent to date/time of last exposure to study treatment. Required for all randomized subjects; null for screen failures or unassigned subjects.	Exp		Quasi Level 2	Offset		
7	Special-Purpose	DM	RFXSTDTC	Date/Time of First Study	Char	20	ISO 8601	Record	First date of exposure to any protocol-specified treatment or therapy, equal	Exp		Quasi Level 2	Offset		
8	Special-Purpose	DM	RFXENDTC	Date/Time of Last Study Treatment	Char	20	ISO 8601	Record Qualifier	Last date of exposure to any protocol-specified treatment or therapy, equal to the latest value of EXENDTC (or the latest value of EXSTDTC if EXENDTC was not collected or is missing).	Exp		Quasi Level 2	Offset		
9	Special-Purpose	DM	RFICDTC	Date/Time of Informed Consent	Char	20	ISO 8601	Record Qualifier	Date/time of informed consent in ISO 8601 character format. This will be the same as the date of informed consent in the Disposition domain, if that protocol milestone is documented. Would be null only in studies not collecting the date of informed consent.	Exp		Quasi Level 2	Offset		
10	Special-Purpose	DM	RFPENDTC	Date/Time of End of Participation	Char	20	ISO 8601	Record Qualifier	Date/time when subject ended participation or follow-up in a trial, as defined in the protocol, in ISO 8601 character format. Should correspond to the last known date of contact. Examples include completion date, withdrawal date, last follow-up, date recorded for lost to follow up, or death date.	Exp		Quasi Level 2	Offset		
11	Special-Purpose	DM	DTHDTC	Date/Time of Death	Char	20	ISO 8601	Record Qualifier	Date/time of death for any subject who died, in ISO 8601 format. Should represent the date/time that is captured in the clinical-trial database.	Exp		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.
12	Special-Purpose	DM	DTHFL	Subject Death Flag	Char	1	(NY)	Record Qualifier	Indicates the subject died. Should be Y or null. Should be populated even when the death date is unknown.	Exp		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.
13	Special-Purpose	DM	SITEID	Study Site Identifier	Char	20		Record	Unique identifier for a site within a study.	Req		Quasi Level 1	Remove	Recode ID variable	If SITEID is required and is recoded as per the alternative rule, it must
14	Special-Purpose	DM	INVID	Investigator Identifier	Char	20		Record Qualifier	An identifier to describe the Investigator for the study. May be used in addition to SITEID. Not needed if SITEID is equivalent to INVID.	Perm		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.
15	Special-Purpose	DM	INVNAM	Investigator Name	Char	20		Synonym Qualifier	Name of the investigator for a site.	Perm		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.
16	Special-Purpose	DM	BRTHDTC	Date/Time of Birth	Char	20	ISO 8601	Record Qualifier	Date/time of birth of the subject.	Perm		Quasi Level 1	Remove		
17	Special-Purpose	DM	AGE	Age	Num	8		Record Qualifier	Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy concerns).	Exp		Quasi Level 1	Derive Age	Aggregate Age	
18	Special-Purpose	DM	AGEU	Age Units	Char	10	(AGEU)	Variable Qualifier	Units associated with AGE.	Exp					
19	Special-Purpose	DM	SEX	Sex	Char	1	(SEX)	Record Qualifier	Sex of the subject.	Req		Quasi Level 1	Keep		
20	Special-Purpose	DM	RACE	Race	Char	100	(RACE)	Record Qualifier	Race of the subject. Sponsors should refer to "Collection of Race and Ethnicity Data in Clinical Trials" (FDA, September 2005) for guidance regarding the collection of race (http://www.fda.gov/RegulatoryInformation/Guidances/ucm126340.htm) See Assumption below regarding RACE.	Exp		Quasi Level 1	Keep		If necessary remap to CDISC code lists and consider races with low frequency into a category "OTHERDI".
21	Special-Purpose	DM	ETHNIC	Ethnicity	Char	20	(ETHNIC)	Record Qualifier	The ethnicity of the subject. Sponsors should refer to "Collection of Race and Ethnicity Data in Clinical Trials" (FDA, September 2005) for guidance regarding the collection of ethnicity (http://www.fda.gov/RegulatoryInformation/Guidances/ucm126340.htm).	Perm		Quasi Level 1	Keep		

Figure 1 - PhUSE De-Identification Standards for SDTMIG 3.2

KEY PRINCIPLES

The PhUSE Data De-Identification Standards for SDTMIG 3.2 provides support in:

- Assessing the role of each variable in term of identification potential, quasi and direct identifiers across the SDTM domains
- Assigning rules for de-identification
- Understanding rationale and address exceptions and special considerations

Direct & Quasi Identifiers are assessed

- **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 provide a consistent substitute value for 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**: Rules maximizing data utility (at the exception of geographic location information that is key to decrease the residual risk) for pro-active data de-identification
- **Alternative rule**: Rules addressing a particular request in the case of 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 **rationale of a given assessment**
- To warn reader for **exceptions or special considerations**

IMPORTANT AREAS TO CONSIDER

This section addresses the different important areas to consider, the rationale for the decisions that were made and the rules associated with the associated SDTM variables.

A number of variables types are at stake with regards data de-identification. HIPAA [10] provides guidance in what type of data to consider (Safe Harbor 18 identifiers). Clinical data typically holds more PII and these 18 identifiers are also extended to variables such as Race, Ethnicity, etc. In addition, Hrynaszkiewicz et al. [11] provides more guidance specific to clinical data.

As part of the development of this standard, the working group reviewed requirements from HIPAA [10], existing and available sponsors' anonymization standards ([1], [12]), ICO's Anonymization Code of Practice [13] and methods proposed in the literature ([9], [11]). Furthermore, there has been continuous collaboration with the TransCelerate Working Group responsible for developing their anonymization guidance document [8].

Although data transparency is outside the scope of submissions of data to health authorities (HA), maintaining CDISC compliance of de-identified data is a nice-to-have, there are a number of scripts that have been and are being developed assuming that datasets are CDISC compliant, and these scripts could then be reused.

The spirit should be to preserve as much data as possible to make sure the data remains usable.

Dates

All time-related information is important in clinical research in particular dates, they present them self in two forms, date and relative days. Conforms with CDISC standards (both dates and relative days are present), it is preferable from a data utility perspective to keep both types of dates, describe how to offset 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.

It was also decided that Date of Death must be offset like any other dates considering its importance in clinical research. It is also flagged in the deliverable what variables can indicate death (e.g. AEOUT) and should be considered if further de-identification is required.

In particular, 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. Dates indicative of age >89 years, e.g., year of disease diagnosis or year a prior medication was started must be replaced by "--redacted--" (e.g. in MH or CM domains).

Recoding of unique identifiers

Subject IDs, Reference IDs, Sponsor IDs, Investigator IDs and Site IDs must be recoded.

In particular, a new random unique subject ID must be created that is not made up of any identifiable information. Site numbers must not be replicated in the recoded subject IDs. The list of original subject IDs and the recoded ones must not have any values in common. Same recoded subject ID must be used in extension study data.

Variables such as Reference ID or Sponsor ID are usually constructed using CRF page numbers or laboratory sample numbers, which are Direct Identifiers and require recoding or deletion (if only operational and are not necessary to link data across datasets). The list of original IDs and the recoded ones must not have any values in common. This applies also to Investigator ID and Site ID, among others, when applicable.

Low Frequency & Rare Events

The concept of low frequency as a means to evaluate re-identification risk is often cited and used. Low frequency is one way to measure risk. The term has the same meaning as `minimal cell size" and "maximum risk" [9]. Maximum risk is a conservative way to measure risk and is more suited to public data release. For non-public data release, which is more congruent with the manner in which data will be disclosed under many clinical trial transparency initiatives, a more suitable way to measure risk is using the concept of "average risk" [9]. Average risk is less

conservative and allows the disclosure of more detailed information. Another important consideration is that the actual value of "low frequency" needs to be computed from the population.

More guidance on the topic is available in Appendix 2 of the PhUSE standard [3].

Handing of Free-text variables and Extensible code-lists

In general, free-text data must be deleted as free-text data 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 variable for future research and as an alternative rule, a non-recoded free-text variable can be removed from the dataset.

For variables that are supported by extensible code lists, extra values can be added as free-text. While free-text is considered in the context of data transparency and sharing data with researchers as uncontrolled and at risk, in this particular case, there is often an extensive list of values available from the CDISC controlled terminology. As part of the data management process, sponsors would detect, query and update free-text values that would not be appropriate to keep if they include PII. Such variables are not marked as Direct or Quasi Identifiers in the SDTMIG tab, but for precaution, we assigned the rule "Review and only redact values with personal information" like for free-text variables.

Geographic Location

Country is an important Level 1 quasi Identifier and in order to decrease residual risk by default (i.e. in the case of proactive data de-identification), country is advised to be aggregated to continent as primary rule unless critical to the analysis (e.g. Country is a fixed-effect factor in a statistical model and the results cannot be reproduced). The alternative rule is to keep country as-is. In particular the rule described in HIPAA [10] within the Safe Harbor method for geographical location is based on an empirical analysis performed by the US Census Bureau and may not be applicable globally.

Site ID and Investigator ID need to be removed because a frequency analysis would likely reveal the most highly recruiting site in a country/region (which by definition would include many of the participants). The alternative rule is to recode Site ID and Investigator ID if required for the analysis and in this case, it should be considered within the risk assessment.

Sensitive data

Although sensitive data (see Definitions tab of [3]) may not necessarily be PII and re-identifying, sensitive data may need to be deleted so that in case of data breach, further measures have already been taken. In principle if the data is not personal data anymore, such data could also be kept. This is the responsibility of the sponsors to decide what risk they are willing to take.

Variables and datasets at stake have a comment associated with such considerations.

PII of third-parties

PII of third parties (Laboratory name or address, Investigator name, etc.) must be removed from all datasets as it may provide geographical information about the patients and also could compromise the privacy of the third parties themselves. However third party roles may be kept as they can be an important factor for the analysis.

Other important quasi identifiers for data analysis

While a variable has been identified as a Quasi Identifier, it may be advisable to keep the variable as-is if it is judged critical for analysis and clinical research in general (e.g. Level 1 quasi Identifier Sex). The rule "Keep" is to document clearly that no action should be taken although the variable should be considered in computing the residual risk and may require in some cases further de-identification (see DI_comment column of SDTMIG tab in [3]).

Also in the case of dates, while visit dates are important information with regards to data privacy, e.g. --DY/VISIT/VISITNUM/VISITDY could help re-identify all visit dates should only one be found out since planned relative dates indicates when during the study the visit is planned to occur. They are classified as Level 2 quasi identifiers and assigned the rule "No further de-identification" meaning that it is not advised to apply further de-identification (they already represent de-identified variable) but are flagged for consistency. This rationale is applied in particular to all relative dates (actual and planned).

RESIDUAL RISK ANALYSIS

The PhUSE approach blends both of the Safe Harbor and Expert Determination perspectives by first identifying a specific set of variables that need to be modified as per the general Safe Harbor approach. Because this 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 [3]. 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.

More details on the residual risk analysis is available in Appendix 2 of the PhUSE de-identification standard for SDTM 3.2 [3] and the IOM report [9].

RESIDUAL RISK ANALYSIS METHODOLOGY

One of the main benefits of the PhUSE standard is that it classifies the variables in a data set into direct identifiers and quasi-identifiers. Direct identifiers (i.e. name, unique identification code or number) which uniquely identify an individual, while quasi identifiers (e.g. age, gender, race, clinical events) can re-identify an individual when used in combination or connection with other information. This classification is a necessary foundation for deciding how to anonymize the information in the data set.

A certain set of transformations are applied to direct identifiers to anonymize them, and a different set of transformations are applied to quasi-identifiers. A primary and secondary de-identification rule for each SDTM 3.2 variable (with comments where appropriate) is provided accordingly.

The application of the rules is a good start, but cannot guarantee that the risk of re-identification is acceptably small. The implication is that there may be residual re-identification risk after the application of the rules. This is where the guidance on residual risk assessment in the PhUSE standard is important.

The evaluation of residual risk is a quantitative exercise. It involves three general steps:

1. Assessing the context of the data sharing. For example, sharing data through a read-only portal access is a different context than providing researchers raw data that they can download onto their own machines.
2. Setting an acceptable threshold for anonymizing the data. This threshold, which is actually a probability of re-identification, is determined by the risk assessment described above.
3. Measuring the actual probability of re-identification in the data. There are multiple models and estimators that have been developed for doing so using the quasi-identifiers. These models and estimators make different assumptions about the data and the plausible adversaries.

Once the risk is measured then it can be compared to the threshold to determine if the actual residual risk is acceptably small. If it is then the data set can be declared to be de-identified. If it is not, then additional manipulations would need to be applied to the data to reduce that measured risk value, for example, further generalization of some of the variables.

An example on how to compute the residual risk analysis is provided in Appendix 2 of the standard document. In Figure 2, the individual risk of re-identification computations are displayed based on population group size associated with the remaining quasi identifiers (here Gender and Year of Birth in 10-year folds). The population group can be the trial population, the similar clinical trials conducted in the same period or the geographic population. While using the trial population would lead to conservative estimates, using other assumptions would enable a finer risk analysis that may enable to release more granular data.

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

Figure 2 - Example of Individual Patient Re-Identification Risk Computation

These individual probabilities of re-identification can be combined using the average risk metrics or the maximum may be considered. It is usually recommended to use the average for controlled data disclosures while the maximum risk metrics is more appropriate to public or semi-public data disclosures. Thresholds of 0.20 and 0.09 are usually recommended for controlled and public/semi-public data disclosures respectively [14].

As has been alluded in recent anonymization guidance from the European Medicines Agency, it is recommended to also produce a residual risk analysis report to document the methods and results used to measure risk and further anonymize the data. If no residual risk was found, then the report would document that as well as justifications for that determination.

The benefits of performing such a residual risk analysis are twofold. First, it provides evidence to regulators that the risk of re-identification for this particular data set is indeed low. This allows the sponsor to meet her legal obligations. Second, it also can allow the release of highly granular data. For example, it may be determined that the release of the country variable is permissible after a residual risk analysis is completed.

LOW FREQUENCY

The concept of low frequency is related to residual risk measurement. Trial participants with quasi-identifier values that are uncommon in the population (i.e., have a low frequency) are easier to re-identify and therefore have a higher risk of re-identification. This means that participants that are part of a small group on their quasi-identifier values are potentially problematic.

How uncommon does a participant have to be? This will depend on two factors. The first is how the actual risk is measured. There are multiple risk metrics, such as maximum risk and average risk. They are still probabilities but they are computed quite differently depending on the data release context. The second factor is the threshold, which was discussed in the section on residual risk analysis.

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 develop 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 (i.e. handling of imputed dates while offsetting across data layers).
- Develop mechanism to evaluate and de-identify derived information that may contain data at risk while preserving data utility.
- Preserve traceability between SDTM and ADaM data layers

CONCLUSION

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.

Hrynaszkiewicz et al. [11] suggests that if more than 2 quasi identifiers are left in the dataset, a risk assessment must be carried out. The recent report from the Institute of Medicine [9] suggests that this rule of 2 may not be enough in certain cases because the re-identification risk may still be high, and therefore suggests a methodology to carry out such a risk assessment in Appendix B "Concepts and Methods for De-identifying Clinical Trial Data".

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.

Following the completion of this first version of the PhUSE De-Identification Standards for SDTMIG 3.2 in May 2015, a number of pilots are conducted in order to assess further the usability and accuracy of the rules defined. Based on the feedback, a second version may be developed and made available to the community.

The Working Group is also currently discussing the next deliverables to work on to support the data transparency initiative in the industry. CDISC ADaM or the CDISC SDTM Therapeutic Areas standards are possibilities as an extension as well as addressing the issue of data de-identification documentation or providing open-source code to de-identify studies according to the PhUSE standards.

REFERENCES

[1] CSDR: Clinical Study Data Request", 2015. [Online]. Available: https://www.clinicalstudydatarequest.com/. [Accessed: 26-Nov-2015].
[2] Clinical Data Interchange Standards Consortium, "CDISC SDTM Implementation Guide (version 3.2)," 2015.
[3] PhUSE De-Identification Working Group, "De-Identification Standards for CDISC SDTM 3.2," 2015. http://www.phuse.eu/Data_Transparency_download.aspx
[4] Nisen, Perry, and Frank Rockhold. "Access to patient-level data from GlaxoSmithKline clinical trials." *New England Journal of Medicine* 369.5 (2013): 475-478.
[5] Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. (2014). Reanalyses of randomized clinical trial data. *Jama*,*312* (10), 1024-1032.
[6] YODA: the Yale School of Medicine's Open Data Access (YODA) Project. https://yoda.yale.edu/ (accessed 02/11/2015).
[7] Project Data Sphere: https://www.projectdatasphere.org/ (accessed 24/08/2015).
[8] TransCelerate Biopharma, "Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach," 2015.
[9] 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.
[10] US Congress, "The Health Insurance Portability and Accountability Act of 1996; 45 Code of Federal Regulations Part 164 - Security and Privacy." 1996.
[11] I. Hrynaszkiewicz, M. L. Norton, A. J. Vickers, and D. G. Altman, "Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers," *BMJ*, vol. 340, no. 181, pp. 304–307, Jan. 2010.
[12] Bradley Malin, "A De-identification Strategy Used for Sharing One Data Provider's Oncology Trials Data through the Project Data Sphere Repository," Project Data Sphere, Jun. 2013.
[13] Anonymisation: managing data protection risk code of practice, ICO, 2012.
[14] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

ACKNOWLEDGMENTS

We would like to thank all the working group participants for their input and contribution to the deliverable.

Vinitha Arumugam & Patricia Coyle (GSK)	Jean-Marc Ferran (Qualiance & PhUSE)	Nancy Freidland (IBM)
Per-Arne Stahl (AstraZeneca)	Nick De Donder and Lauren Shinaberry (Business & Decision Life Sciences)	Gene Lightfoot (SAS Institute)
Sherry Meeh (Johnson & Johnson)	Cathal Gallagher (d-Wise)	Jacques Lanoue & Benoit Vernay (Novartis)
Kim Musgrave (Amgen)	Nate Freimark (Theorem)	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)	Ravi Yandamuri (MMS Holdings)

And the reviewers who provided feedback and comments during the elaboration of the deliverable.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Jean-Marc Ferran
Enterprise: Qualiance
E-mail: JMF@qualiance.dk
Web: www.qualiance.dk

Name: Khaled El Emam
Enterprise: Privacy Analytics Inc.
E-mail: kelemam@privacy-analytics.com
Web: www.privacy-analytics.com

Name: Sarah Nolan
Enterprise: University of Liverpool
E-mail: sjn16@liverpool.ac.uk
Web: www.liverpool.ac.uk

Name: Boris Grimm
Enterprise: Boehringer Ingelheim
E-mail: boris.grimm@boehringer-ingelheim.com
Web: www.boehringer-ingelheim.com

Name: Nick De Donder
Enterprise: Business & Decision Life Sciences
E-mail: nick.dedonder@businessdecision.com
Web: http://www.businessdecision-lifesciences.com/