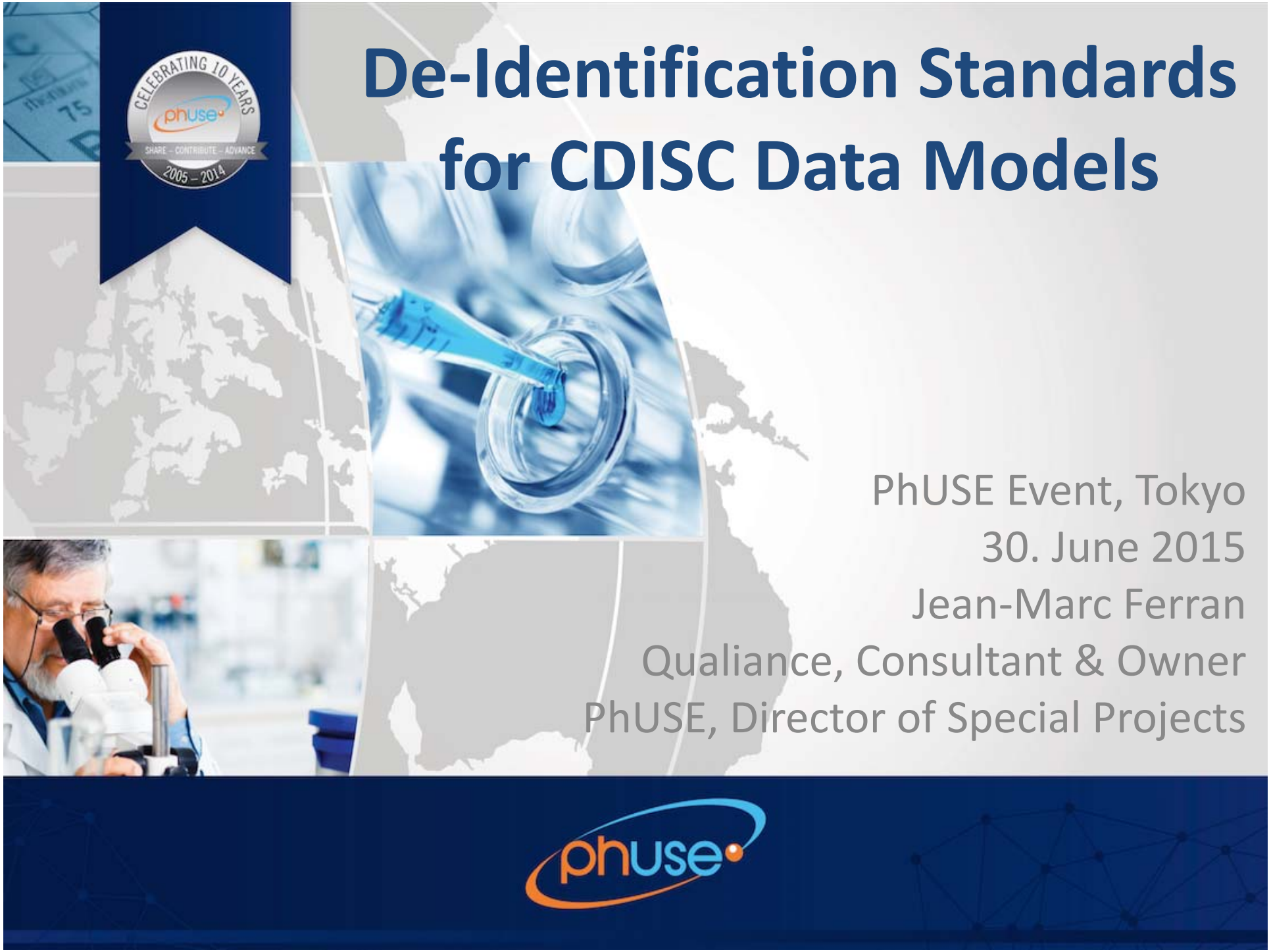


De-Identification Standards for CDISC Data Models

PhUSE Event, Tokyo

30. June 2015

Jean-Marc Ferran

Qualiance, Consultant & Owner

PhUSE, Director of Special Projects

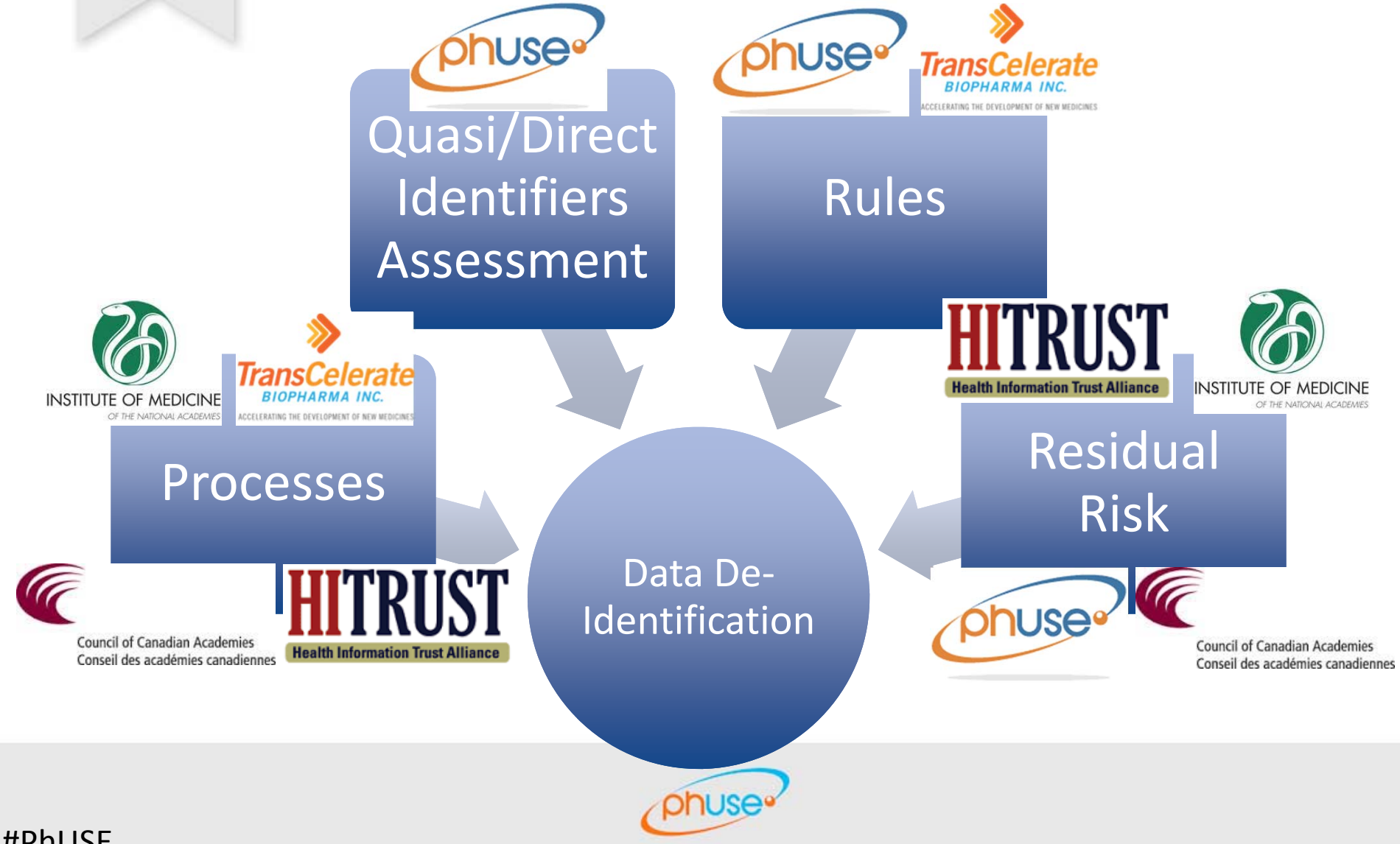
Agenda

- Vision and Goals of the Working Group
- Data De-identification Standards for SDTM 3.2
- Learnings & Potential Future Extensions

PhUSE Data Transparency Initiative Background

- There are current efforts by **regulators such as EMA** to examine how to **make Individual Patient Data (IPD) from clinical trials shared more widely**
- **Sponsors** have started **sharing IPD based on request proposals** from researchers and…
  - Data in **different data models** is available
  - **Each company seems to be defining their own high-level guidelines** for data de-identification
  - It is possible to **request data from different companies** within same research proposal

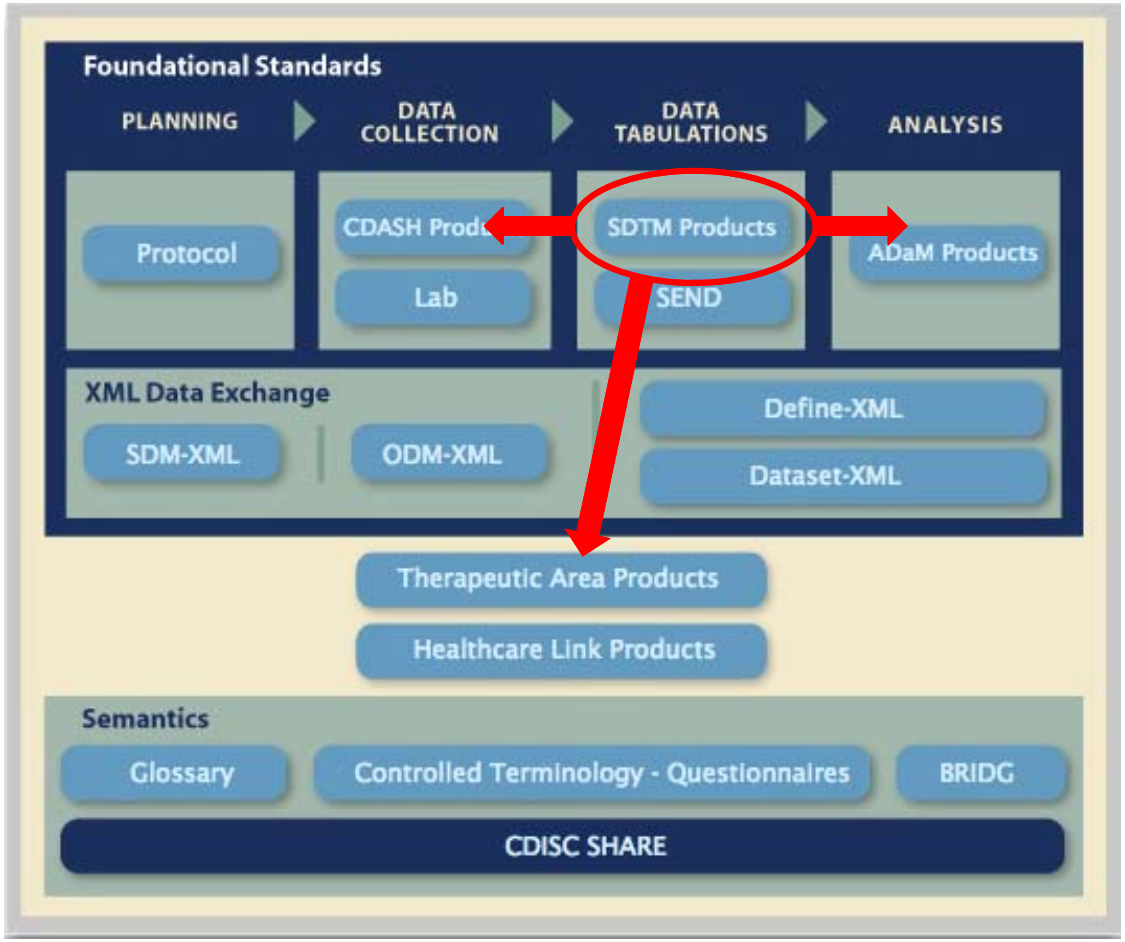
Data De-Identification Guidelines



PhUSE De-Identification Working Group Vision

"Develop data de-identification standards for CDISC data models"

Overview of CDISC standards



Source: CDISC website

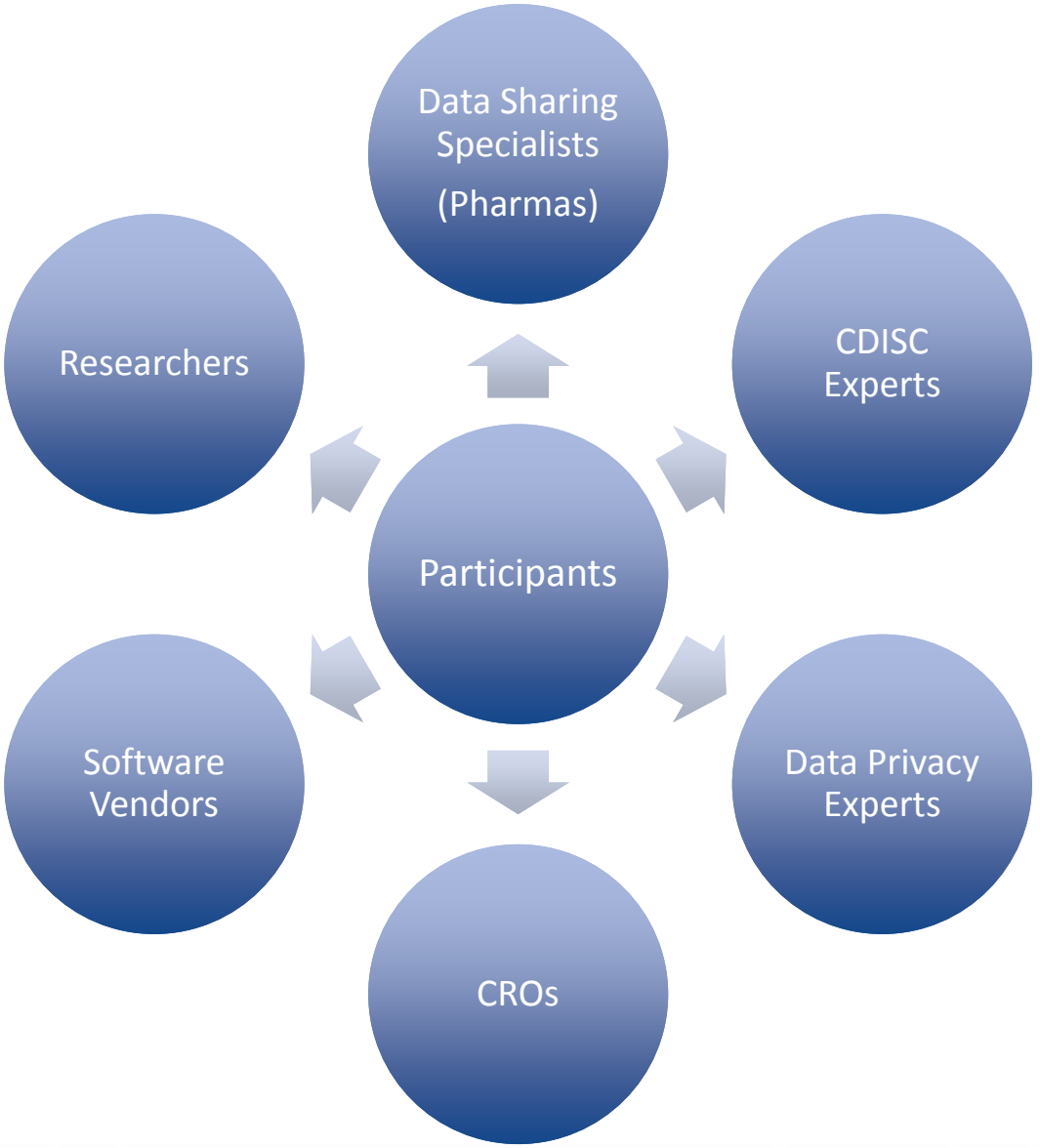
Goals

Provide peer-reviewed de-identification standards for CDISC data models to the industry

- Facilitate the **assessment of direct and quasi identifiers** in CDISC datasets
- **Ensure consistency** in de-identified data shared across sponsors
- *Provide **guidance on handling of low frequency** and residual risk assessment **in different data release contexts** – See Appendix 2*

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 & Cochran)
Boris Grimm (Boehringer Ingelheim)	Shafi Chowdury (Shafi Consultancy)	


Data Sharing Specialists (Pharmas)
Researchers
CDISC Experts
Participants
Software Vendors
Data Privacy Experts
CROs

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**

Key Areas and Rules

Dates

- Must be offset
- Date of Birth – Derive into Age at baseline and aggregate patients over 89 years old or derive into age folds (10-15, 15-20, 20-25 etc., 18-20, 20-22, 22-24, etc.)
- Date of Death - Offset

Low frequency & rare events

- Methodology such as one described in IOM report is recommended to be used
- Variables and datasets at stake have a comment associated with such considerations

Recoding of unique identifiers

- Subject IDs
- Investigator ID
- Site IDs
- Reference ID and Sponsor ID

Handling of free-text variables and extensible code lists

- If critical to the analysis, and not recoded in the dataset. Review and only redact values with personal information. Otherwise remove.
- Extensible code lists variables are flagged as a warning as free-text may be added

Geographical location

- Aggregation of country to continent unless country is critical to analysis.
- Site and Investigator names and IDs. must be deleted. Site/Investigator ID may be recoded in some cases.

Sensitive data

- This is the responsibility of the sponsors to define how to handle such data
- Variables and datasets at stake have a comment associated with such considerations

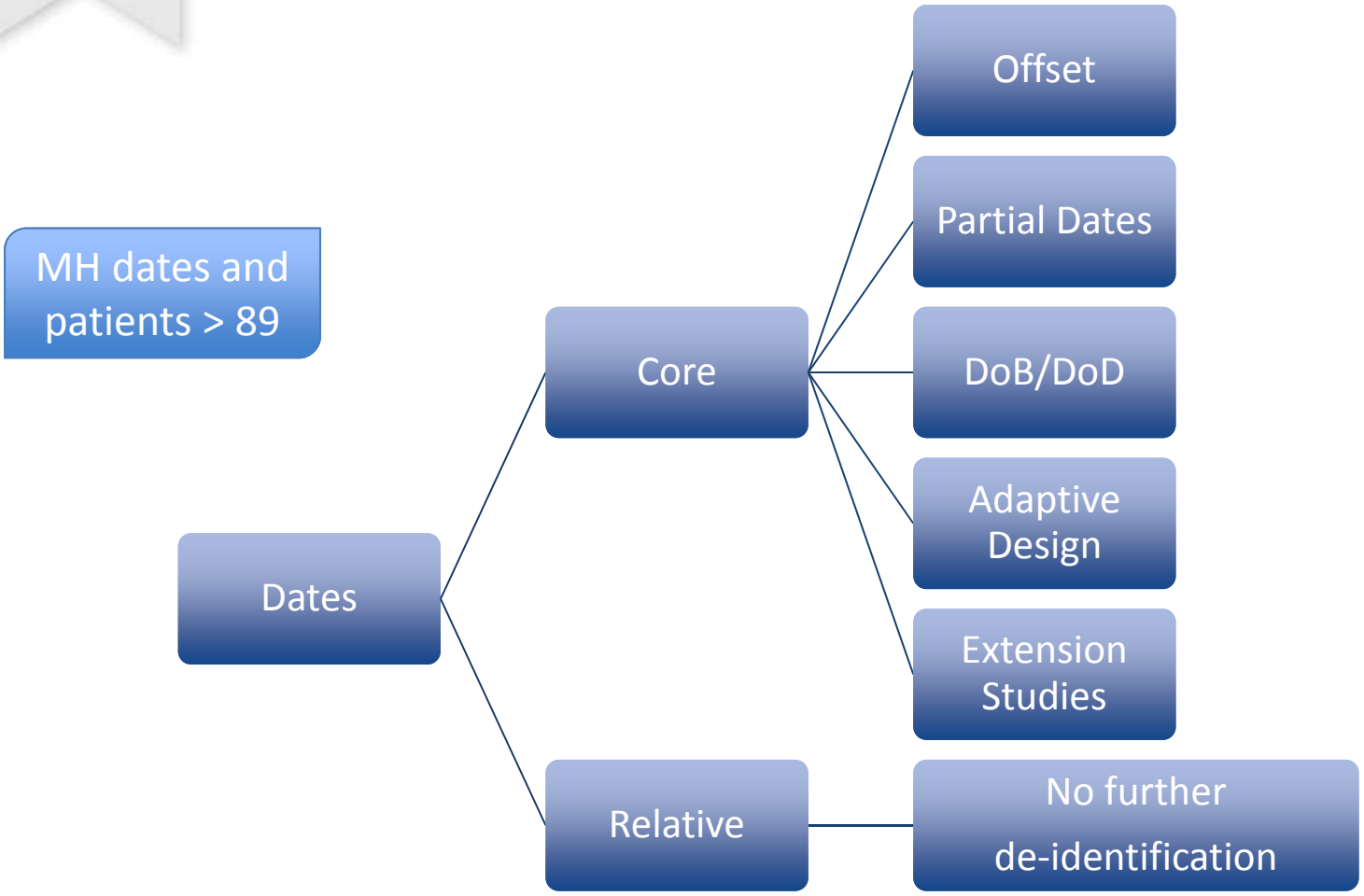
Some quasi identifiers are advised to be kept as-is

- Important variable for analysis. E.g. Gender
- De-identification is already in place. E.g. relative dates such as study day

PII of third-party

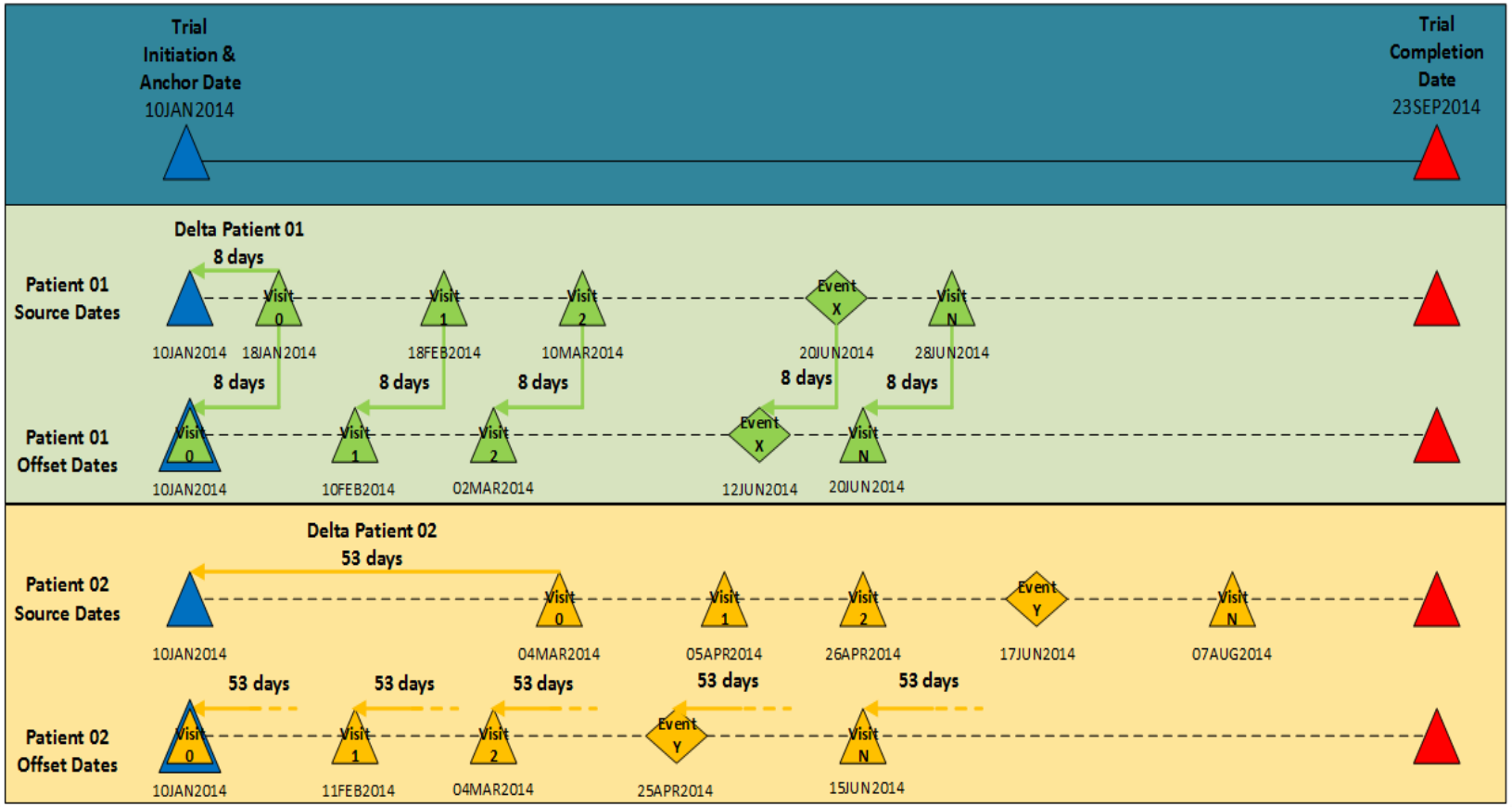
- Must be removed as they can provide geographical information
- Information such a evaluator type is however advised to be kept

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

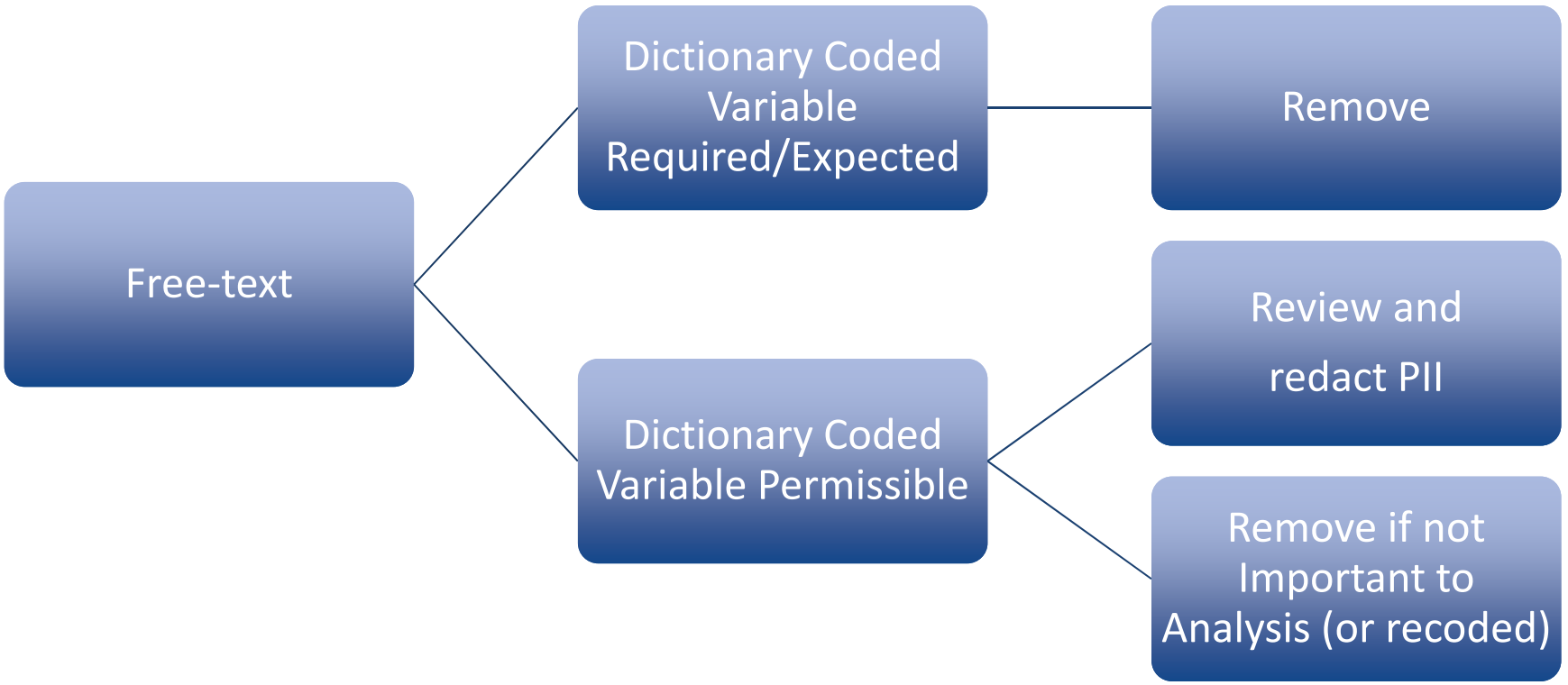


Issue with Partial Dates

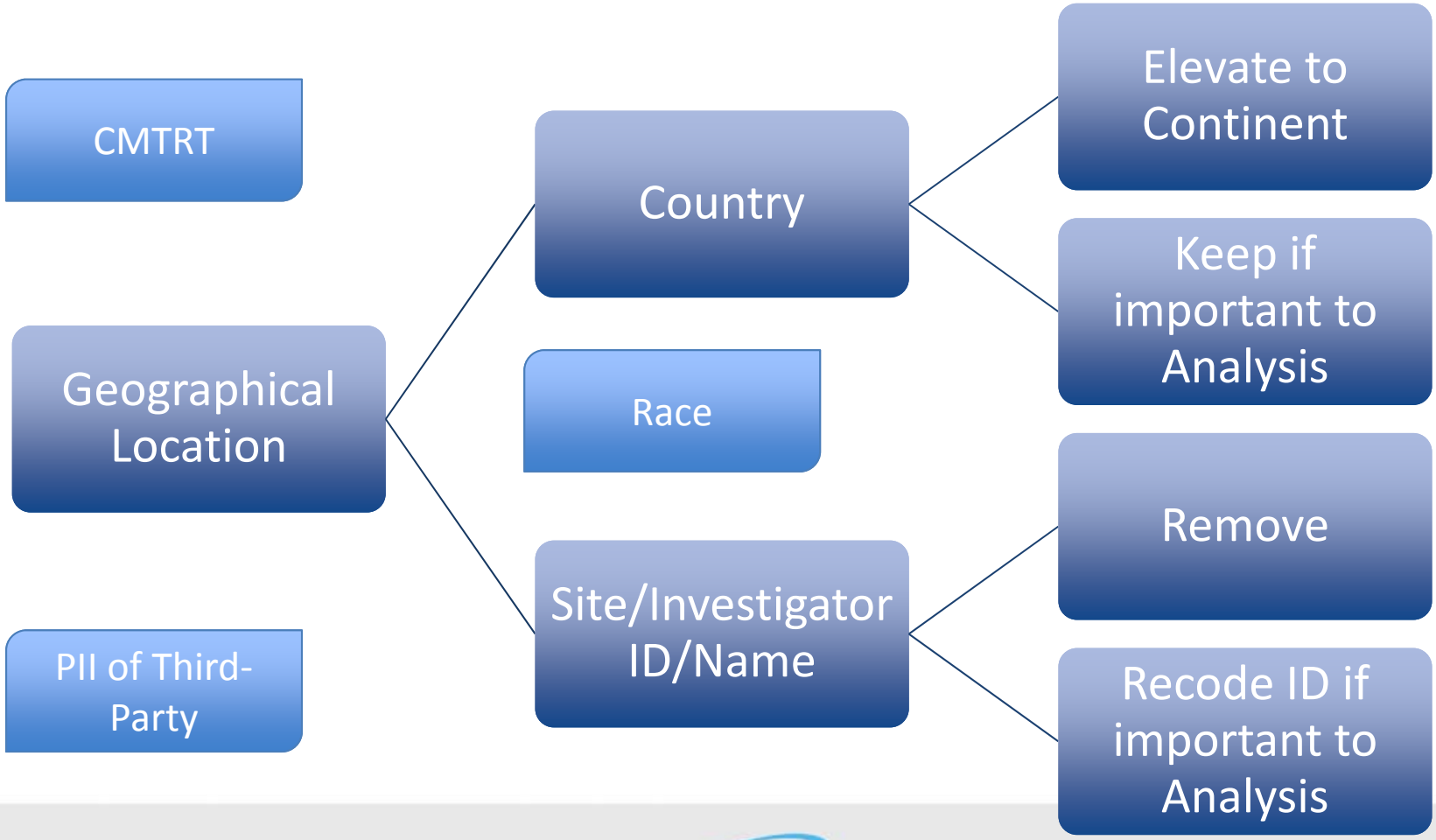
Ex: Delta applied of -14 days

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

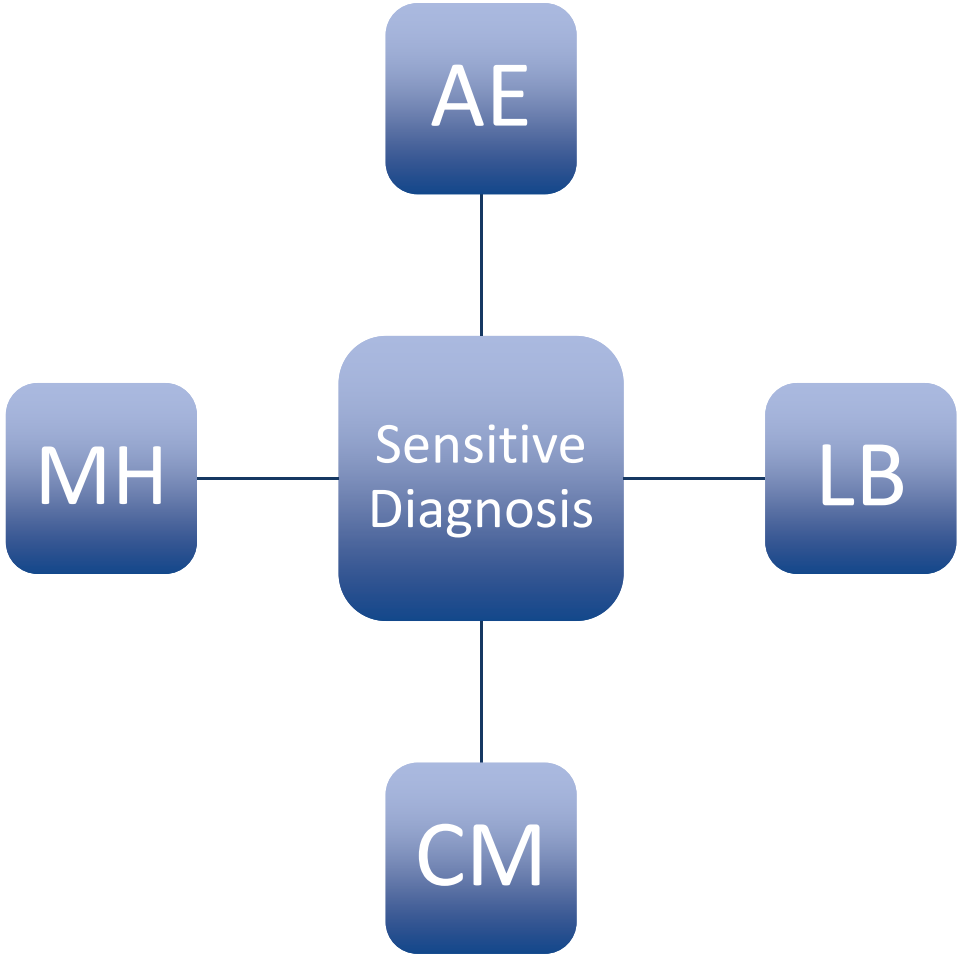
Free-text


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

Geographical Location



Sensitive Diagnosis



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

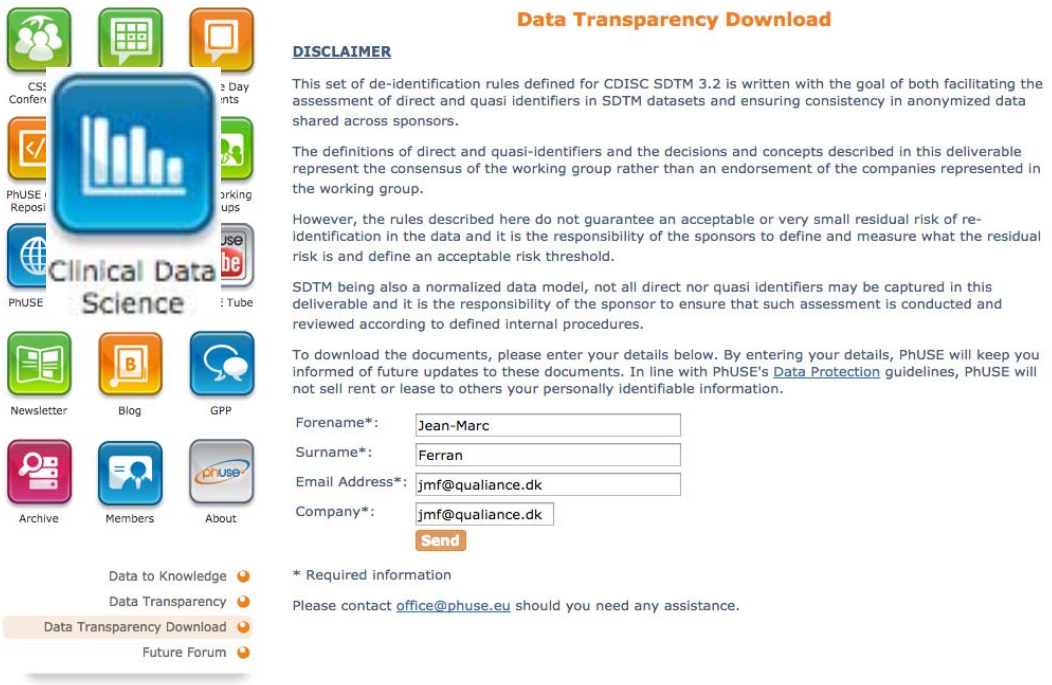
#PhUSE

200+ comments from External Review

Pharmaceuticals	CROs/Consultancies	Professional Organizations and Non-for-Profit	Academia
GSK	Quintiles	Project Data Sphere	Liverpool University
AstraZeneca	Privacy Analytics	European Medical Writer Association (EMWA)	Queen Mary University of London
Sanofi	MMS Holdings	TransCelerate	Clinical Trials Unit at UCL
Shire	d-Wise		University of Oxford
UCB			
Pfizer			
Roche			
Boehringer Ingelheim			
Novartis			
EISAI			
Abbvie			
Merck Serono			
J&J			

Available for Download

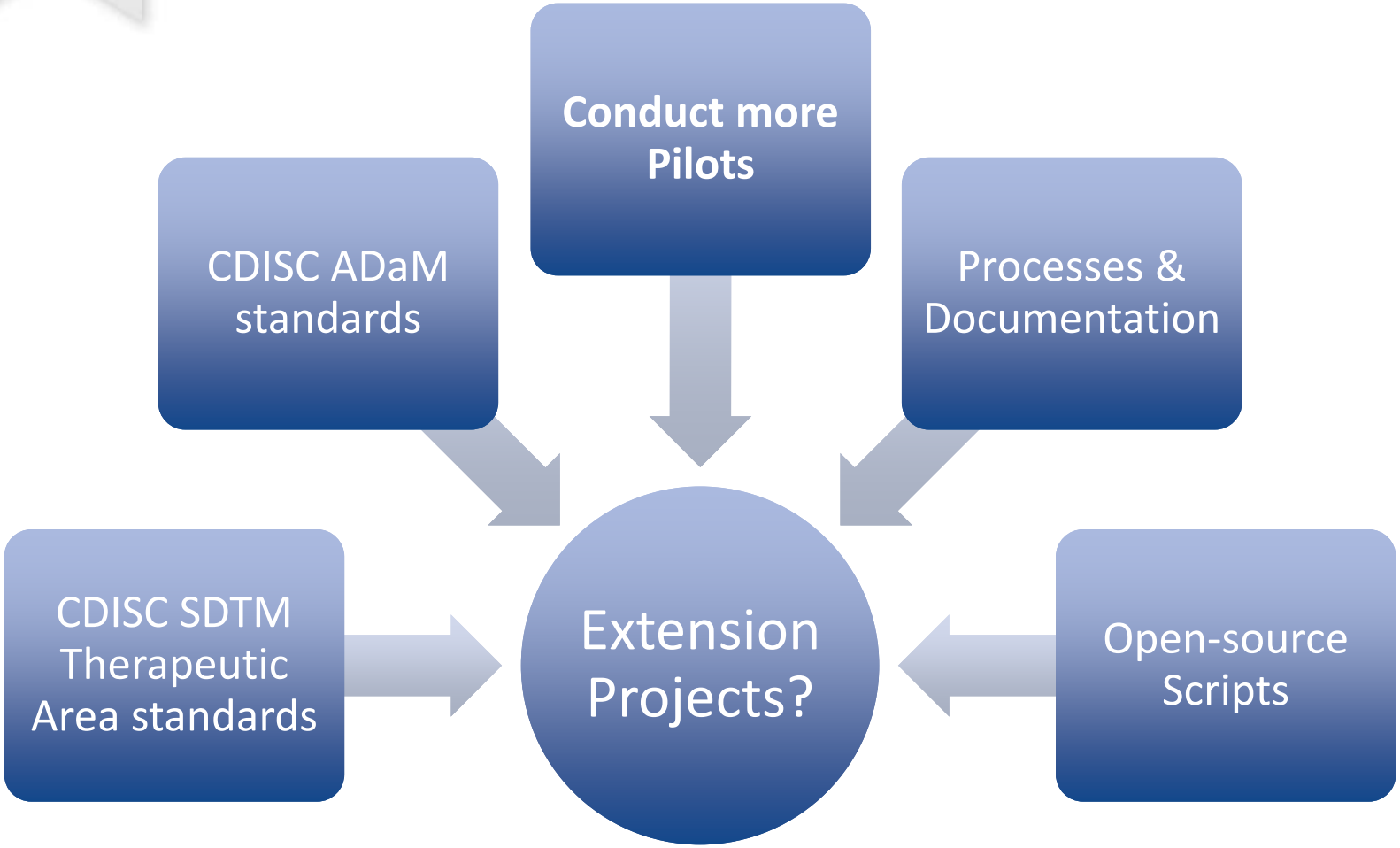
Published on 15. May 2015



Learnings

- **SDTM highly normalized and implemented differently**
  - Difficult to consider all cases
  - Notes and warnings are issued
  - As much as possible, we could keep CDISC compliance
- **Consensus not always simple to reach**
  - The variety of participants' background helped making informed decisions
  - One rule per scenario is defined – No flavor!

Next steps?


Conduct more Pilots
CDISC ADaM standards
Processes & Documentation
CDISC SDTM Therapeutic Area standards
Extension Projects?
Open-source Scripts

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

Appendix 2 – Low Frequencies

- The PhUSE approach blends both of these perspectives *(safe harbor and expert determination)* 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. 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.