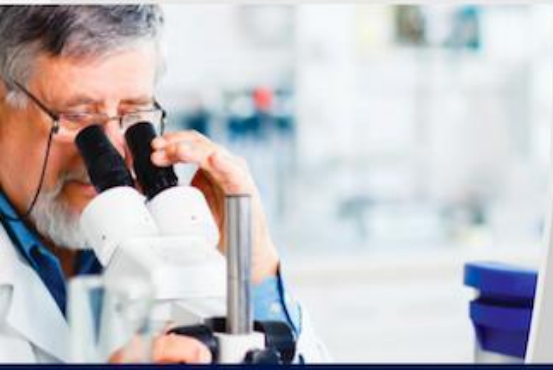
PhUSE Data De-Identification Working Group: Providing De-Identification Standards for CDISC data models



Frankfurt SDE
15. April 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

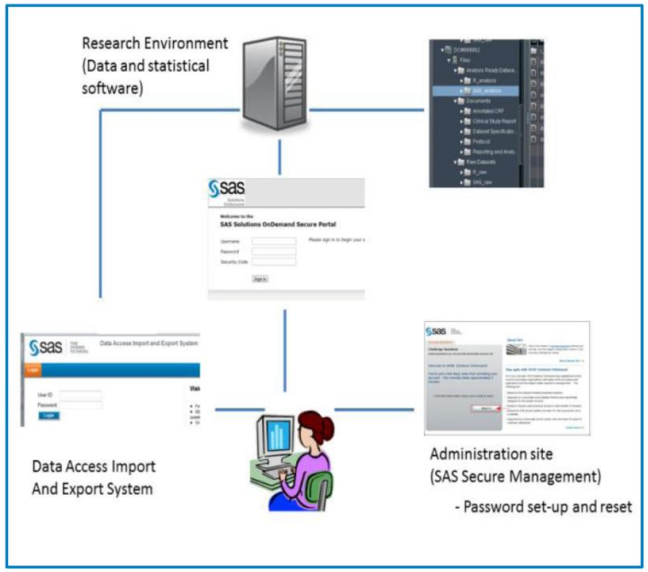
- There are current efforts by **regulators such as EMA** to examine how to **make IPD from clinical trials shared more widely**
- **Sponsors** have started **sharing data based on request proposals** from researchers and…
  - Data in **different data models** is available
  - **Each company seems to 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 Transparency Concept

Researcher writes research requests

Independent Review Board (IRB) evaluates requests

Secure Statistical Computing Environment



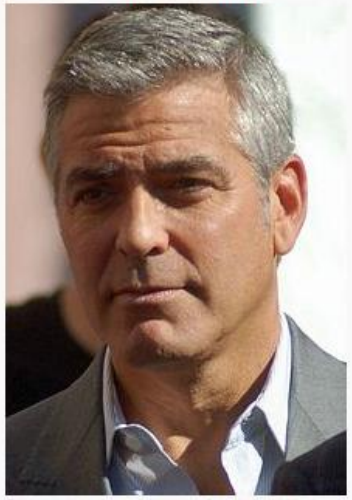
Data de-identification

QUALIANCE

Who is Georges?

Patient ID	DoB	Age	G		Country	Partner Age
1	12APR1963	51	M		nada	48
2	28MAY1974	40	M		ance	41
3	06MAY1961	53	M		ited States	36
4	28MAY1954	60	Fe		ain	65
5	14JUL1969	45	M		azil	41
6	13AUG1964	50	Fe		gentina	45
7	18MAR1961	53	M		ited States	48
8	22JAN1961	53	M		ited States	37
9	27SEP1924	90	M		nada	73
10	07FEB1956	58	M		nada	62

George Clooney



Clooney at a ceremony for John Wells to receive a star on the Hollywood Walk of Fame in January 2012

Born	George Timothy Clooney May 6, 1961 (age 53) Lexington, Kentucky, U.S.
Occupation	Actor, filmmaker
Years active	1978–present
Spouse(s)	Talia Balsam (m. 1989–93) Amal Alamuddin (m. 2014)
Parents	Nick Clooney Nina Warren
Relatives	Rosemary Clooney (aunt) Jose Ferrer (uncle) Miguel Ferrer (cousin) Rafael Ferrer (cousin)

QUALIANCE

Who is Georges?

Patient ID	Age Category	Age	Gender	Race	Country	Partner Age
1	<89	51	Male	White	Canada	
2	<89	40	Male	Asian	France	
3	<89	53	Male	White	United States	
4	<89	60	Female	Black	Spain	
5	<89	45	Male	Black	Brazil	
6	<89	50	Female	White	Argentina	
7	<89	53	Male	White	United States	
8	<89	53	Male	White	United States	
9	≥89	.	Male	White	Canada	
10	<89	58	Male	White	Canada	

QUALIANCE

Who is Georges?

Patient ID	Age Category 2	Age	Gender	Race	Continent	Partner Age
1	50-59		Male	White	North America	
2	40-49		Male	Asian	Europe	
3	50-59		Male	White	North America	
4	60-69		Female	Black	Europe	
5	40-49		Male	Black	South America	
6	50-59		Female	White	South America	
7	50-59		Male	White	North America	
8	50-59		Male	White	North America	
9	≥89		Male	White	North America	
10	50-59		Male	White	North America	

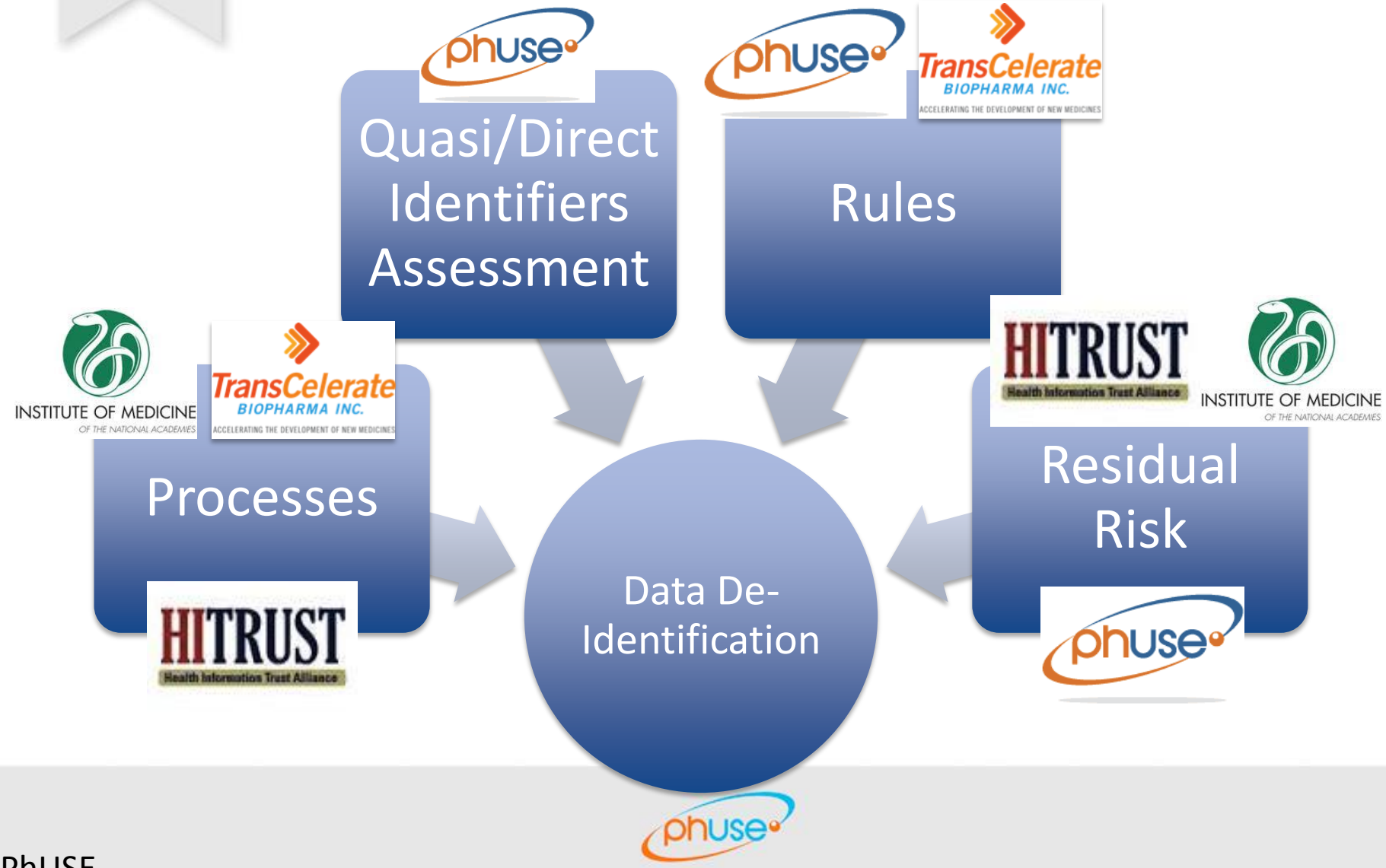
QUALIANCE

Who is Georges?

Patient ID	DoB	Age	Gender	Race	Country	Partner Age
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

QUALIANCE

Data De-Identification Guidelines



PhUSE De-Identification Working Group Vision

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

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

PhUSE Data De-Identification Working Group Participants

Vinitha Arumugam & Patricia Coyle (GSK)	Jean-Marc Ferran (Qualiance & PhUSE)	Nancy Freidland (IBM)
Per-Arne Stahl (AstraZeneca)	Lauren Shinaberry (BDLS & CDISC)	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)	Sarvesh Singh (ICON)	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)	
Adel Salem (Novo Nordisk)	Khaled El Emam (Privacy Analytics)	


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 deliverable is in **final draft state, being finalized** and content may change
- The rules described **do not guarantee an acceptable or very small residual risk** of re-identification

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
- **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 reader 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

- Critical to the analysis, and not recoded in the dataset. Review and only redact values with personal information
- 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

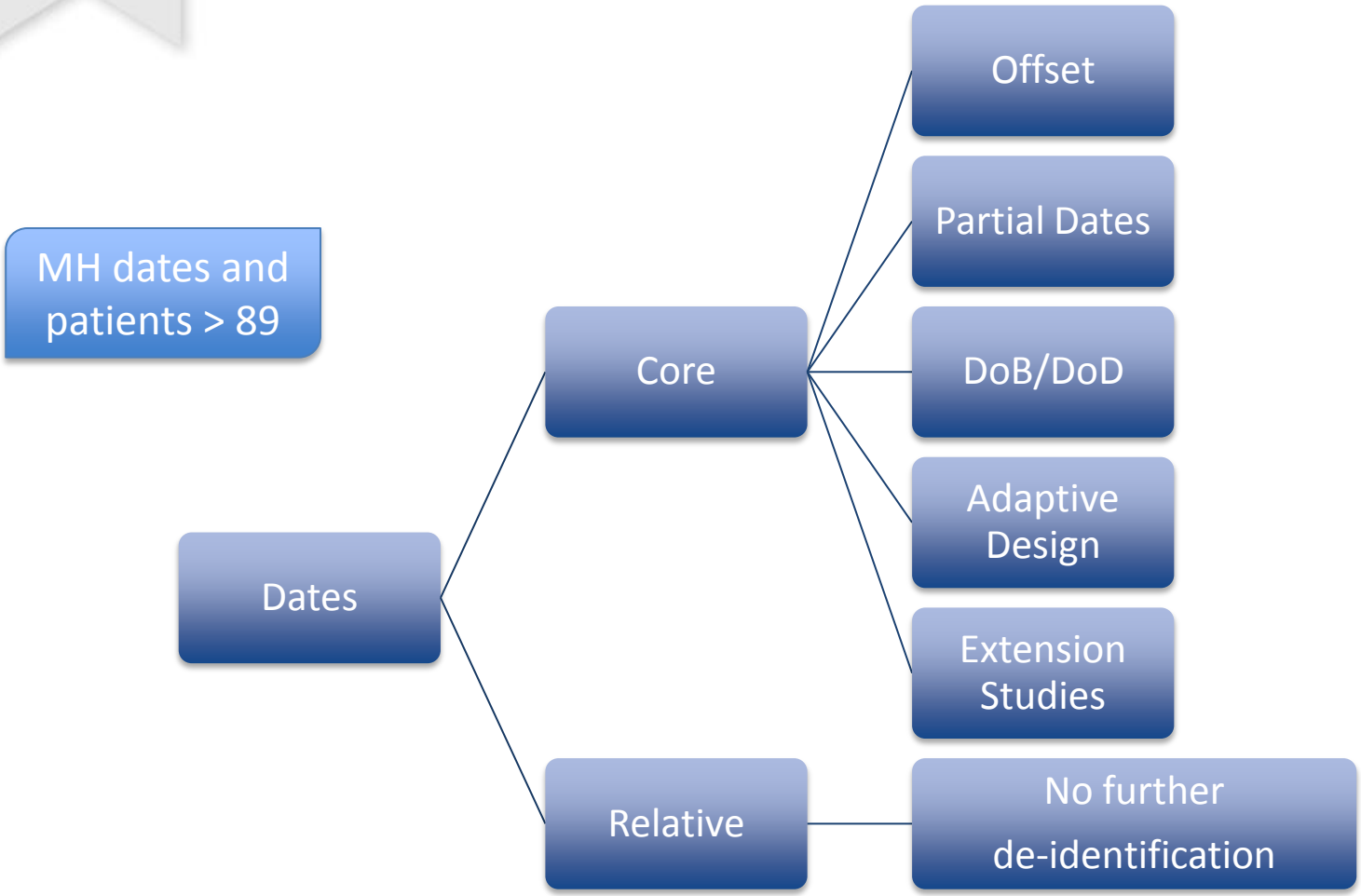
Some quasi identifiers are advised to be kept as-is

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

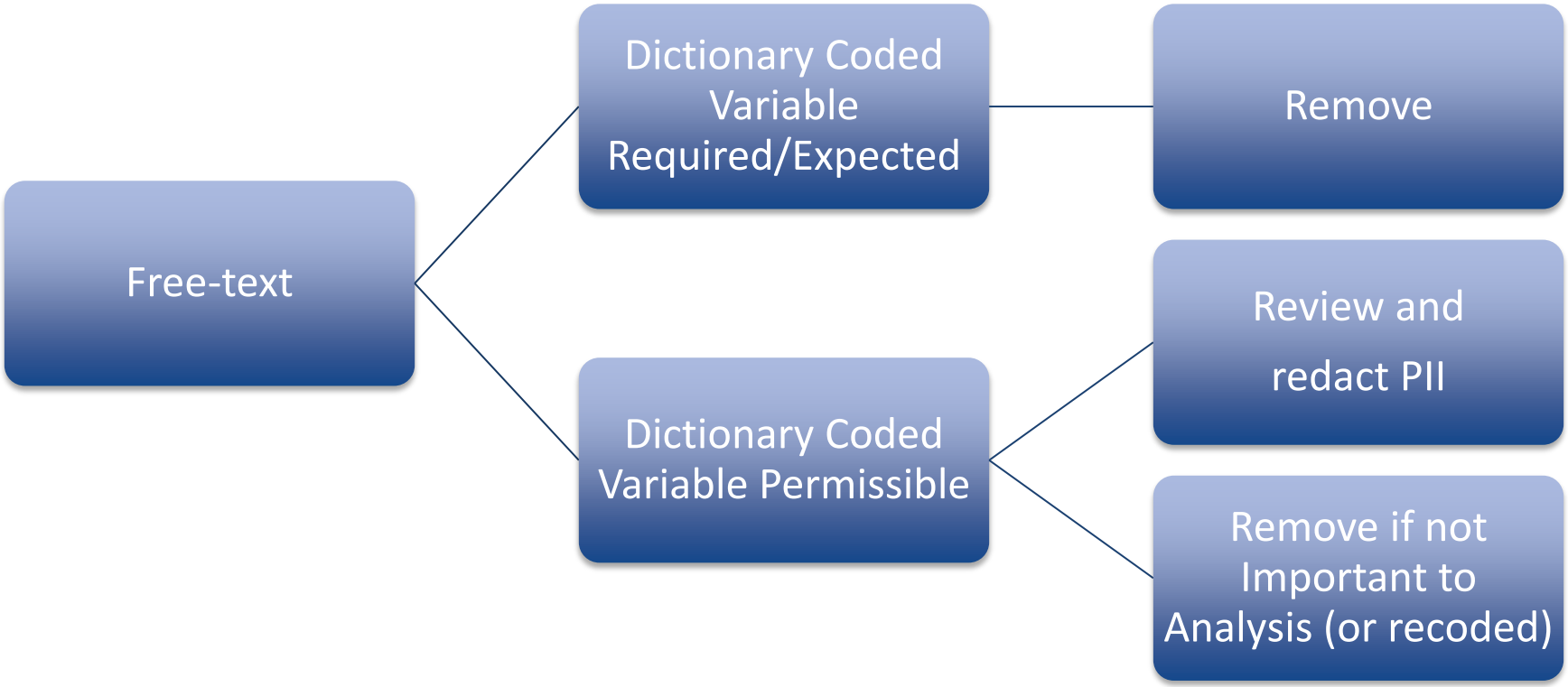
PII of third-party

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

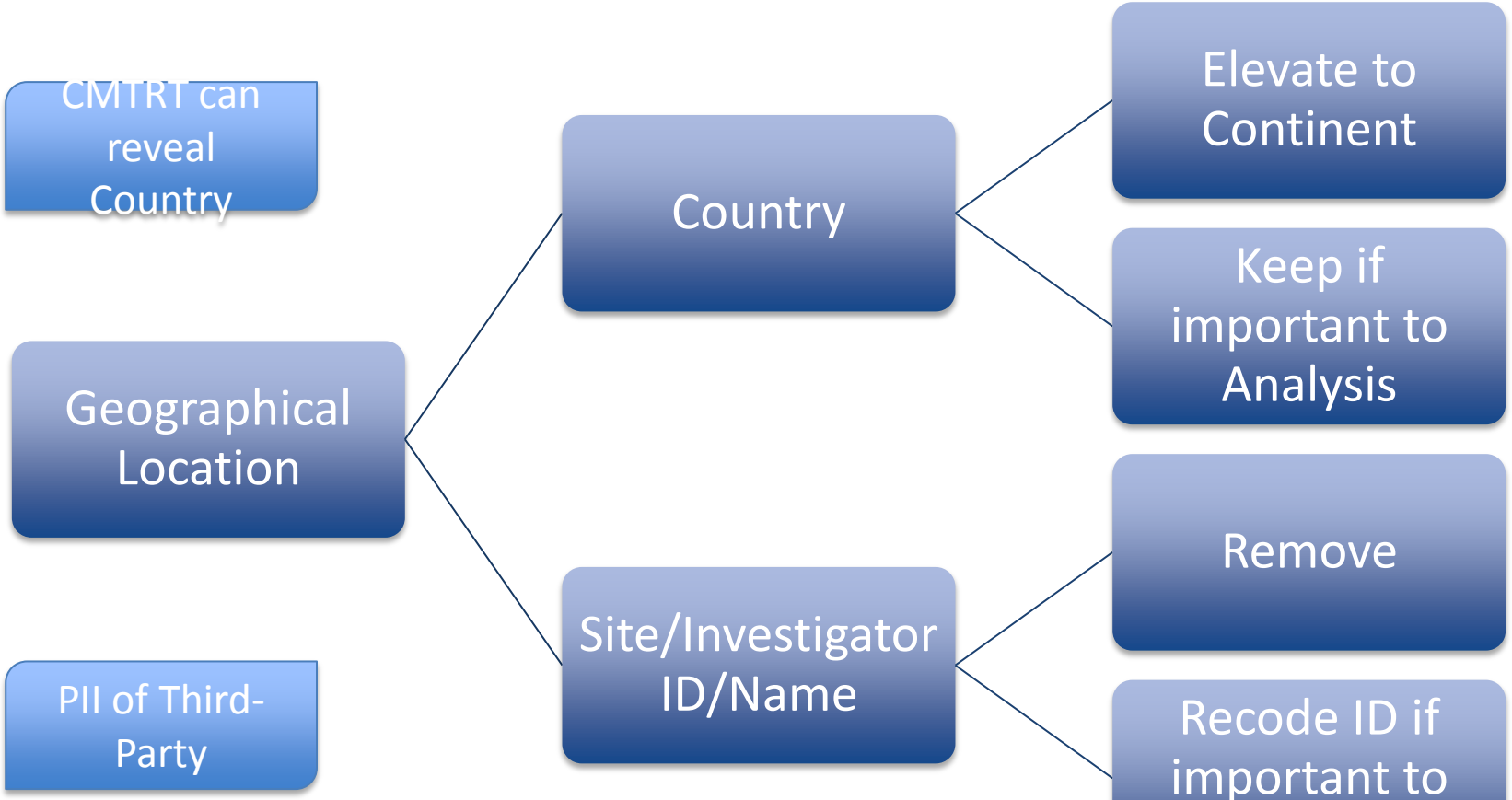
Dates


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

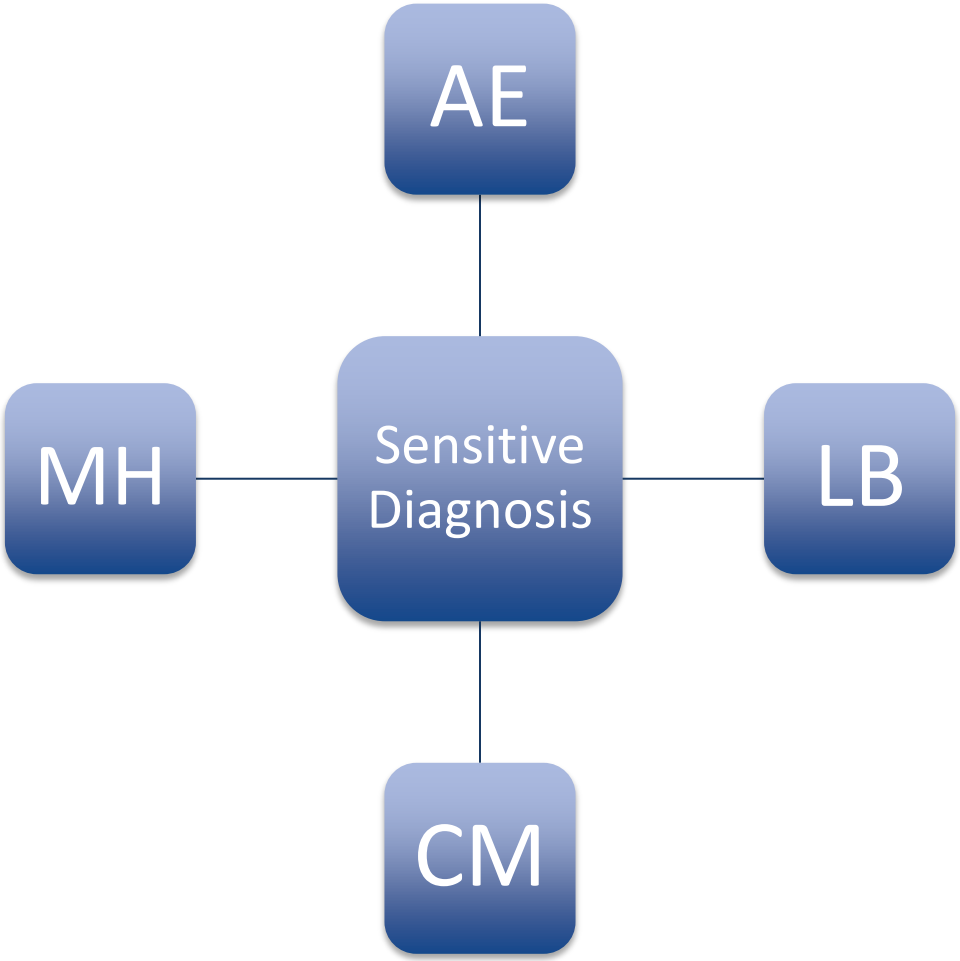
Free-text


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



"The Deliverable"

De-Identification Standards for CDISC SDTM 3.2 (In Review)

	Rule Description										Impact on Data Utility			
1	Seq_For_Order	Observation_Class	Domain_Prefix	Variable_Name	Variable_Label	Type	Length	Controlled_Terms_or_Format	Role	CDISC_Notes	Core	Direct_Quasi_Identifier (Direct/Quas	DI_Primary_Ru	DI_Alternative_Rule	DI_Comment
170	968	All classes		--ENRF	End Relative to Reference Period	Char	20		Timing	Identifies the end of the observation as being before, during or after the sponsor-defined reference period. The sponsor-defined reference period is a continuous period of time defined by a discrete starting point and a discrete ending point represented by RFSTDTC and RFENDTC in Demographics. Note: This variable will be deprecated (phased out) in a future (post SDTM v1.4) release. The functionality of this variable can be replaced by the use of --ENRTPT with --ENTPT – RFENDTC.					
171	969	All classes		--EVLINT	Evaluation Interval	Char	20		Timing	Duration of interval associated with an observation such as a finding -- TESTCD, represented in ISO 8601 character format. Example: -P2M to represent a period of the past 2 months as the evaluation interval for a question from a questionnaire such as SF-36.		Quasi Level 2	Keep		
172	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	Keep		Low frequencies in a very special (not standardised) interval could lead to a higher probability of identification
173	1	Special-Purpose	DM	STUDYID	Study Identifier	Char	20		Identifier	Unique identifier for a study.	Req				
174	2	Special-Purpose	DM	DOMAIN	Domain Abbreviation	Char	2	(DOMAIN)	Identifier	Two-character abbreviation for the domain.	Req				
175	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		
176	4	Special-	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		
177	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; will be 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		
178	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		
179	7	Special-	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		
180	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		
181	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		
182	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		
183	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	Direct	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.
184	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.
185	13	Special-	DM	SITEID	Study Site Identifier	Char	20		Record	Unique identifier for a site within a study.	Req	Quasi Level 1	Recode ID variable		Merge sites with less than 10 patients into one SITEID (e.g. 9999)
186	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	Recode ID variable		
	15	Special-	DM	INVNAM	Investigator Name	Char	20		Synonym Qualifier	Name of the investigator for a site.	Perm				Such information is related to other individuals than the patients and can also reveal geographic location of site. In addition, it holds little

Cover | Review Guidance | Intro | Definitions | Decisions | Rules | SDTMIG | References | Change Log

Cover | Review Guidance | Intro | Definitions | Decisions | Rules | SDTMIG | References | Change Log

Cover | Review Guidance | Intro | Definitions | Decisions | Rules | SDTMIG | References | Change Log

#PhUSE

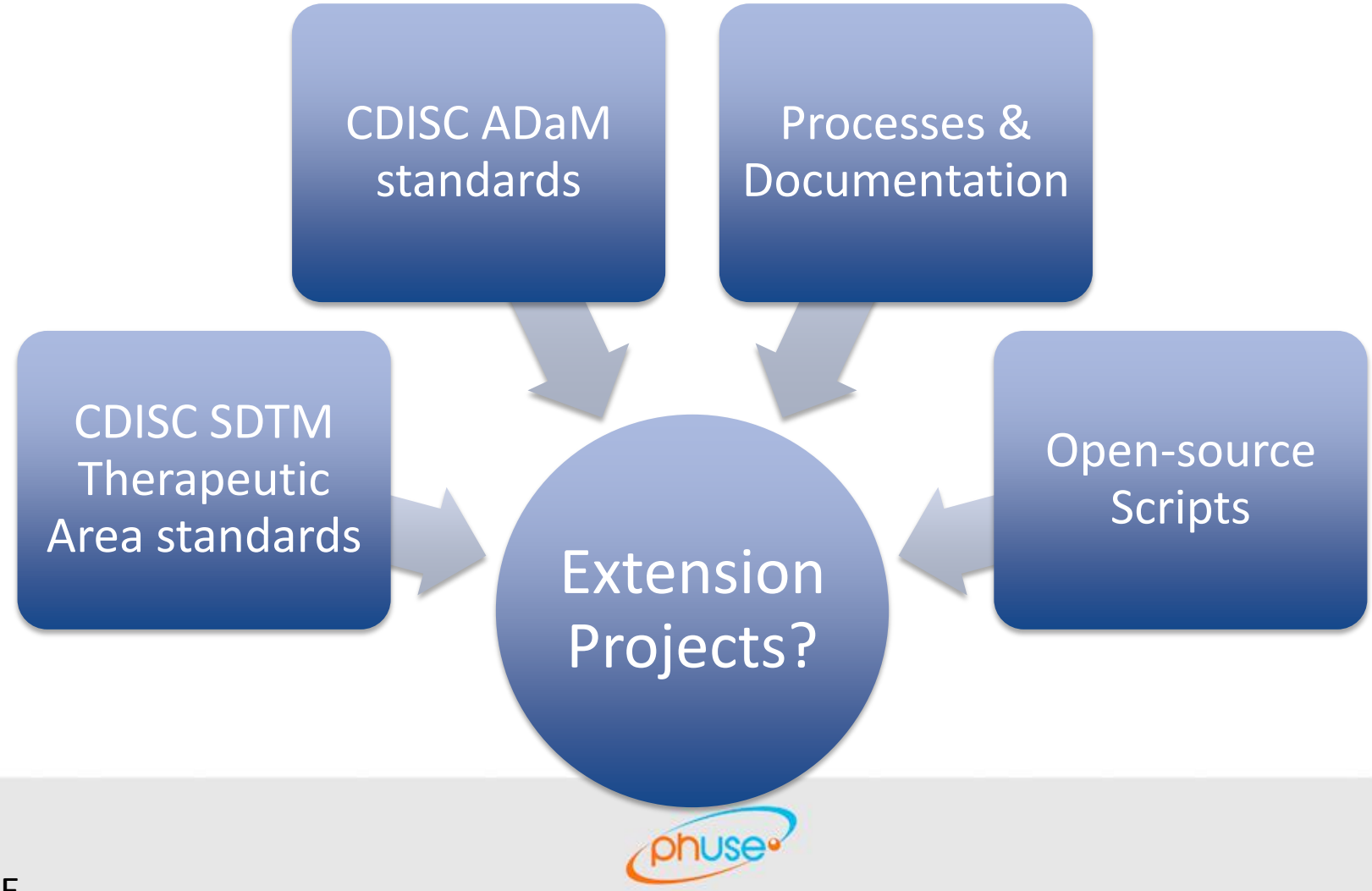
200+ comments from External Review

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

Learnings

- **SDTM highly normalized and implemented differently**
  - Difficult to consider all cases
  - Notes and warnings are issued
- **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?


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

Paris PhUSE Single Day Events

- Friday 29. May
- Sanofi offices
- Clinical Trial Data Sharing: Methods and Experiences with De-Identification
- Presenters from:
  - Industry
  - Data De-Identification Experts
  - Data Protection Agencies
  - Professional Organizations developing standards

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