



QUALIANCE

Anonymising Clinical Data – Key principles, Methods and Considerations

EFSPI/PSI Webinar: Anonymising Clinical Data

17. November 2017

Jean-Marc Ferran

Consultant & Owner

About the Speaker



Jean-Marc Ferran

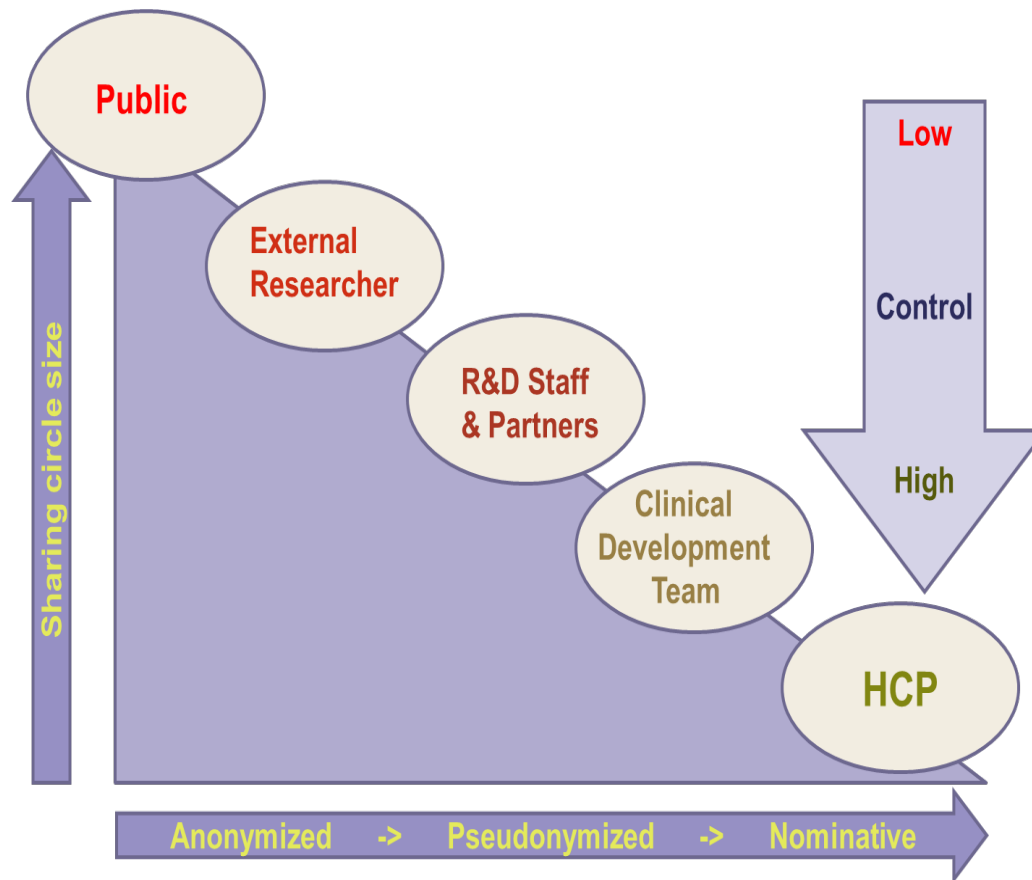
- Consultant & Owner – Qualiance
- Data Transparency Working Group Lead, PhUSE
- 15-year experience in the Life Sciences Industry
- Member of EMA Technical Anonymization Group and Health Canada Reference Group on Public Release of Clinical Information

Agenda

- Data Sharing and De-Identification Guidelines
- Residual Risk Assessment
- Context & Probability of Attack
- PhUSE De-Identification Standard



Data Sharing, Anonymization & Context

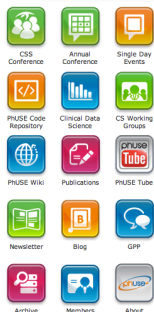


Source: PY Lastic, 37th Int. Privacy Conference, Amsterdam, 2015

Data De-Identification Guidelines Published in 2015

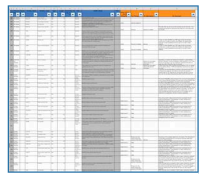


Pharmaceutical Users Software Exchange

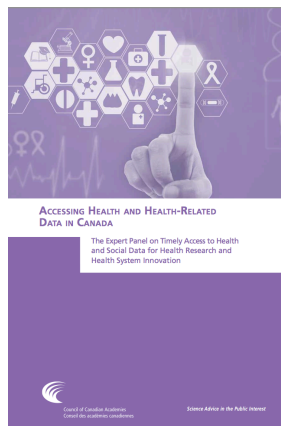


Data Transparency Material

Thank for your requesting access to the PHUSE De-identification Standard for SDTM 3.2, version 1.0.1. Click on the image below to download the document.



Guidance can be found in the "Introduction" tab. Please contact office@phuse.eu should you need any assistance.



Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk

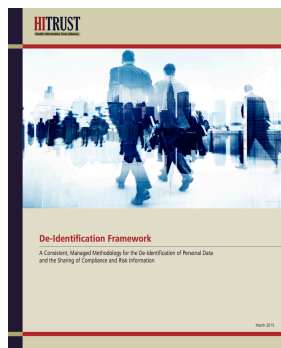
ISBN
978-0-309-31629-3

280 pages
6 x 9
PAPERBACK (2015)

Committee on Strategies for Responsible Sharing of Clinical Trial Data;
Board on Health Sciences Policy; Institute of Medicine



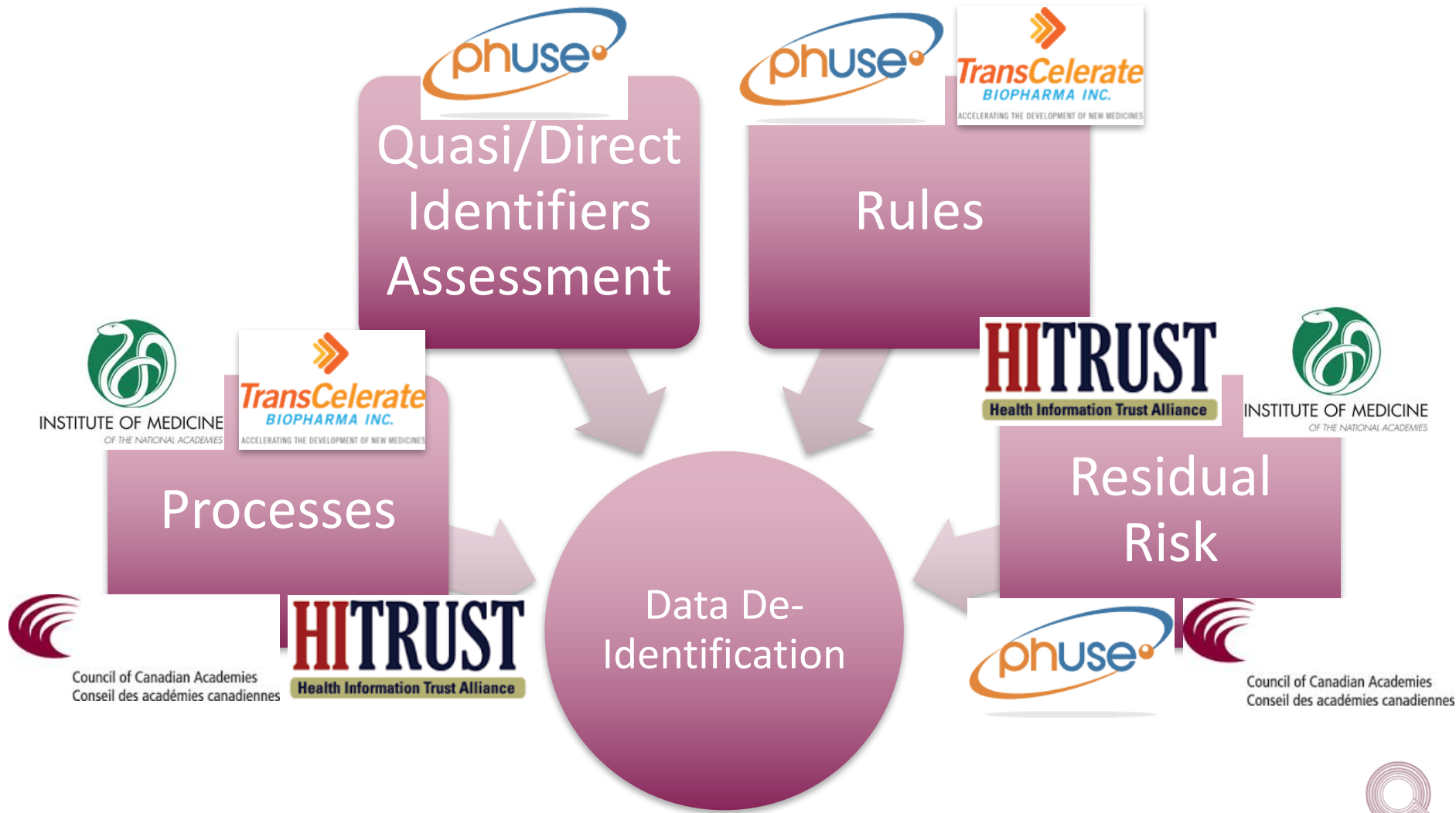
Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach



INTERNATIONAL
PHARMACEUTICAL
PRIVACY CONSORTIUM

IPPC White Paper on Anonymisation of Clinical Trial Data Sets

Data De-Identification Guidelines



Disclosure Process

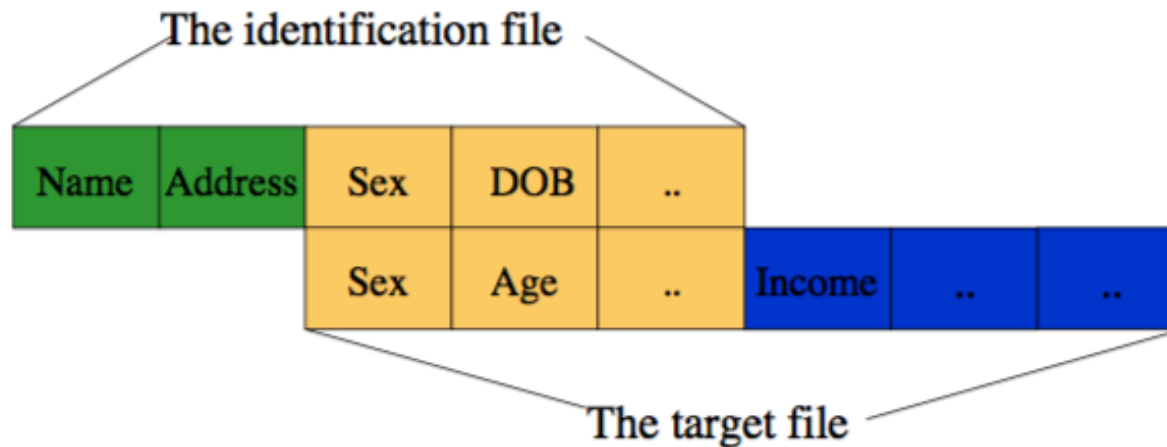


Figure 2.1: An illustration of the key variable matching process leading to disclosure. From Duncan et al (2011).

Source: The Anonymisation Decision-Making Framework, Elliott et al., 2016

EMA Policy 0070 Guidance

Adversary for Public Data Releases



1. Financial interest

An organisation sees a financial interest in finding out who are the trial participants in the clinical trial. Usually it would require some strategy to identify accurately a fair number of trial participants.

2. Demonstration attacks

A group or individual, possibly for academic reasons, in order to embarrass the data controller, or to undermine the public support for release of data, wishes to identify just one trial participant without regard to which trial participant it might be.

3. Event in which an acquaintance examine a report

A random event in which an individual happens to examine a report including data on a trial participant with whom they are well acquainted to the extent that they can accurately guess that certain information relates to that trial participant.

4. Participant of special interest

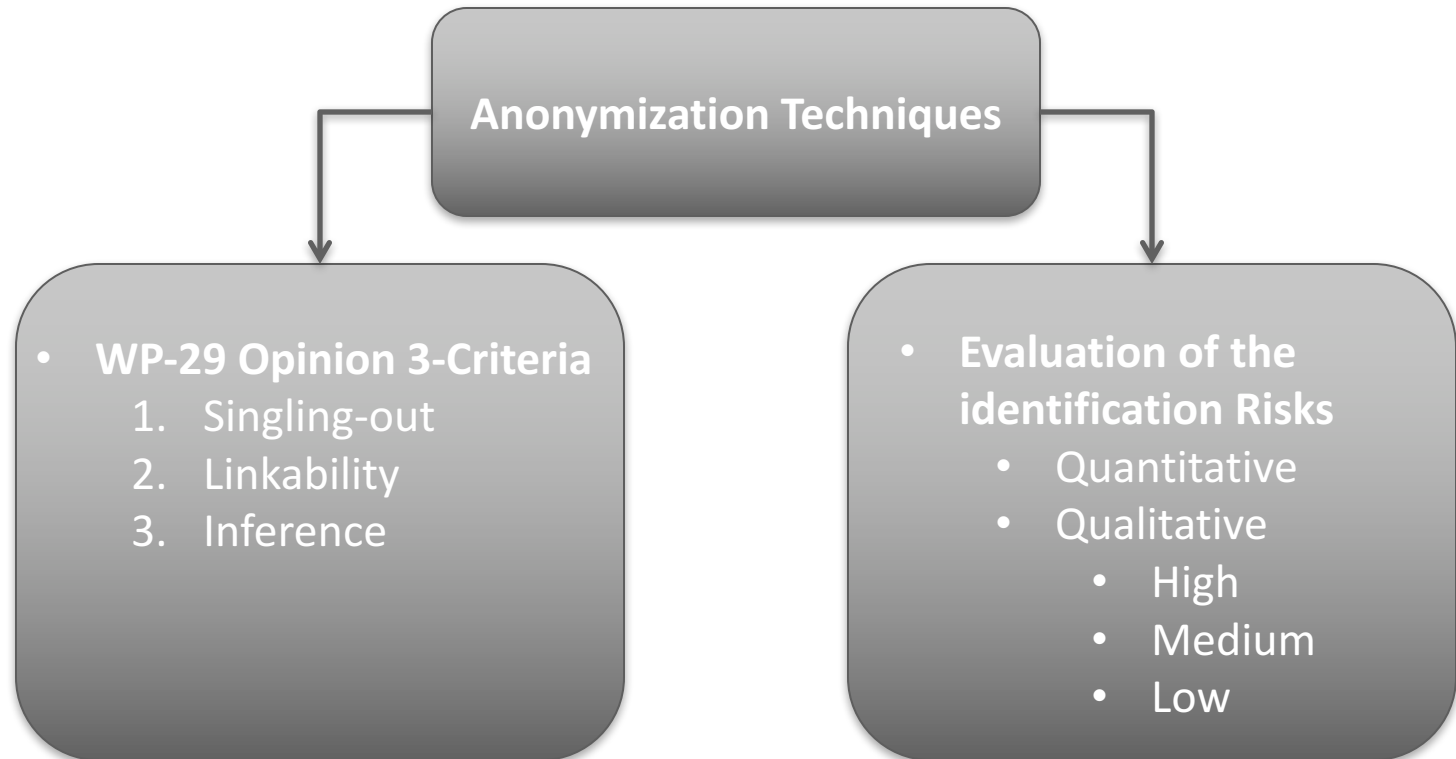
One trial participant is of particular public interest and is focused on by the press or other body

*“Applicants/MAHs should identify **possible adversaries and plausible attacks** on the data and evaluate the impact on the risk of re-identification.”*



EMA Policy 0070 Guidance

Anonymization Techniques




*“In order to achieve a maximum usefulness of the data published, **it is unlikely that for clinical reports all three criteria can be fulfilled by any anonymisation solution**, it is EMA’s view that a thorough evaluation of the risk of re-identification needs to be performed”*

Who is Georges?



Patient ID	DoB	Age	Gender
1	12APR1963	51	M
2	28MAY1974	40	M
3	06MAY1961	53	M
4	28MAY1954	60	F
5	14JUL1969	45	M
6	13AUG1964	50	F
7	18MAR1961	53	M
8	22JAN1961	53	M
9	27SEP1924	90	M
10	07FEB1956	58	M

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)

Country	Partner Age
Canada	48
France	41
United States	36
Spain	65
Brazil	41
Argentina	45
United States	48
United States	37
Canada	73
Canada	62

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	



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	



Who is Georges?



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



Equivalence Classes



Patients having same characteristics for important quasi identifiers

Size 1:
100.0%

Patient ID	DoB	Age	Gender	Race	Country	Partner Age
1	12APR1963	51	Male	White	Canada	48
2	28MAY1974	40	Male	Asian	France	41
3	06MAY1961	53	Male	White	United States	36
4	28MAY1954	60	Female	Black	Spain	65
5	14JUL1969	45	Male	Black	Brazil	41
6	13AUG1964	50	Female	White	Argentina	45
7	18MAR1961	53	Male	White	United States	48
8	22JAN1961	53	Male	White	United States	37
9	27SEP1924	90	Male	White	Canada	73
10	07FEB1956	58	Male	White	Canada	62



Equivalence Classes



Patients having same characteristics for important quasi identifiers

Size 3:
33.3%

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	








Equivalence Classes



Patients having same characteristics for important quasi identifiers

Size 5:
20.0%

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	

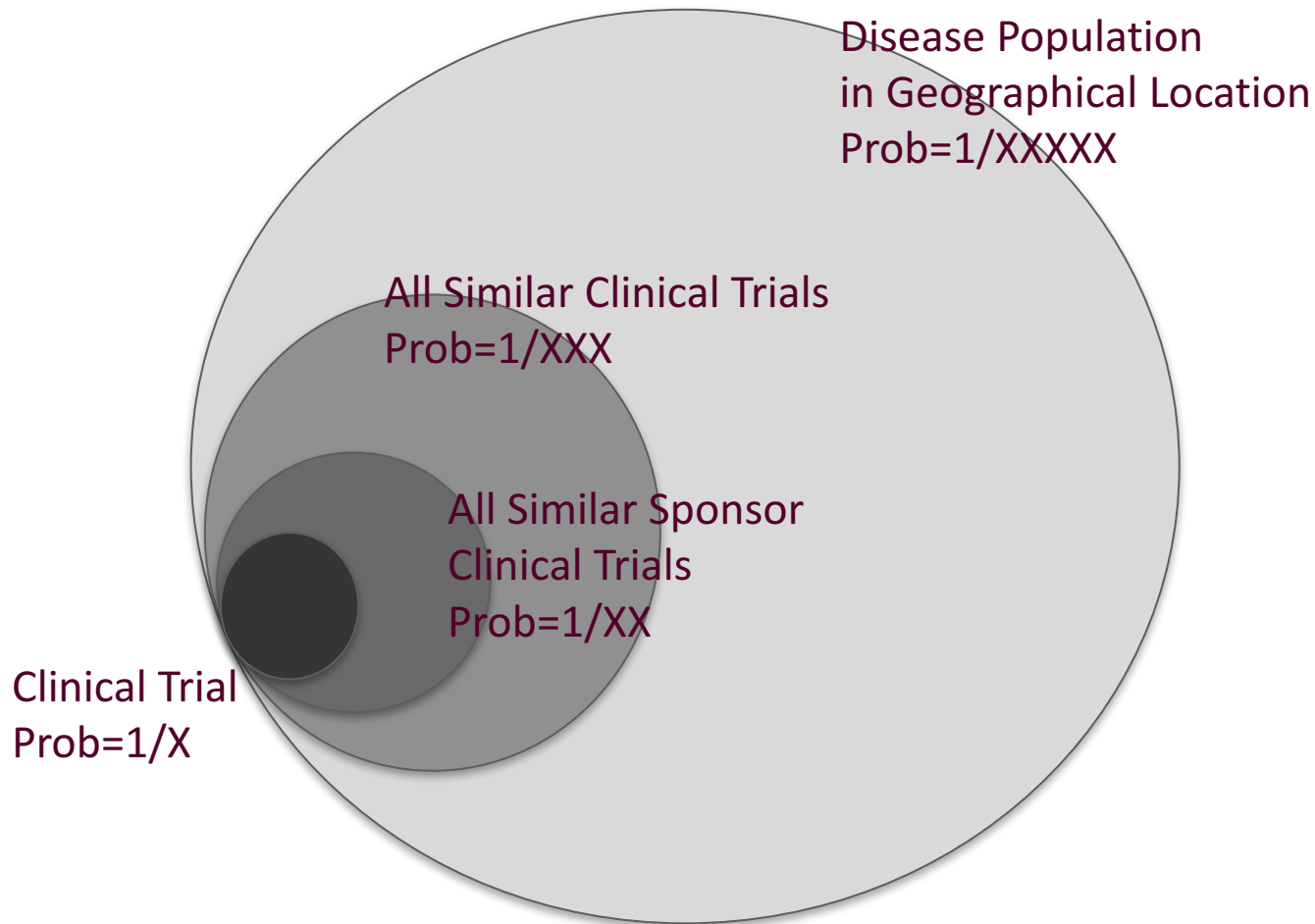
Simple Risk Measures

$$\text{Average}_{\text{Patients}} \left(\frac{1}{\text{Size}(\text{EquivalenceClass}[\text{Patient}])} \right)$$

$$\text{Max}_i \left(\frac{1}{\text{Size}(\text{EquivalenceClass}[i])} \right)$$



Residual Risk & Population



Probability of Successful Attack

For all i , $P(\text{ReID} \cap \text{Attack}_i) = P(\text{ReID} / \text{Attack}_i) \times P(\text{Attack}_i) \leq \text{Threshold}$

Attack	Example	Factors influencing $P(\text{Attack } i)$
1: Attempt	Researcher attempts at re-identifying patients	Mitigating Controls Motives & Capacity
2: Acquaintance	Researcher spontaneously recognizes patients	Study Patients Prevalence
3: Breach	A rogue organization “hacks” in the portal and retrieve the data	Security Practice at Data Recipient

- $P(\text{ReID} / \text{Attack } i)$ is controlled through data de-identification
- $P(\text{Attack } i)$ is dependent on disclosure context





PhUSE De-Identification Standards





Agenda

- **Vision and Goals of the Working Group**
- **Data De-identification Standards for SDTM 3.2**





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





PhUSE De-Identification Working Group Vision

“Develop data de-identification standards for CDISC data models”

20+
Participants
from Pharma,
CROs,
Software and
Academia

Focus first on
SDTM

Data Privacy
Rules &
Rational
Data Utility



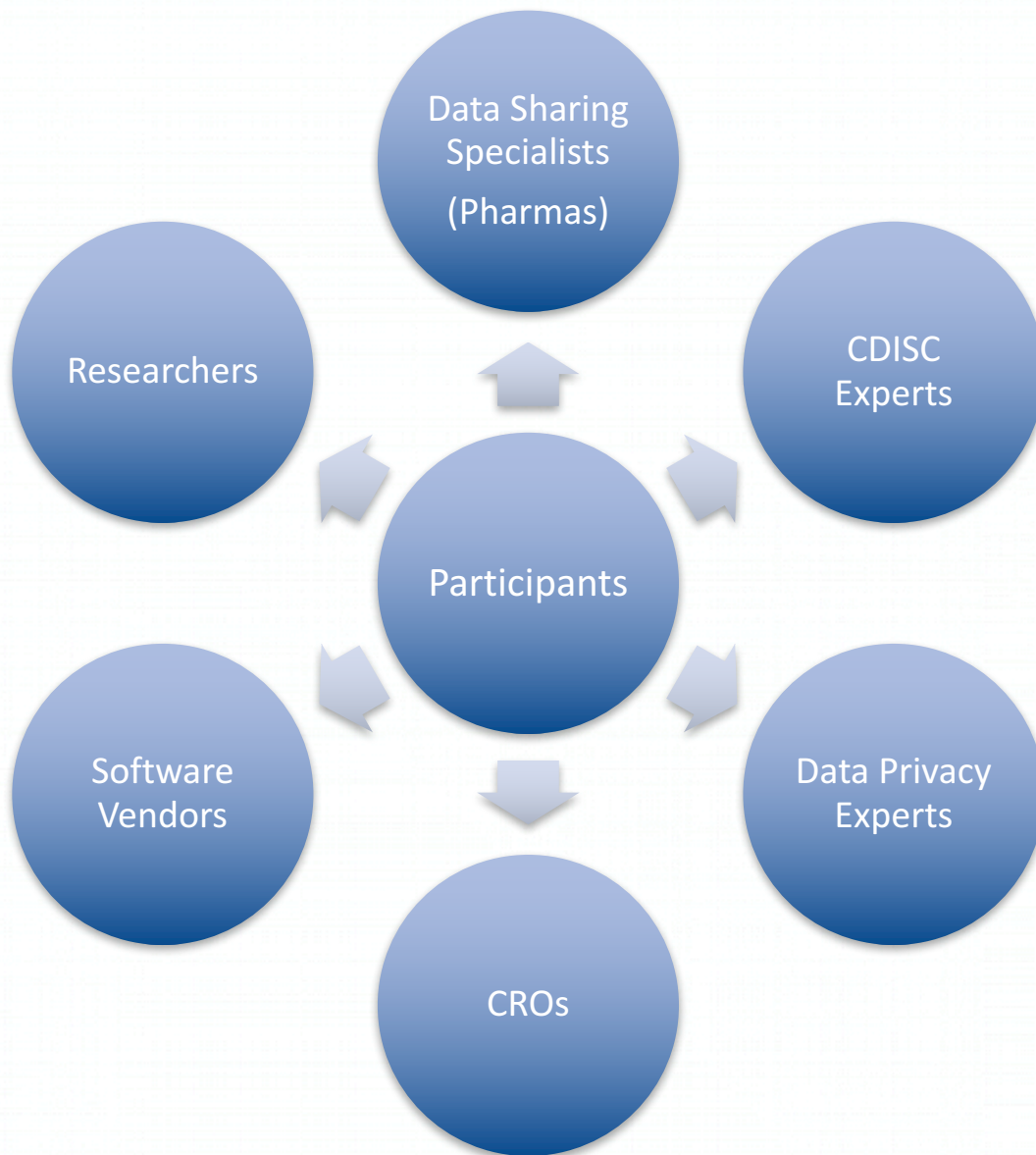


Goals

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*







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

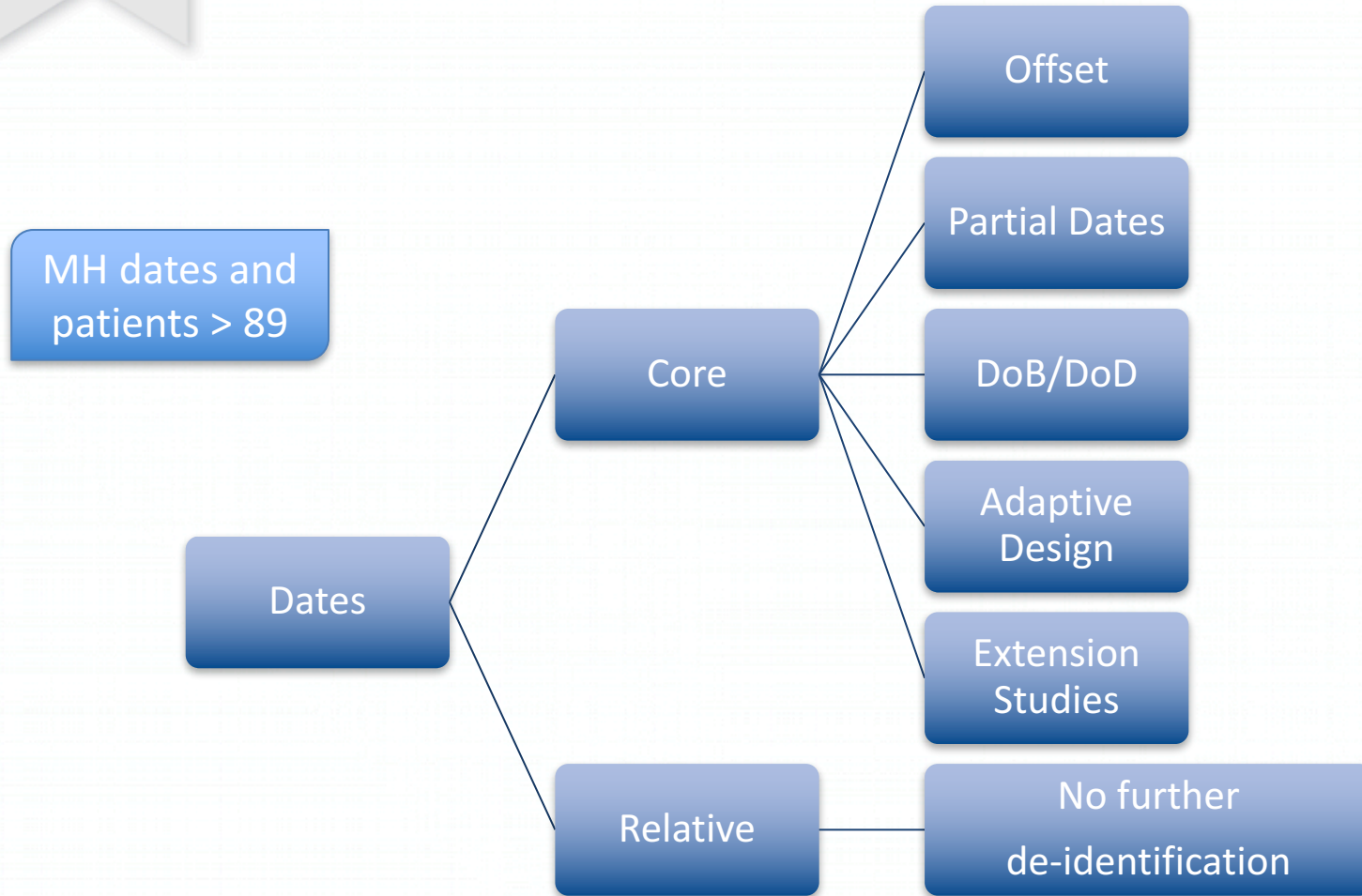
PII of third-party

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





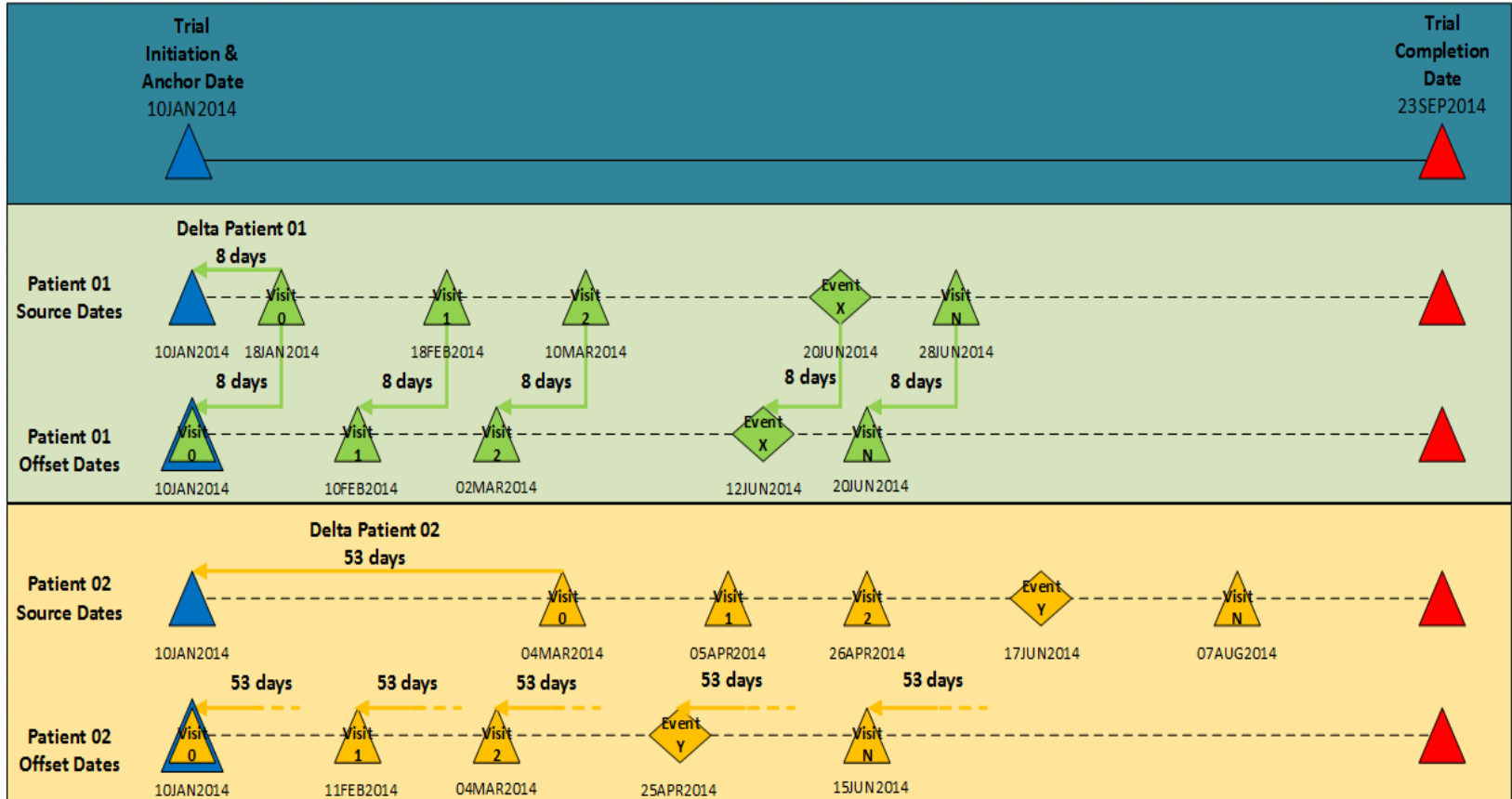
Dates





Dates Offset Recommended Algorithm

(Appendix 1)





Issue with Partial Dates

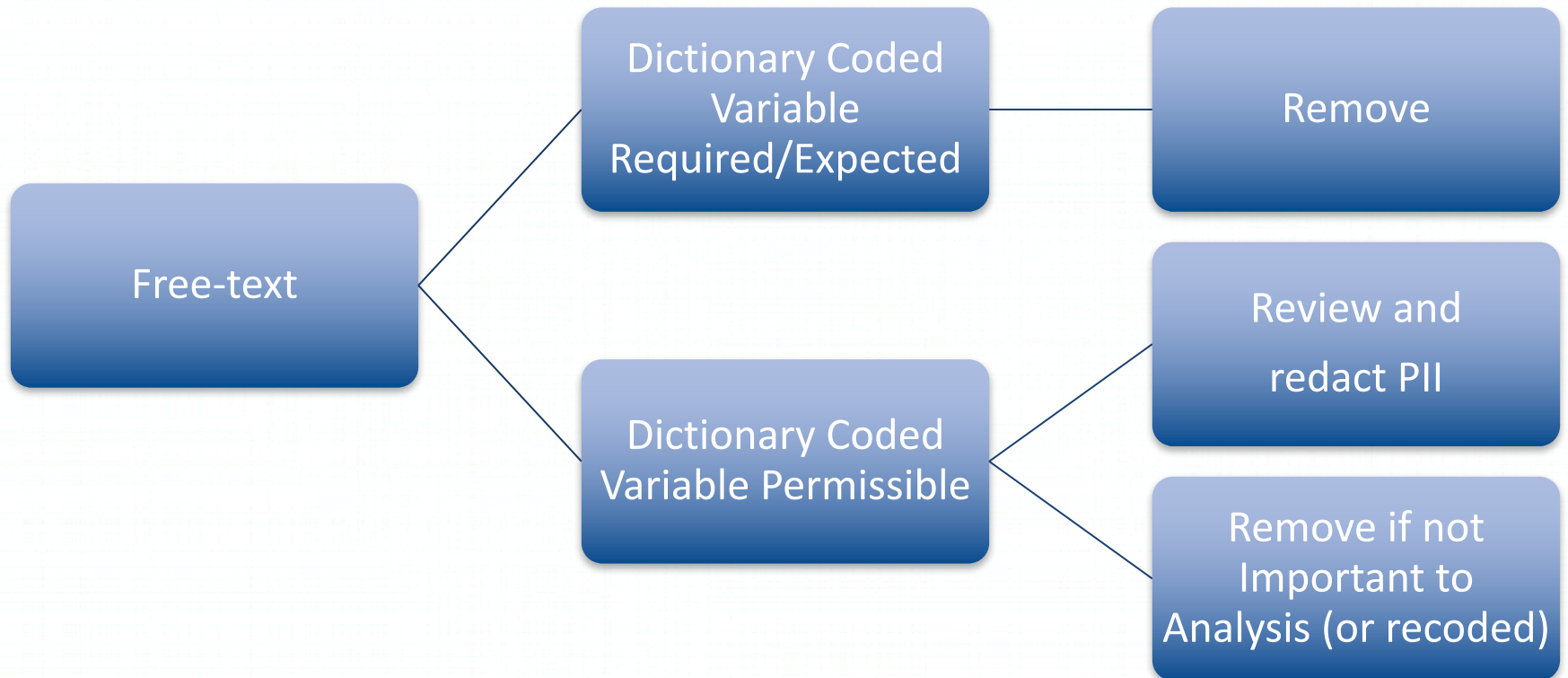
Ex: Delta applied of -14 days

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



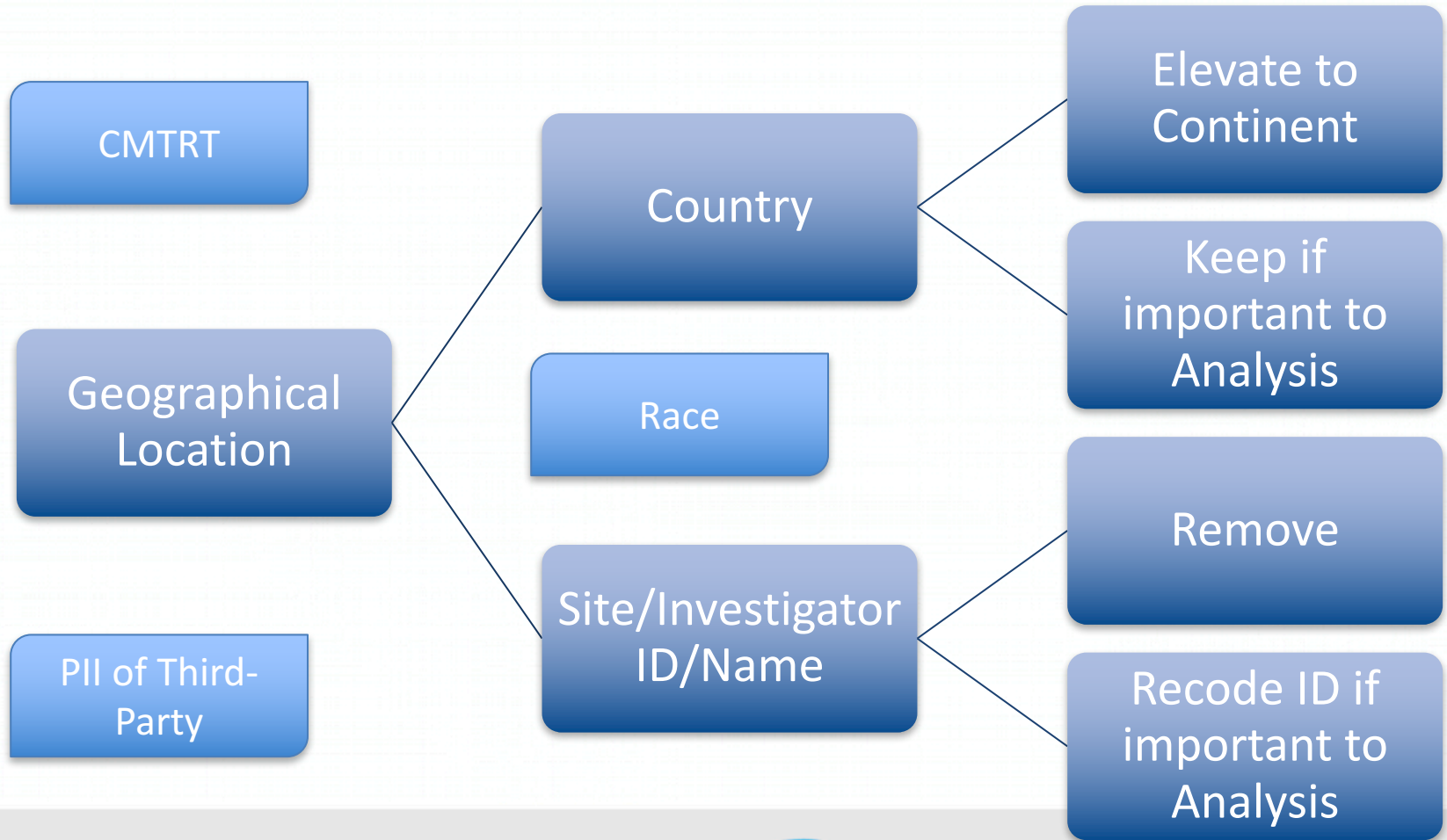


Free-text



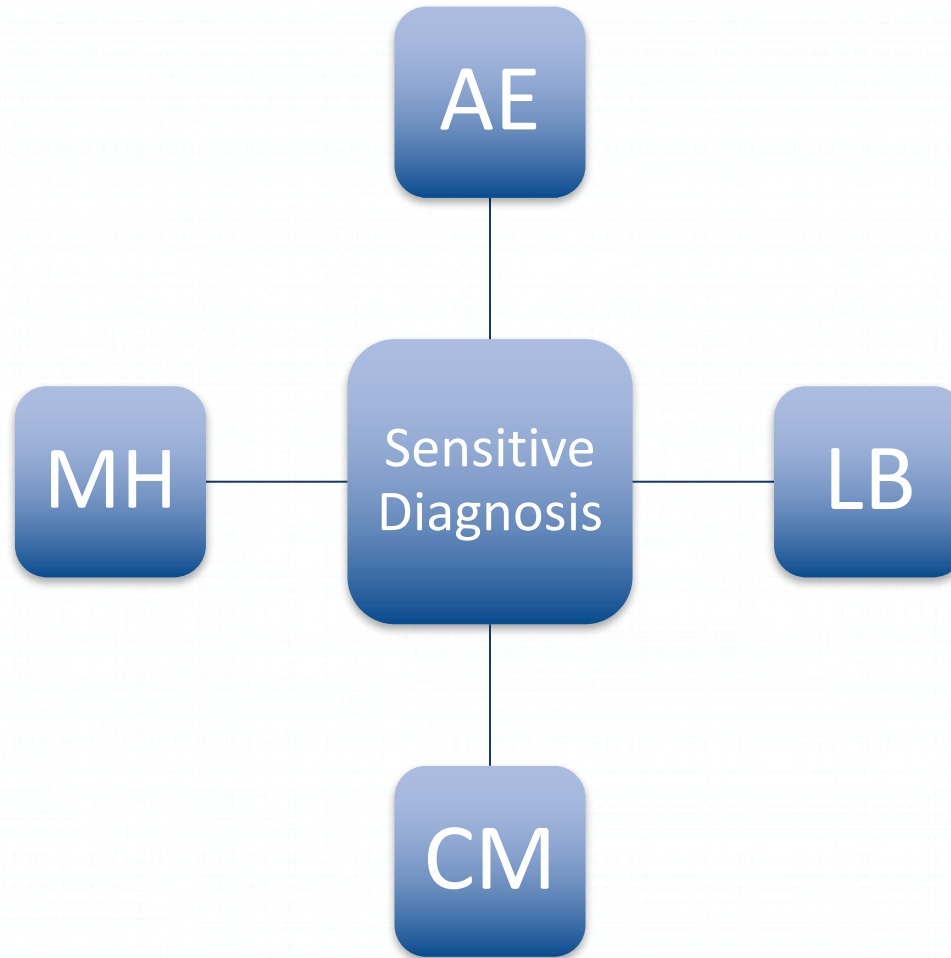


Geographical Location





Sensitive Diagnosis





Deliverable

De-Identification Standards for CDISC SDTM 3.2

700+ downloads

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	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 occurrence 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 occurrence 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





EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

2 March 2016
EMA/90915/2016

External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use

Once a variable has been determined to be an identifier it is necessary to establish whether it should be classified as a direct identifier or a quasi-identifier. This is important because the techniques used to protect direct identifiers are different from those used for quasi identifiers.

PhUSE has defined a set of rules developed to facilitate the assessment of direct and quasi identifiers in the data. These rules help pharmaceutical companies to establish the various categories of personal data that can be found in the clinical reports.



Conclusions

- Many **De-Identification Standards** are available
- Identification of **Direct and Quasi Identifiers** requires detailed understanding of study data and its structure
- Analysis of **Data Sharing Context and Plausible Attackers** is key to Quantitative Risk Assessment
- **Data Utility** is key and must be considered for both research requests and public disclosure



Recommended Readings

- PhUSE De-Identification Standard for SDTM 3.2
 - <http://www.phuse.eu/data-transparency-download>
- PhUSE De-identification Working Group: Providing De-identification Standards to CDSIC Data Models
 - Ferran, El Emam, Nolan, Grimm & De Donder
 - PhUSE 2016 (DH01)
- Calculating the Risk of Re-Identification of Patient-level Data using a Quantitative Approach
 - Kniola
 - PhUSE 2016 (DH09)
- EMA Policy 0070: Data Utility in Anonymized Clinical Study Reports
 - Ferran & Nevitt
 - PhUSE 2017 (DH04)
- Plausible Adversaries in Re-Identification Risk Assessment
 - Kniola
 - PhUSE 2017 (DH09)

Thanks!

Jean-Marc Ferran

Consultant & Owner, Qualiance ApS



dk.linkedin.com/in/jeanmarcferran/



@QualianceTwitta