cdisc

2023
CHINA
INTERCHANGE
BEIJING | 25-26 AUGUST

An Introduction to Privacy Methodology

陆欢 Huan LU, Associate Manager, Statistical Programming
Biostatistics and Programming, Sanofi

Meet the Speaker

陆欢 Huan LU

Title: Associate Manager, Statistical Programming

Organization: Sanofi

5+ years in clinical trial research in data science, clinical data sharing and transparency and innovation development in analytics and reporting tools for submission.

Graduated from George Washington University majoring in Mathematics, Applied Mathematics and Statistics, be with Sanofi since 2018 in Clinical Science and Operation department.

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

Disclaimer

The presentation below discusses a proposed privacy methodology developed by TransCelerate for use with clinical trial data.

Nothing in the methodology or this presentation should be construed to represent or warrant that persons using the methodology have complied with all applicable laws and regulations.

All individuals and organizations using this methodology bear responsibility for complying with the applicable laws and regulations for the relevant jurisdiction.

Agenda

1. WHAT is Privacy Methodology?
2. WHY is Privacy Methodology needed?
3. HOW to apply Privacy Methodology?

WHAT is

Privacy Methodology?

TransCelerate
BIOPHARMA INC.

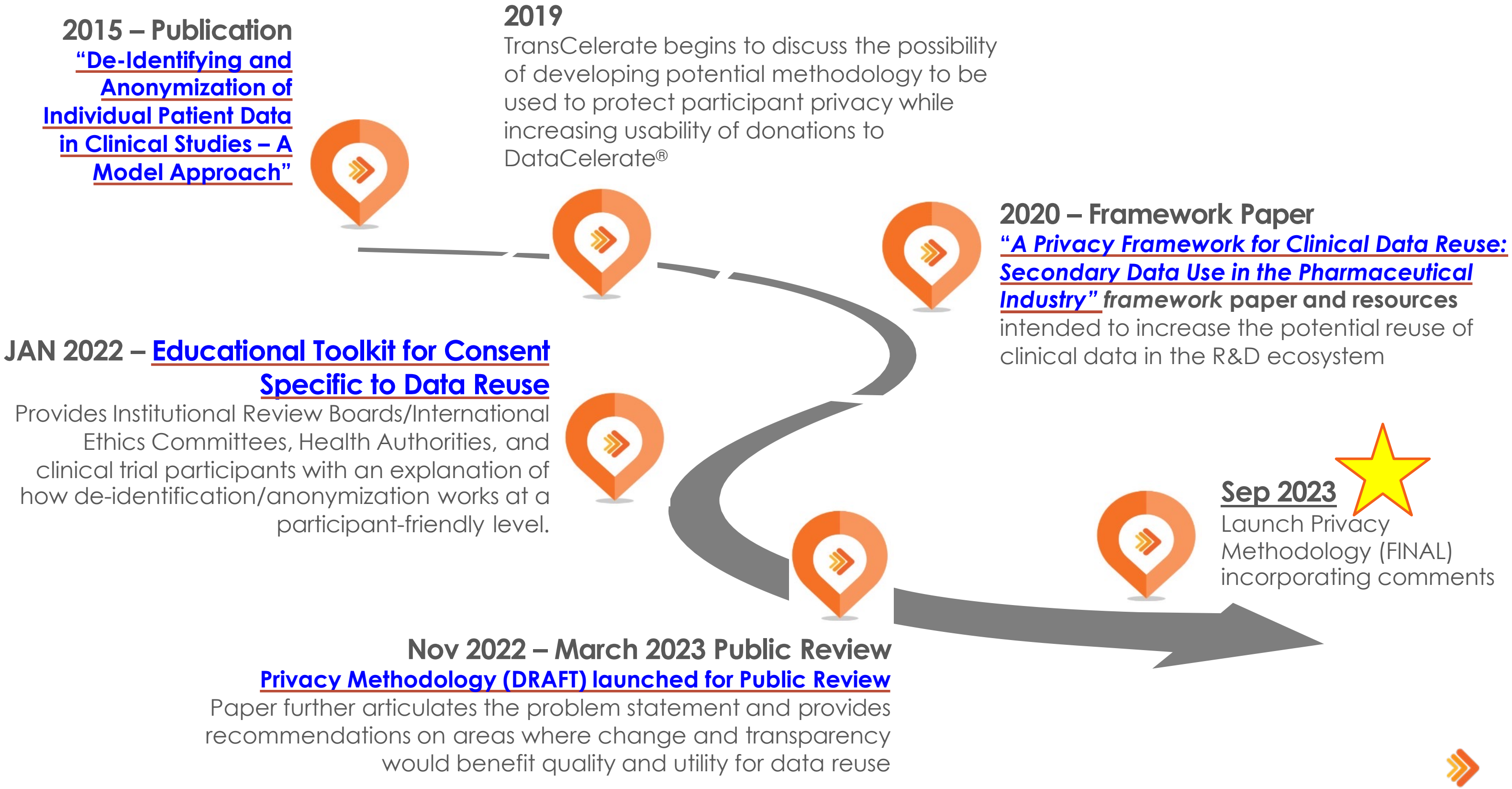
Proposed Privacy Methodology to Improve Cross-Industry Data Sharing

August 25-26 2023

CDISC 2023 China Interchange

Huan Lu

TransCelerate Solutions in Data Privacy / Transparency



2015 – Publication
“De-Identifying and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach”

2019
TransCelerate begins to discuss the possibility of developing potential methodology to be used to protect participant privacy while increasing usability of donations to DataCelerate®

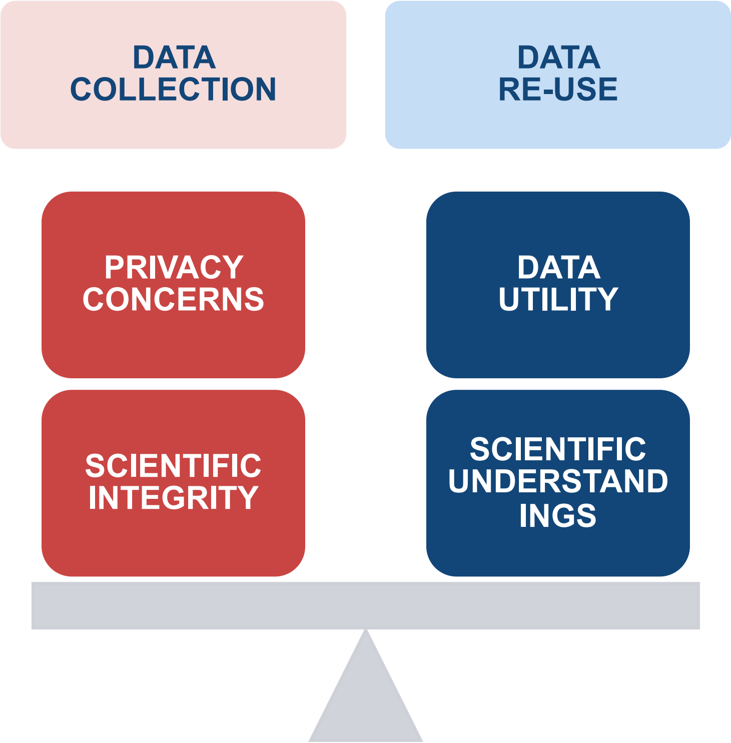
2020 – Framework Paper
***“A Privacy Framework for Clinical Data Reuse: Secondary Data Use in the Pharmaceutical Industry” framework* paper and resources** intended to increase the potential reuse of clinical data in the R&D ecosystem

JAN 2022 – Educational Toolkit for Consent Specific to Data Reuse
Provides Institutional Review Boards/International Ethics Committees, Health Authorities, and clinical trial participants with an explanation of how de-identification/anonymization works at a participant-friendly level.

Nov 2022 – March 2023 Public Review
Privacy Methodology (DRAFT) launched for Public Review
Paper further articulates the problem statement and provides recommendations on areas where change and transparency would benefit quality and utility for data reuse

Sep 2023
Launch Privacy Methodology (FINAL) incorporating comments

Background



DATA COLLECTION

DATA RE-USE

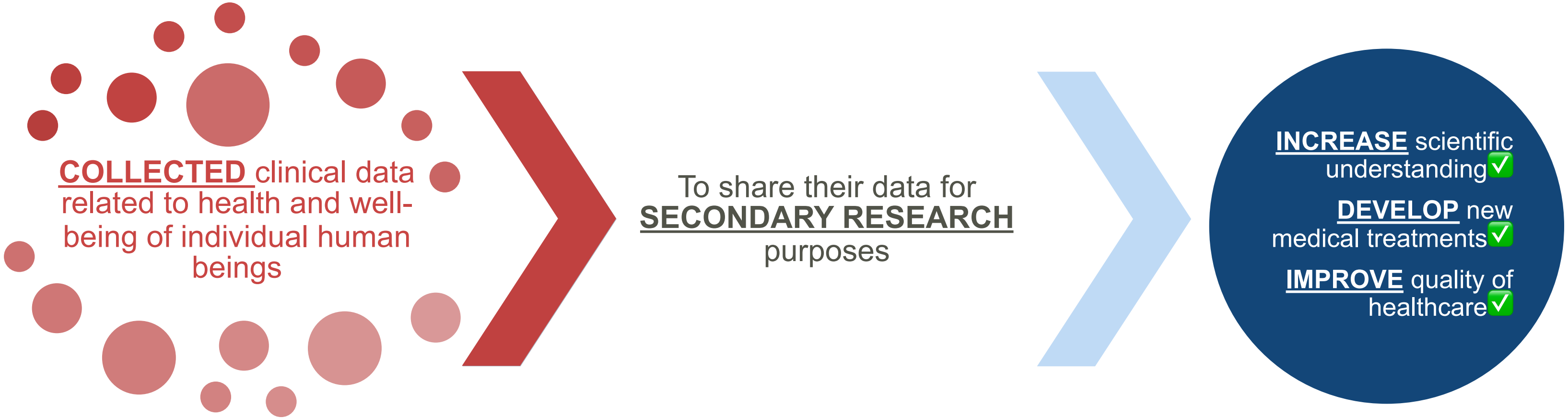
PRIVACY CONCERNS

SCIENTIFIC INTEGRITY

DATA UTILITY

SCIENTIFIC UNDERSTAND INGS

Background (cont'd)



COLLECTED clinical data related to health and well-being of individual human beings

To share their data for **SECONDARY RESEARCH** purposes

INCREASE scientific understanding✅

DEVELOP new medical treatments✅

IMPROVE quality of healthcare✅

What your data may look like at *Study Site*

1. Black Smith
2. 2140 L Street NW, Washington, DC
3. bsmith@aol.com
4. Male
5. 36 y/o
6. Blood Pressure: 136/96 mmHg

Your data is *collected* by the Study Site. They are not the same organisation as the Sponsor.

What your data may look like at *Sponsor Site*

1. *--removed--*
2. USA
3. *--removed--*
4. Male
5. 36 y/o
6. Blood Pressure: 136/96 mmHg
7. *Participant Number: 123-369-001*

Your data is *transformed* by the Study Site so as not to share your identity with the Sponsor.

What your data may look like *Beyond the Sponsor Site*

1. *--removed--*
2. *North America*
3. *--removed--*
4. Male
5. *30 - 39 y/o*
6. Blood Pressure: 136/96 mmHg
7. *Participant Number: 999-888-128*

Changing elements of your data, such as the Participant Number, makes it very difficult to identify you from the study data.

WHY is

Privacy Methodology needed?

Rationale

- **PRESERVE PRIVACY** and confidentiality of research participants
- Reduce research participant burden through **EASIER REUSE** of existing study data
- **DELIVER FASTER** scientific insights
- Provide **GREATER TRANSPARENCY** in the privacy safeguards applied to study data
- **INCREASE** overall data **UTILITY**

HOW to

apply Privacy Methodology?

Unique Identifiers

- **DIRECT IDENTIFIERS**: Values that could be directly linked to the individual participant, such as Unique Subject ID (USUBJID) or serial number. These values should always be scrambled;
- **INDIRECT IDENTIFIERS**: Values which are not directly linked to the individual participant, such as site identifier, vendor identifier, batch/lot numbers. These values should be assessed if de-identification/anonymization is needed.

Unique Identifiers (cont'd)

	BEFORE	AFTER	BEFORE	AFTER	NO CHANGE
STUDYID	**USUBJID**	**USUBJID**	**SITEID**	**SITEID**	**xxSEQ**
1234-5678	1234-5678-10001	1000-0010-00056	USA-0024	011-0110	1
1234-5678	1234-5678-10001	1000-0010-00056	USA-0024	011-0110	2
1234-5678	1234-5678-10002	1000-0010-00301	FRA-0007	023-0074	1

Example of Scrambling Unique Identifiers That Are Linked to the Participant and Retaining the Unique Identifier That Is Not Directly Linked to the Participant

Dates

- **RELATIVE DAY** is a method that uses a specific date (e.g., each participant's randomization date) as a reference day (e.g., day 0, day 1) for a participant, and transforms all other days for that specific participant into the number of days relative to the reference date. The original date variables are to be redacted afterward;
- **DATE OFFSET** is a method where the original date is transformed by adding or subtracting a defined number of days from the original date. This method can be used with different date ranges and includes an option to use a different randomized offset on an individual basis.

Dates (cont'd)

		BEFORE	AFTER	BEFORE	AFTER	DATE OFFSET
STUDYID	**SUBJID**	**DATE1**	**DATE1**	**DATE2**	**DATE2**	
1234-5678	1234-5678-11	2021-03-02	2021-03-24	2022-07-08	2022-07-30	22
1234-5678	1234-5678-12	2021-08-29	2021-09-24	2022-07-31	2022-08-26	26
1234-5678	1234-5678-13	2021-11-06	2021-11-01	2022-07-29	2022-07-24	-5

Example of Date Offset

Verbatim/Free Text

- **VERBATIM TEXT** (otherwise known as "**FREE TEXT**") may be collected across a wide range of case report form (CRF) pages and thus may be present in many datasets. Since any text strings could be captured within these fields, it is likely that personal data are collected.

Verbatim/Free Text (cont’d)

		BEFORE	AFTER
STUDYID	**SUBJID**	**COVAL**	**COVAL**
1234-5678	1234-5678-10001	Comment 1	—redacted—
1234-5678	1234-5678-10002	Comment 2	—redacted—
1234-5678	1234-5678-10003	Comment 3	—redacted—

Example of Redaction of Verbatim/Free Text

Banding of Variables

- **STATIC BANDS**: Static bands have the same cut-off points for bands for each study (e.g., 20–29, 30–39);
- **SEMI-FIXED BANDS**: Semi-fixed bands support the combination of selected static bands to increase the number of participants within a band to reduce privacy risks (e.g., 20–39 when combining bands from static banding example above);
- **FLEXIBLE BANDS**: Flexible bands are individual bands created for each set of data to preserve scientific utility as much as possible. Flexible bands come in two varieties: single-dimensional banding, which generates independent bands on each relevant variable within the dataset, and multi-dimensional banding, which consists of creating bands on two or more variables (e.g., based on age and some prior medical history diagnosis).

Banding of Variables (cont’d)

		BEFORE	AFTER
STUDYID	**USUBJID**	**AGE**	**AGE**
1234-5678	1234-5678-10001	46	46-48
1234-5678	1234-5678-10002	47	46-48
1234-5678	1234-5678-10003	48	46-48
1234-5678	1234-5678-10004	50	50-50
1234-5678	1234-5678-10005	50	50-50
1234-5678	1234-5678-10006	50	50-50
1234-5678	1234-5678-10007	50	50-50

Example of Single-Dimensional Flexible Banding of Age

Patient Demographics

- Demographic data collected as part of the clinical study provide an essential element to describe the study population as part of the planned analyses. Information such as sex, race, ethnicity, and age are valuable to retain for data sharing and further data reuse. However, combined and correlated together with other quasi-identifying information, the overall risk of re-identification might increase if no further measures are applied.
- Assuming a well-balanced study population and following re-identification risk assessments, most Data Providers retain information about sex, race, and ethnicity in the dataset, to the extent that **NO LOW FREQUENCY GROUPS** are present in the data.
- **OUTLIERS** are interesting and important factors in the analyses but must be given special consideration because there may be an increased risk of re-identification for the specific participant.

Data with Low Frequencies

- Variables that contain some **LOW FREQUENCY** values may **LEAD TO AN INCREASED RISK OF RE-IDENTIFICATION** (e.g., data that, after grouping of several variables and their expression levels, applies to a very small cell/group size). This may be further compounded by data with multiple variables of low frequencies.

Data with Low Frequencies (cont’d)

		BEFORE	AFTER	BEFORE	AFTER
USUBJID	**DOMAIN**	**SEX**	**SEX**	**RACE**	**RACE**
1234-5678-USA003-10001	DM	F	-- Redacted --	WHITE	WHITE
1234-5678-DNK001-10002	DM	F	-- Redacted --	WHITE	WHITE
1234-5678-POL002-10003	DM	M	-- Redacted --	WHITE	WHITE
1234-5678-GER002-10004	DM	F	-- Redacted --	UNKNOWN	-- Redacted --

Example of Redacting Low Frequency Sex and Race

Data with Low Frequencies (cont'd)

		BEFORE	AFTER
USUBJID	**DOMAIN**	**AEDECOD**	**AEDECOD**
1234-5678-POL002-10003	AE	Oligospermia	-- Redacted --
1234-5678-POL002-10003	AE	Headache	Headache
1234-5678-POL002-10003	AE	Diarrhea	Diarrhea
1234-5678-POL002-10003	AE	Rhinitis	Rhinitis

Example of Redacting Additional Information Revealing Trial Participants' Sex Through a Rare Event (e.g., Event of Oligospermia)

Sensitive Information

- Sensitive information is highly personal in nature and **DISCLOSURE MAY CAUSE HARM** to the individual participant (e.g., data related to alcohol abuse, drug use, conditions such as HIV/AIDS, mental health information).

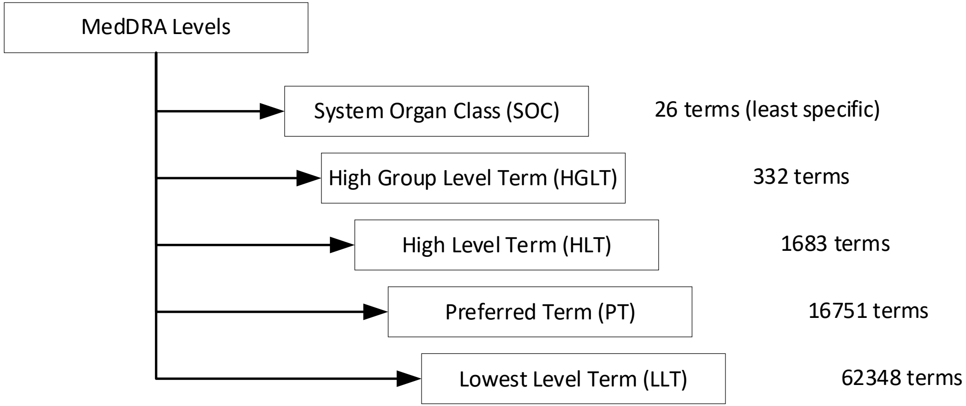
Sensitive Information (cont’d)

			BEFORE	AFTER
USUBJID	DOMAIN	SUCAT	SUOCCUR	SUOCCUR
1234-5678-USA003-10001	SU	ALCOHOL HISTORY	N	-- Redacted --
1234-5678-USA003-10001	SU	TOBACCO HISTORY	Y	-- Redacted --
1234-5678-GER002-10004	SU	ALCOHOL HISTORY	Y	-- Redacted --
1234-5678-GER002-10004	SU	TOBACCO HISTORY	Y	-- Redacted --

Example of Removing Sensitive Substance Usage Records

Adverse Events

- Collection of adverse events during the clinical study plays a pivotal role to assess the safety of the investigational product. The MedDRA dictionary provides codes to describe the adverse events terms to 5 levels:



MedDRA Levels

System Organ Class (SOC)	26 terms (least specific)
High Group Level Term (HGLT)	332 terms
High Level Term (HLT)	1683 terms
Preferred Term (PT)	16751 terms
Lowest Level Term (LLT)	62348 terms

Adverse Events (cont'd)

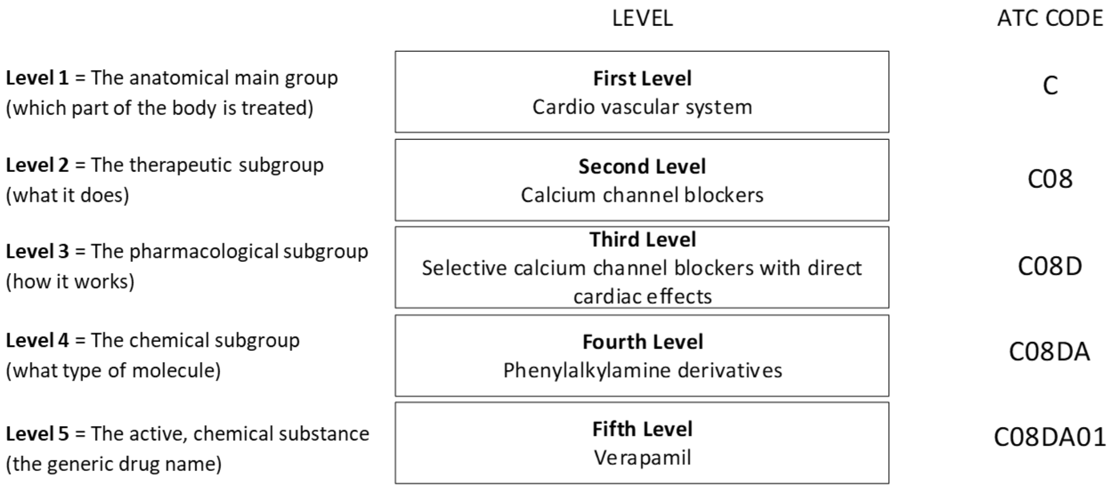
- The more levels of MedDRA codes that are shared with respect to the adverse event, the greater the level of utility that can be ascertained from the information. However, the more granular the description of the adverse event (using the MedDRA codes), the greater the possibility that a participant could be re-identified using these data or in combination with other data pertaining to them.
- Consideration needs to be made regarding **THE NATURE OF THE ADVERSE EVENT** and whether other information, when combined, increases the identifiability of a specific participant.

Medications

- Information collected during clinical studies includes the medication history of research participants; this consists of the current and concurrent medications that the participant is taking at the time of the study as well as medications taken in the past. Accurate documentation of this information is critical for researchers to understand whether a participant’s medication history should be treated as a confounding factor in the original clinical study or in subsequent analyses.

Medications (cont’d)

- **VERBATIM** medication term (CMTRT/CMMODIFY) is removed and replaced with the corresponding WHO ATC drug code



	LEVEL	ATC CODE
Level 1 = The anatomical main group (which part of the body is treated)	**First Level** Cardio vascular system	C
Level 2 = The therapeutic subgroup (what it does)	**Second Level** Calcium channel blockers	C08
Level 3 = The pharmacological subgroup (how it works)	**Third Level** Selective calcium channel blockers with direct cardiac effects	C08D
Level 4 = The chemical subgroup (what type of molecule)	**Fourth Level** Phenylalkylamine derivatives	C08DA
Level 5 = The active, chemical substance (the generic drug name)	**Fifth Level** Verapamil	C08DA01

Source: Verapamil example from WHO website

Geographic Location

- Information related to geographic location can serve as **INDIRECT IDENTIFIERS** that may increase the risk of re-identification when combined with other available information in participant-level data. When associated with an individual participant’s data, location information that is specific to the study site could allow for links between data and facilitate the recreation of an individual participant’s profile.

Geographic Location (cont’d)

	BEFORE	AFTER	BEFORE	AFTER	AFTER
SUBJID	**SITEID**	**SITEID**	**COUNTRY**	**COUNTRY**	**REGION**
10001	USA-0001	011-0110	USA	USA	NORTHERN AMERICA
10002	USA-0001	011-0110	USA	USA	NORTHERN AMERICA
10015	USA-0002	011-0221	USA	USA	NORTHERN AMERICA
10037	POL-0001	054-0080	POL	-- Redacted --	EUROPE
10042	GER-0001	036-0073	GER	-- Redacted --	EUROPE
10068	DNK-0001	088-0004	DNK	DNK	EUROPE
10070	DNK-0002	088-0013	DNK	DNK	EUROPE

Example of Retaining Country Information for USA and DNK While Generalizing Country Information to the Continent Level for Poland and Germany (Only One Site per Country)

Records of Participants Who Have Died

- Some research participants may die during the conduct or post-treatment phases of the study. In some countries, privacy regulations may apply post-mortem (e.g., 10 years after death). While records of participants who have died is of analytical value, care must be taken to recognize the personal nature of this information and the need to show courtesy to the next of kin.
- The Data Provider is responsible to ensure that use of data from participants who have died is compliant with any applicable regulations.

What the future holds?

Considerations of Novel Areas for Privacy Measures

- Data Derived From **GENOMIC** Data

  While the information collected from clinical trial participants offers potentially great scientific utility, certain types of genetic/genomic data may represent high re-identification risk and, as such, must be removed.

- **SEASONALITY**

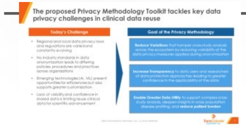
  Sharing of seasonality-related information requires further exploration to understand the number of datasets where seasonal information is relevant, and the impact of different de-identification/anonymization approaches for such datasets to accommodate original month and hemisphere.

Conclusion

- De-identification/anonymization of clinical data is an **EVOLVING** area, in both regulations and practice. As clinical data sharing and reuse matures and becomes more common in the pharmaceutical industry due to advances in data analysis technology, data privacy methodologies (including the one described in this paper) must be revised to ensure the continued **BALANCE** between protecting the **PRIVACY** of research participants and optimizing scientific data **UTILITY**.

Solution Overview: Privacy Methodology Toolkit

The Privacy Methodology Toolkit consists of 4 components

Resource	Name	Description
 CLINICAL DATA SHARING – A PROPOSED METHODOLOGY TO ENABLE DATA PRIVACY WHILE IMPROVING DATA UTILITY NOVEMBER 2022, DRAFT VERSION	**Privacy Methodology for Clinical Data Reuse FINAL** *(methodology paper)* *[Core Solution]*	Building upon existing solutions in the industry, this proposed methodology identifies data privacy approaches transforming *emergent thinking from the TransCelerate MCs into practical recommendations & considerations for high-value variable types in clinical data reuse.*
 APPENDIX 2: DATA TRANSPARENCY CHECKLIST	**Data Transparency Checklist** *(template)* *[Core Solution]*	Standalone template for adopting data providers to provide valuable information and transparency to their data transformation activities and enable reuse of the data.
	Public Review Response Document *[Supporting Resource]*	Document of the consolidated public responses received during the public review period held Nov 2022 to March 2023.
 The proposed Privacy Methodology Toolkit tackles key data privacy challenges in clinical data reuse Today's Challenge - Regional and local data privacy laws and regulations are varied and constantly evolving - No industry standards in data anonymization leads to differing policies, procedures and practices across organizations - Emerging technologies (AI, ML) present opportunities for efficiencies but also supports greater customization - Lack of visibility and confidence in shared data is limiting reuse clinical data for scientific advancement Goal of the Privacy Methodology **Reduce Variations** that hamper cross-study analysis across the ecosystem by reducing variability of the data privacy measures applied during anonymization **Increase Transparency** to data users and researchers of data protection approaches leading to greater confidence in the applicability of the data **Enable Greater Data Utility** to support complex cross-study analysis, deepen insights in cross population disease profiling, and **reduce patient burden**	**Informational Resource** *[Supporting Resource – PPT and short video]*	A change resource that provides an overview of the solution toolkit – what is it, how it was developed, the value for the ecosystem, how it can be implemented, etc.

New References

Available Sept 5 2023 - Revised Data Privacy Methodology Paper, Transparency Checklist and Educational Toolkit.

Additional References

2015 – Publication: "De-Identifying and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach"

2020 - Framework Paper: *"A Privacy Framework for Clinical Data Reuse: Secondary Data Use*
in the Pharmaceutical Industry"

Jan 2022 - Educational Toolkit for Consent Specific to Data Reuse

Nov 2022 - March 2023 Public Review: Privacy Methodology (DRAFT) launched for Public Review

Data Privacy Education poster: TransCelerate Privacy Page Educational Poster_final (transceleratebiopharmainc.com)

Thank You!

For further questions, please contact huan.lu@sanofi.com