



Real Life Sciences

15Nov2022

PHUSE EU Connect 2022

Presentation DH04:

Processing Data and Documents with Natural Language

*Processing and Quantitative Anonymization Strategies in Support
of CTIS Disclosure Requirements*

Presenter:

Elliot Zimmerman, CEO Real Life Sciences

Agenda

CTIS Highlights and What to Know

Are You Ready?

Anonymizing Documents and Data

Risk Based De-Identification Use Cases

Q&A

***Processing Data and Documents
with Natural Language
Processing and Quantitative
Anonymization Strategies in
Support of CTIS Disclosure
Requirements***



CTIS Highlights - What to Know

Through July 2022, 195 Clinical Trial Applications have been submitted via the CTIS.

Regulation (EU) No 536/2014 The Clinical Trials Regulation (CTR)

The CTR aims to create an environment that is favourable for **conducting clinical trials in the EU** with the highest standards of safety for participants and increased **transparency of CTs information**. It replaces and expands the scope of the existing EU Clinical Trials Directive 2001/20/EC.

The regulation aim is to benefit **Authorities** (member states, Ethics Committees, etc.), **Sponsors** and the **General Public** including patients, scientists, HCPs, media, citizens.

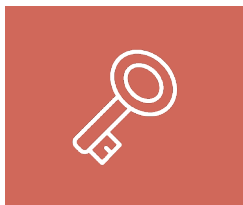
Clinical Trial Information System (CTIS)

The implementation of the CTR is supported by the **Clinical Trials Information System (CTIS)**.

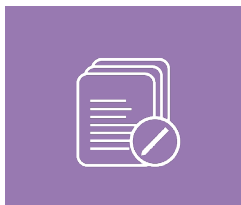
The rules established in the Regulation is supported through a dedicated EU Portal and Database (the CTIS) that will* become the **single entry point** for submitting CTs information in the EU and will support the **daily business processes** of Member States and sponsors throughout the life cycle of their CTs.

Sources: ema.europa.eu and [link](#)

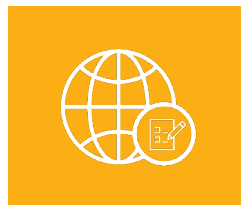
Note: EudraCT can continue to be used through Jan. 2025.



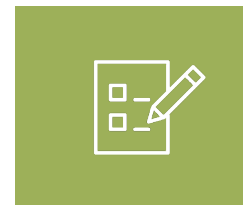
Personal Data and
**Commercially
Confidential
Information** are exempt
from disclosure



Even in the case of a
deferral request, public
and non-public version
documents **must be
submitted
simultaneously**



Following a decision
(of any kind) by Member
State(s), public version
data and documents will
be made available to the
public in CTIS - unless a
deferral was requested
and approved



**A Plain Language
Summary** is required
for each Phase 2-4
trial within 12 mo. of
the close of the trial



1

Each document will be submitted in anonymized / redacted form ("for publication") and unredacted form ("not for publication") simultaneously.

2

Example documents for publication include but are not limited to:

- Protocol
- Investigator Brochure
- Informed Consent Information Sheet

3

Expectation is for sponsors to submit high quality dossiers in CTIS to minimise need to raise RFI.

Example Document	For Publication	Not For Publication
Initial Clinical Trial Application - Cover Letter	✓	✓
Protocol and synopsis	✓	✓
EU CT Form	✓	✓
Multi-Trial Substantial Modification form	✓	✓
Investigator Brochure	✓	✓
Informed Consent Information Sheet	✓	✓
Quality related documentation / quality assessment reports		✓
Financial arrangements		✓

Listing represents a subset of documents. Please refer to [Table I](#) for more details.



1

The CTR is in effect and the CTIS is live as of 31Jan2022 with a phased rollout schedule.

2

"For publication" documents must be uploaded at the time the "Not for Publication" document is uploaded.

3

Deferral options are available and vary based on trial category.

4

A Plain Language Summary is required within 12 months of completion of each Phase 2-4 trial.

5

Protecting subject data and privacy and commercially confidential information is critical for your business.



The CTR has taken transparency and disclosure of in-process and completed trials to another level.

The resulting impact to study teams and risks associated with disclosure of subject data and CCI must be managed.

How can I manage this?

1

Minimization during authoring; templates and style guides

2

Clearly defined processes / SOPs for preparing "for publication" documents

3

Redaction and anonymization processes – and proper tools

4

Consistent and timely identification of CCI

5

Expertise, tools and processes to author PLS



Anonymizing Documents and Data

Preparation of 'Not For Publication Documents'

Personal data should be pseudonymized

Principles of minimization should be applied

Example documents that are Not For Publication only
include: IB, IMPD Section S+E, Assessment reports,
Inspection reports, etc.

Preparation of 'For Public Documents'

Personal data of trial participants must be anonymized

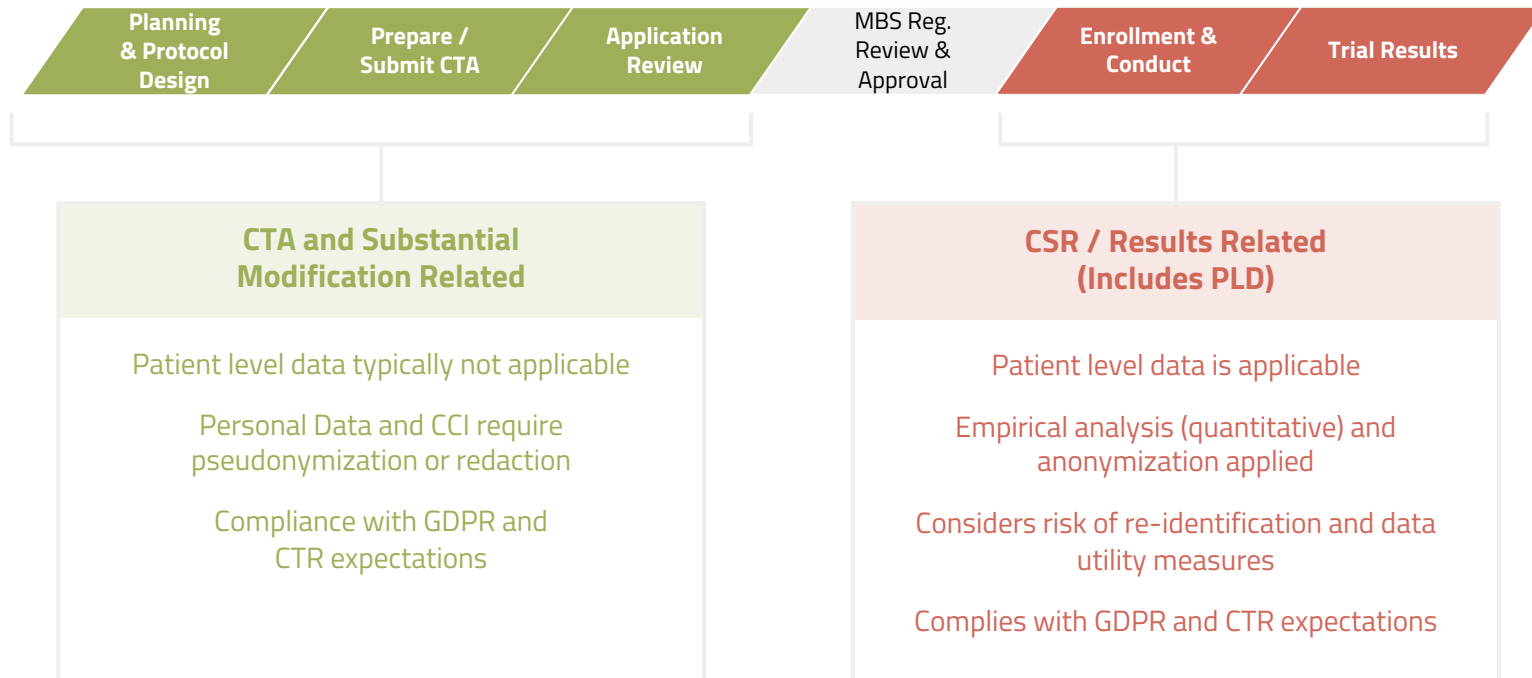
Removal of entire sections / pages is not
considered appropriate

CTR's expectation is to maintain data utility while
ensuring adequate data subject protection

Akin to Policy 0070 guidance, the anonymization must
ensure that the risk of re-identification is acceptably low
and transformations may not result in a different
interpretation of the trial documentation

CCI redactions must consider the timing of publication
of the document itself*

*Refer to "[CTIS Guidance document on how to approach the protection of personal data and commercially confidential information in documents...](#)" Section 4.





Illustrative Example - For Discussion Purposes Only

Original Document Text No Transformations

Patient **123456**, visited on **130ct2020** was a **33** year old **male** in Treatment Group A (200mg) from the **United States** with a medical history of **diabetes type II** and **hypertension**. Drug was safely shipped and stored at 35.2 degrees celsius.

Rules Based Method All personal data and CCI are marked

Patient **438393**, visited on **PPD** was a **30-40** year old **PPD** in Treatment Group A (**CCI** mg) from the **PPD** with a medical history of **PPD** and **PPD**. Drug was safely shipped and stored at **CCI** degrees celsius.

Risk based Anonymization aka Quantitative Risk Assessment

Patient **438393**, visited on **29Nov2020** was a **30-35** year old **male** in Treatment Group A (**CCI** mg) from the **North America** with a medical history of **(Med. History)** and **(Med. History)**. Drug was safely shipped and stored at **CCI** degrees celsius.

*Risk-based patient level date offsetting applied.

Risk Modeling: Risk of Re-identification (ROR) Metrics

Data transformation - Attribute metadata

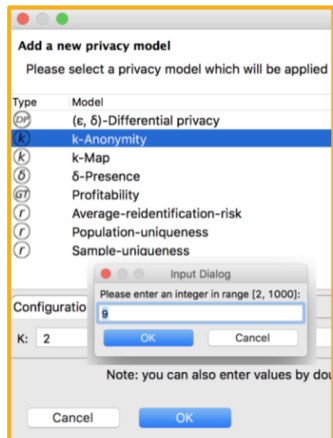
Type: Quasi-identifying Transformation: Generalization

Minimum: All Maximum: All

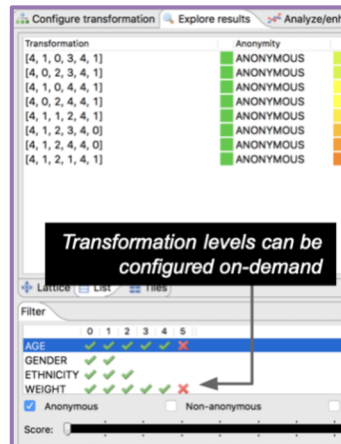
	Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
18	[15, 20]	[10, 20]	[0, 20]	[0, 40]	*	
19	[15, 20]	[10, 20]	[0, 20]	[0, 40]	*	
20	[20, 25]	[20, 30]	[20, 40]	[0, 40]	*	
21	[20, 25]	[20, 30]	[20, 40]	[0, 40]	*	
22	[20, 25]	[20, 30]	[20, 40]	[0, 40]	*	
23	[20, 25]	[20, 30]	[20, 40]	[0, 40]	*	
24	[20, 25]	[20, 30]	[20, 40]	[0, 40]	*	
25	[25, 30]	[20, 30]	[20, 40]	[0, 40]	*	
26	[25, 30]	[20, 30]	[20, 40]	[0, 40]	*	
27	[25, 30]	[20, 30]	[20, 40]	[0, 40]	*	
28	[25, 30]	[20, 30]	[20, 40]	[0, 40]	*	
29	[25, 30]	[20, 30]	[20, 40]	[0, 40]	*	
30	[30, 35]	[30, 40]	[20, 40]	[0, 40]	*	
31	[30, 35]	[30, 40]	[20, 40]	[0, 40]	*	
32	[30, 35]	[30, 40]	[20, 40]	[0, 40]	*	
33	[30, 35]	[30, 40]	[20, 40]	[0, 40]	*	
34	[30, 35]	[30, 40]	[20, 40]	[0, 40]	*	
35	[35, 40]	[30, 40]	[20, 40]	[0, 40]	*	
36	[35, 40]	[30, 40]	[20, 40]	[0, 40]	*	

Example transformation levels/hierarchy for the AGE identifier

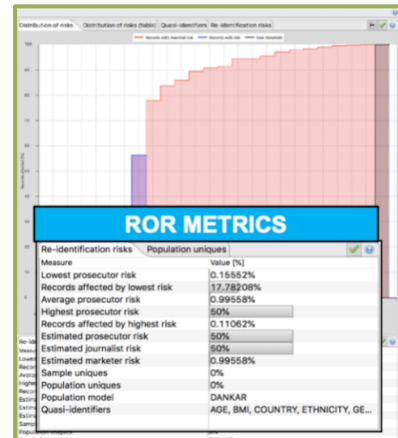
Default or custom templates across identifiers containing different transformation "levels" can be used...



...in combination with the selected privacy model / threshold...



...transformation options that meet the privacy thresholds are presented and can be configured...



...prior to transforming data, different types of ROR measures are generated across individual records and in aggregate.



As shown on the table at the bottom of Figure 14-1, there was a disproportionate number of subjects in the xanomeline treatment groups who experienced a special interest dermatologic adverse event with 74% (CCI) of xanomeline low dose and 73% (CCI) of xanomeline high dose subjects with at least one event of special interest compared to 34% (CCI) of placebo subjects. This high rate of dermatologic events most likely contributed to the larger proportion of subjects in the active treatment groups who discontinued the study due to an adverse event (Table 14-1.02).

An analysis of the time to the first dermatologic event indicated that the median time to first event was significantly different (PPD) between treatment groups with a median time of 33 days in the xanomeline low dose group (PPD CI: 27 – 48 days) and 36 days in the xanomeline high dose group (PPD CI: 24 – 46 days) compared to placebo, in which the median time to first event was not estimable (Figure 14-1).

12.3.4. Narratives of Deaths, Other Serious Adverse Events and Certain Other Significant Adverse Events

Narratives for subjects who died while on-study are included below. Narratives for subjects who reported other serious adverse events or other significant adverse events are not included in this study report due to the nature of this CDISC Pilot Project.

Subject 01-701-9535, a 70-79-year-old woman with Alzheimer's dementia, began receiving xanomeline low dose on 21 November 2012. The subject was diagnosed with Alzheimer's dementia in 2010. The subject also had non-insulin dependent (Adverse

Commercially Confidential Information and sensitive terms have been marked and redacted

From Quantitative Risk Assessment, transformations have been applied

Subject ID	Pseudonymized
Subject Age	Transformed - 10 year bands
Visit Dates	Transformed - offset per patient
Adverse Events	Retained, transformed to SOC or categorically transformed
CCI & Sensitive Terms	Redacted



Risk Based De-identification Use Cases



1

Use Case 1

Anonymizing Small Patient Population
Trial Data

2

Use Case 2

Anonymizing Outlier Patient Records

*A risk based, advanced
quantitative risk assessment
will inform the data controller of
the risk of re-identification and
data usefulness while
empirically evaluating the trial
datasets and its characteristics.*



Situation

With a small patient population, anonymizing trial data for rare diseases could lead to a large volume of redaction and suppression; therefore sacrificing clinical utility.

Illustrative Example – Trial Characteristics

- Phase 2
- # Subjects: 42
- Disease: Amyloidosis
- Trial locations: USA, Poland, Canada
- Race: Caucasian and non-caucasian
- Global subjects with disease: 12,000 in US

Challenge

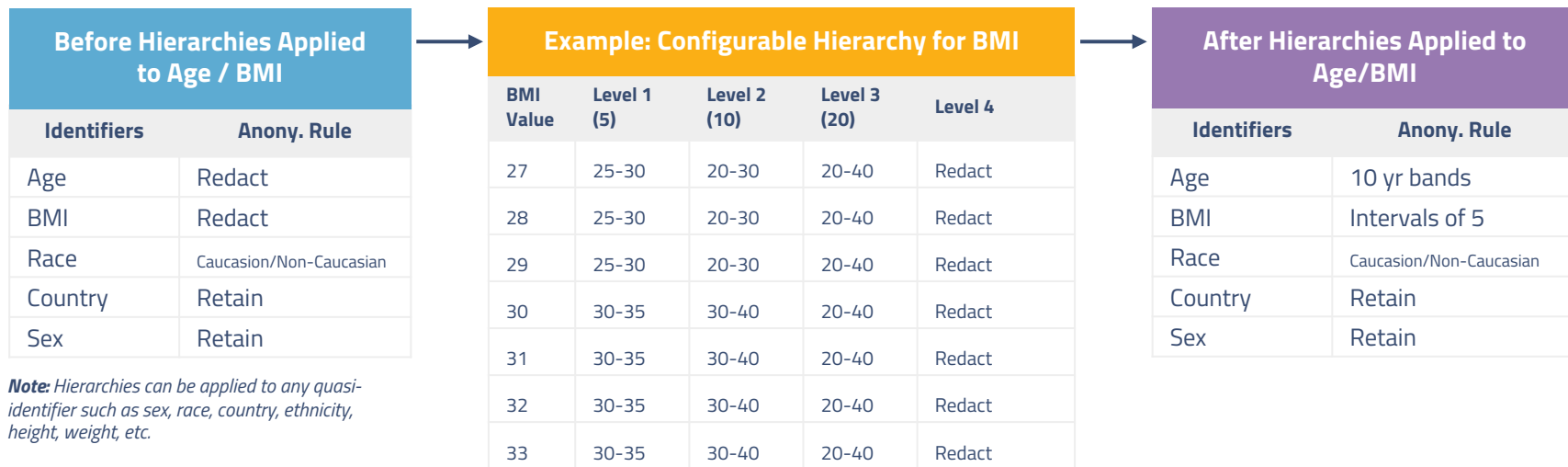
A 9% (or less) risk of re-identification (ROR) measure will be expected by the regulator. Internal sponsor expectations may be 5% ROR.

A 9% risk threshold groups trial subjects into groups of 11 (equivalence class size).

- **Age** and **BMI** are desirable to retain and fully disclose due to the nature of the study, however, establishing class sizes of 11 is a challenge
- **BMI** values are extremely unique and is captured at the decimal level
- As a result, for a study with a 9% ROR, retaining **Age/BMI is not possible without additional risk based techniques**

Solution

Generalizing is an anonymization technique used to help preserve data utility across attributes which otherwise would be redacted. Generalization takes into consideration the target risk threshold (e.g., 9%) and leverages **anonymization hierarchies** to provide options for preserving data utility while adhering to the desired threshold. Hierarchies are established for each quasi-identifier to allow a risk based system to systematically evaluate these generalization options.



Results Comparison			
	Rules Based	Risk Based	Description
Target Risk of Re-Identification	9%	<9%	
Actual ROR	8.42%	6.92%	
Information Loss	38.60%	29.70%	Measures the amount of data that is redacted. This value decreases as more information is retained/generalized.
Granularity	54.07%	65.60%	Captures the granularity of the data by measuring the extent to which the data was preserved. The higher the granularity, the closer the transformed value was to the original value.
Age	Redacted	10 yr Band	
BMI	Redacted	Interval of 5	



1

Use Case 1

Anonymizing Small Patient Population
Trial Data

2

Use Case 2

Anonymizing Outlier Patient Records

*A risk based, advanced
quantitative risk assessment
will inform the data controller of
the risk of re-identification and
data usefulness while
empirically evaluating the trial
datasets and its characteristics.*

Situation

Adverse Events and Serious Adverse Event data at the participant level record in a trial dataset may result in a high number of suppressed records to achieve the risk of re-identification threshold.

Illustrative Example – Trial Characteristics

- Phase 3
- # Subjects: 309
- Disease: Antisocial Personality Disorder
- Trial locations: Canada, S. Korea, Italy
- Race: White, Asian, African-American
- Global subjects with disease: 400,000+ in U.S

Challenge

Record suppression as a means to protect ROR can limit clinical usefulness – information loss will be impacted. Record suppression technique does not target specific identifiers, rather it redacts / suppresses the entire participant record from the dataset.

The primary focus and concern is optimization of clinical utility by limiting full suppression, not ROR itself.

Solution

Targeted Suppression is a de-identification technique that aims to significantly reduce suppression of broad sections of clinical data by targeting only the specific identifier(s) that negatively impact the risk of re-identification. This allows for higher retention or generalization of a participant's record as opposed to full record suppression. Hence, data utility is optimized while still achieving the targeted ROR.

Consider Participants A-G shown below and their indirect identifiers – age, sex, race and adverse events. The combination of indirect identifiers of **Participants 3 and 4** are unique, thus the two participants will be considered outliers. With **full record suppression**, all four identifiers would have been suppressed. Suppression is limited to race and adverse event only when applying the targeted technique and these two participants cannot be distinguished from each other as they comprise their own equivalence class.

PARTICIPANT	ORIGINAL DATA	FULL RECORD SUPPRESSION	TARGETED SUPPRESSION
1	23 F White Abdominal Pain	20-25 F White Abdominal Pain	20-25 F White Abdominal Pain
2	23 F White Abdominal Pain	20-25 F White Abdominal Pain	20-25 F White Abdominal Pain
3	23 M Asian [Myocardial Infarction]	[Age] [Sex] [Race] [Adverse Event]	20-25 M [Race][Adverse Event]
4	23 M African American [Atrial Fibrillation]	[Age] [Sex] [Race] [Adverse Event]	20-25 M [Race][Adverse Event]
5	35 M Asian Abdominal Pain	35-40 M Asian Abdominal Pain	35-40 M Asian Abdominal Pain
6	35 M Asian Abdominal Pain	35-40 M Asian Abdominal Pain	35-40 M Asian Abdominal Pain
7	35 M Asian Abdominal Pain	35-40 M Asian Abdominal Pain	35-40 M Asian Abdominal Pain

Results Comparison			
	Prior to Targeted Suppression	After Targeted Suppression	Description
Target Risk of Re-Identification	<9%	<9%	
Full Record Suppression	22 Subjects	1 Subject	Significant improvement in full record suppression count resulting from targeted suppression logic
Actual ROR	7.48%	6.16%	ROR is reduced as a byproduct of targeted suppression
Information Loss	36.70%	30.90%	
Granularity	65.60%	70.05%	



The CTR has taken transparency and disclosure of in-process and completed trials to another level.

The resulting impact to study teams and risks associated with disclosure of subject data and CCI must be managed.

1

The CTR has imposed specific requirements for public and non-public version documents

2

Robust and purpose-built tools for identifying and anonymizing personal data are critical to avoid trial delays

3

Advanced quantitative risk assessment techniques will protect ROR while optimizing clinical utility

Questions?



Real Life Sciences, LLC

565 E. Swedesford Road, Suite 205
Wayne, PA 19087

rlsciences.com
+1-484-537-5700
inquiry@rlsciences.com