



PHUSE 2023

Beyond k-anonymity for Clinical Trial Data anonymization

Presenter Name: Parveen Kumar
Email: parveen.kl2@geninvo.com

www.geninvo.com
North America | UK | Europe | Asia



AGENDA

- Updates of Recent Initiatives
- Categorization of Clinical Data
- Anonymization
- Risk Approaches
 - K-anonymity
 - Challenges with K-anonymity
 - L-diversity
 - Challenges with L-diversity
 - T-closeness
 - Challenges with T-closeness
- Points to Consider



www.geninvo.com
North America | UK | Europe | India



Recent Updates on DT Initiatives



Government of Canada
Gouvernement du Canada

MENU

Canada.ca

> Health

> Drug and health products

> Licensing, authorizing and manufacturing drug and health products

> Drug and health product review and approval

> Guidance document on Public Release of Clinical Information: profile page

Public Release of Clinical Information: guidance document

Foreword

Guidance documents are meant to provide assistance to industry and health care professionals on how to comply with governing statutes and regulations. Guidance documents also provide assistance to staff on how Health Canada mandates and objectives should be implemented in a manner that is fair, consistent and effective.



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

21 March 2019
EMA/144064/2019

European Medicines Agency policy on publication of clinical data for medicinal products for human use

POLICY/0070
Status: Adopted
Effective date: 21 March 2019
Review date: No later than March 2021
Supersedes: Policy/0070, dated 2 October 2014 (EMA/240810/2013)



www.geninvo.com
North America | UK | Europe | India



U.S. FOOD & DRUG
ADMINISTRATION

Home

/ Regulatory Information

/ Search for FDA Guidance Documents

/ Civil Money Penalties Relating to the ClinicalTrials.gov Data Bank

GUIDANCE DOCUMENT

Civil Money Penalties Relating to the ClinicalTrials.gov Data Bank

Guidance for Responsible Parties, Submitters of Certain Applications and Submissions to FDA, and FDA Staff

AUGUST 2020

Download the Final Guidance Document

Read the Federal Register Notice

Final



Clinical Trials

English EN | CT

About

Search clinical trials and reports

CTIS for sponsors

CTIS for authorities

Support



All initial applications must be submitted through CTIS from 31/01/23



Clinical trials in the European Union

This website supports the undertaking and oversight of clinical trials in the European Union (EU) and European Economic Area (EEA).

It is part of a broad initiative to transform the EU/EEA clinical trials environment in support of large clinical trials in multiple European countries, to the benefit of medical innovation and patients.



Categorization of Clinical Data

Quasi-Identifiers	Direct Identifiers:	Sensitive information
Demographic: Race, Ethnicity, Date of Birth, Age, Sex, Country Longitudinal: Adverse Events (serious): Death, Hospitalization;Start/End of AE – SITEID	USUBJID, INVNAME	Substance abuse, mental health disorders, HIV, reproductive health, sexually transmitted diseases, and criminal activity) – MHTERM (text) – medical history
Risk: This information shared may lead to identification of subjects who participated in study which is against EU legislations. Personal data must be protected. Remedy: Anonymisation of clinical data		

Data Considered for Sharing

Age	Gender	Site	Medical History
18	Male	BL, Illinois	Diabetes
20	Female	BL, Illinois	Influenza
36	Female	BL, Illinois	Cancer
44	Male	BL, Illinois	Diabetes

Vote Registration Records from BL, Illinois

Birthdate	Gender	Name
2/2/2002	Female	Alice Smith
3/3/1986	Female	Bob Jones
4/4/1978	Male	Charlie Doe

Assessment of Anonymization

Two options to establish if dataset is anonymised:

Demonstrated that after anonymization, it is no longer possible to:

- 1.Singling out: Is it still possible to single out an individual?
- 2.Linkability: Is it still possible to link records relating to an individual and identify that individual?
- 3.Inference: Can information be inferred concerning an individual?

Note: It is clear that it is very difficult to meet the above criteria and maintain data utility at the same time. Only possibility is that if CSRs do not contain any information on direct and quasi-identifiers then above criteria can be met.

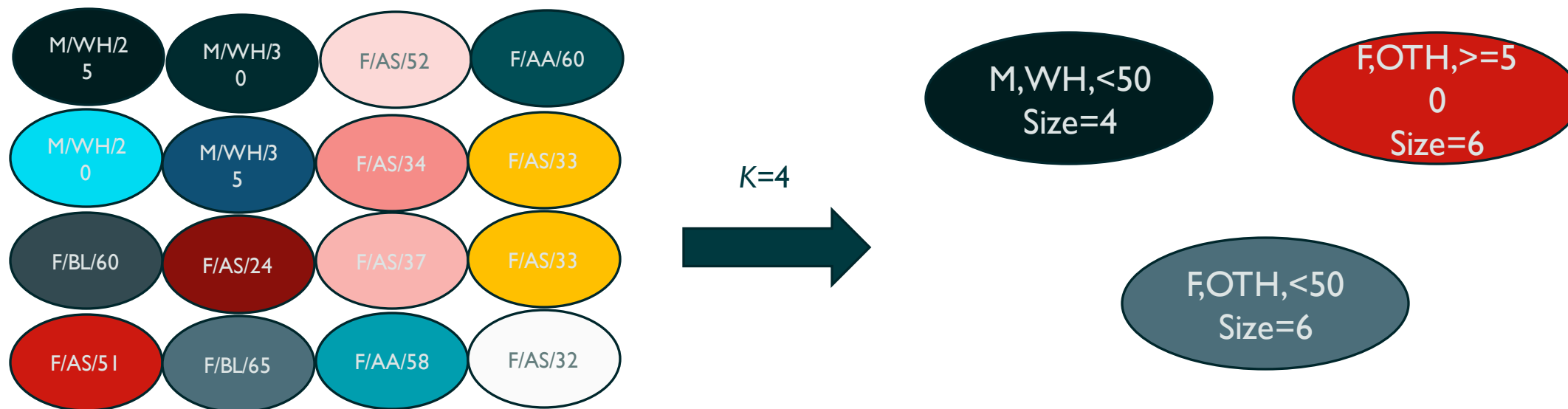
OR

Perform an analysis of re-identification risk. Measuring the risk of re-identification involves:

- 1.An appropriate risk metric
- 2.A suitable threshold
- 3.Actual measurement of risk

K-anonymity

- Common methods for achieving k -anonymity for some value of k
- Suppression
- Generalization



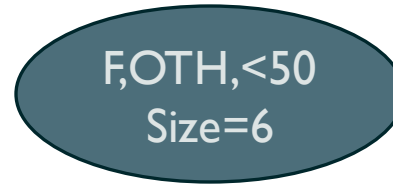
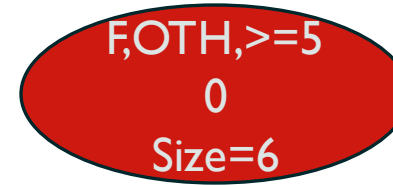
K-anonymity - Issues

- **Homogeneity Attack** - all the values for a sensitive value within a set of k records are identical
- **Background Knowledge Attack** - leverages an association between one or more quasi-identifier attributes with the sensitive attribute to reduce the set of possible values for the sensitive attribute

Hypertension, Suicide Attempt,
Heart Attack



Hypertension, Suicide Attempt,
Cardiac Arrest



Heart Attack

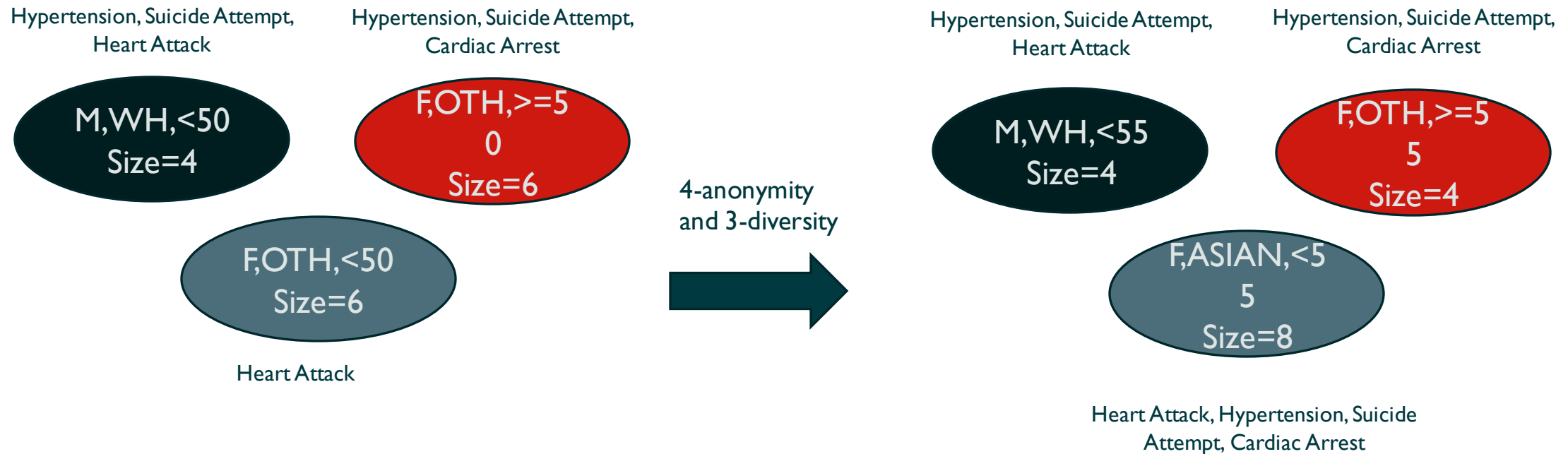
L-diversity

- **Distinct ℓ -diversity** – The simplest definition ensures that at least ℓ distinct values for the sensitive field in each equivalence class exist.
- **Calculation of Distinct L-Diversity:** After the calculating the k-anonymity, we get equivalence classes. For each equivalence class we get N distinct sensitive attributes S , where $N \subseteq S$, $N \geq 1$. For equivalence classes;
 - class 1, ℓ_1 -diversity= n_1 ,
 - class 2, ℓ_2 -diversity= n_2 ,
 - class 3, ℓ_3 -diversity= n_3 , and so on.
- **ℓ -diversity**= $\min (n_1, n_2, n_3, \dots)$.

L-diversity

Updating the Generalization

- Age generalized to <55 , ≥ 55 years
- Race generalized to White, Asian and Other categories



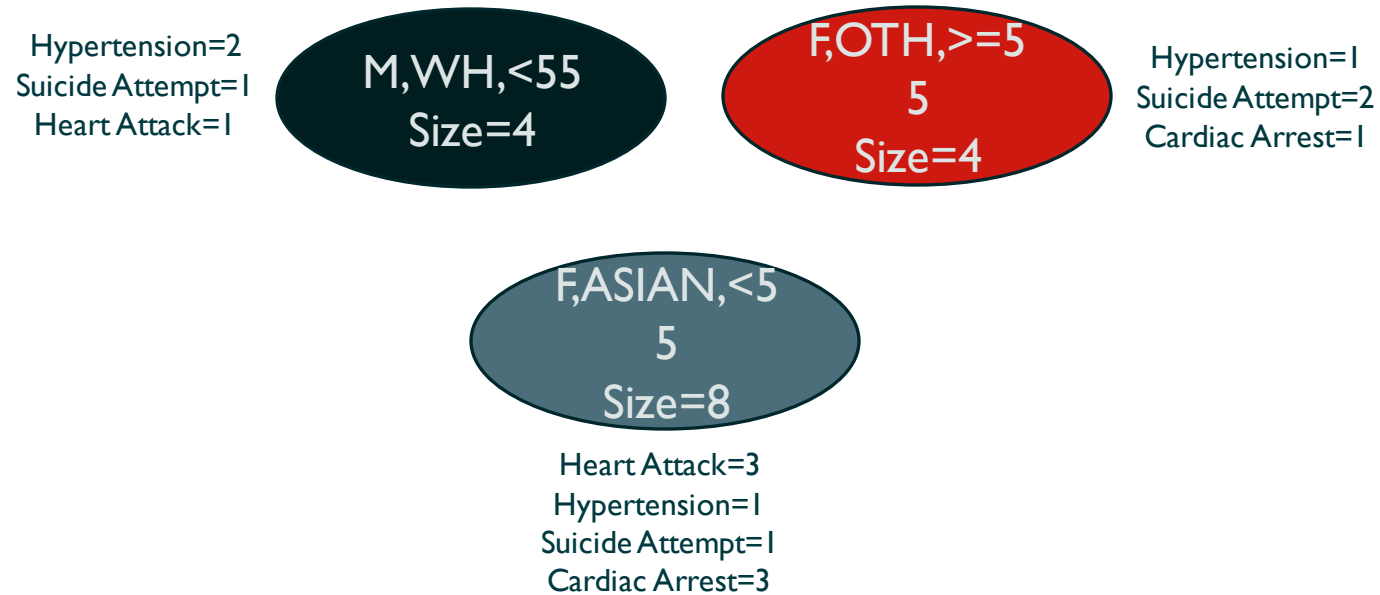
L-diversity

- L-diversity isn't a definition that definitely addresses a particular threat model. Rather, it's a "fix" for one of k-anonymity's most obvious flaws
- L-diversity does not even require the data publisher to have as much information as the adversary
- Parameter L protects against more knowledgeable adversaries
- Larger the value of L, the more information is needed to rule out possible values of the sensitive attribute
- L-diversity may be difficult or unnecessary to achieve:
 - Single Sensitive Attribute with single value
 - Attribute categories are skewed with low frequency for few categories



Need for t-closeness

- *Skewness attack*
- *Similarity attack*



T-closeness

- A betterment of l-diversity is a t-closeness technique by decreasing the granularity of the interpreted data.
- Reduces the correlation between the quasi-identifier attributes and the sensitive attributes
- The distance between the distributions is measured using **Earth Mover's Distance (EMD)**
- It interrupts attribute disclosure that protects data privacy
- Protects against homogeneity and background knowledge attacks mentioned in k-anonymity
- It identifies the semantic closeness of attributes, a limitation of l-diversity

T-closeness

Definitions and Objective

Q: distribution of sensitive attribute value in the whole table

P: distribution of sensitive attribute value in specific equivalence class

B0, B1, B2: Adversary belief on individual's sensitive attribute before looking at published data, after looking at full data and after looking at equivalence classes of data, respectively

Objective of t-closeness is to limit the gain of information by adversary from **B1** to **B2** by limiting the distance between **P** and **Q**

if $P = Q$, then **B1** and **B2** should be the same.

If **P** and **Q** are close, then **B1** and **B2** should be close as well, even if **B0** may be very different from both **B1** and **B2**.

T-closeness

Earth Mover's Distance (EMD)

For numerical sensitive attribute:

$$E(\mathbf{P}, \mathbf{Q}) = \frac{1}{m-1} \sum_{i=1}^m \left| \sum_{j=1}^i (p_j - q_j) \right|$$

For Categorical sensitive attribute:

$$E'(\mathbf{P}, \mathbf{Q}) = \frac{1}{2} \sum_{i=1}^m |p_i - q_i|$$

Where p and q are probabilities for values in distribution P and Q

Cons of T-closeness

- Sometimes, it may not be possible to achieve ϵ -diversity and t -closeness for recommended values at all
- Loss in data utility to achieve the recommended value of $t=0.2$
- Using Earth Mover's Distance (EMD) measure in t -closeness, it is hard to identify the closeness between t -value and the knowledge gained
- Necessitates that sensitive attribute spread in the equivalence class to be close to that in the overall table

Points to Consider

- Individual Patient Data (IPD) will be used at least for following sensitive information
 - Medical History
 - Adverse Events
- There can be multiple sensitive attributes for same patient
- Order of event occurrence could be of importance to consider
- Non sensitive events can be removed



www.geninvo.com
North America | UK | Europe | India



Ideas
Delivered
Intuitively

Thank You!

+1-833-GEN-INVO (436-4686) | info@geninvo.com | www.geninvo.com

