

Applying *k*-anonymity and *l*-diversity in OMOP CDM databases

João Almeida

INTRO:

- The official OMOP CDM guidelines can prevent the re-identification of patients stored in the databases.
- However, literature has shown that procedures that omitted key identifiers are not robust anonymisation procedures.

METHODS

- k*-Anonymity limits the information released, based on generalisation and suppression of data concepts, as well as the number of repetitive elements.
- l*-Diversity technique was proposed aiming to fill some gaps of the *k*-anonymity model.
- These techniques can be applied to the OMOP CDM schema requiring the characterisation of each field.
- When sharing a sample of the database, these techniques increase the anonymization levels.

RESULTS

- We only considered some tables of the OMOP CDM.
- Tables in the groups “Standardized vocabularies”, “Standardized health economics”, “Standardized derived elements” and “Standardized metadata” were omitted.
- Key attributes, quasi-identifiers and sensitive attributes were mapped to apply the privacy-preserving techniques.

Sharing aggregate data? Robust anonymization through privacy-preserving techniques.



Scan QR to download the full paper

Name	Day of Birthday	Sex	Zip Code	Health Condition
João	01/11/1989	Male	3810-015	Diabetes
Maria	04/05/1979	Female	3810-012	Obesity
José	11/02/1968	Male	3810-018	Hypertension
Marta	23/07/1988	Female	3810-027	HIV
Jorge	30/01/1987	Male	3810-002	Diabetes
Márcia	21/08/1990	Female	3810-024	Hypertension
Júlio	17/03/1972	Male	3810-027	Chest pain

Key Attribute | Quasi-identifier | Sensitive Attribute

The key attributes were defined as the unique identifiers directly associated with a patient. We considered all fields in the format “*_source_value” as sensitive attributes. All of the remaining fields were considered quasi-identifiers because these represent the equivalence classes in the view.

Medical Data Released as Anonymous

SSN	Name	Age	Sex	Zip Code	Health Condition
		30	Male	3810	Diabetes
		40	Female	3810	Obesity
		50	Male	3810	Hypertension
		30	Female	3810	HIV
		30	Male	3810	Diabetes
		30	Female	3810	Hypertension
		50	Male	3810	Chest pain

Annotations: 2-anonymity with 1-diversity (left), 2-anonymity with 2-diversity (right)

Medical Data Released as Anonymous

SSN	Name	Date of Birth	Sex	Zip Code	Marital Status	Health Condition
		25/10/1993	Male	3750-403	Single	Alcoholism
		02/11/1991	Female	3750-402	Married	Obesity
		05/03/1995	Female	3810-193	Married	Hypertension
		15/07/1985	Female	3810-193	Divorced	Diabetes
		09/01/1989	Male	3810-192	Married	Hypertension
		13/04/1992	Male	3810-192	Widow	Chest pain
		22/03/1994	Female	3810-195	Married	Shortness of breath

Voter List

Name	Address	City	Zip Code	Date of Birth	Sex	...
João Almeida	Rua dos Barros	Águeda	3750-403	25/10/1993	Male	...

João Rafael Almeida and José Luís Oliveira