Grant Agreement No.: 101080718
Call: HORIZON-HLTH-2022-STAYHLTH-01-two-stage
Topic: HORIZON-HLTH-2022-STAYHLTH-01-05-two-stage
Type of action: HORIZON-RIA

BIO-STREAMS

D4.2 Anonymisation & Pseudonymisation Algorithms

Revision: v.1.0

Work package	WP 4
Task	Task 4.1
Due date	31/10/2025
Submission date	31/10/2025
Deliverable lead	UOI
Version	1.0
Authors	Orestis Papagiannopoulos (UOI) Eleni Georga (UOI) Dimitrios Fotiadis (UOI)
Reviewers	Billy Langlet (KI) Anastasios Gogos (INTRA) Marios Logothetis (INTRA)

Abstract	BIO-STREAMS clinical data are pseudonymised at source by the participating healthcare organisations and do not contain direct identifiers when transferred to the consortium. D4.2 provides an operational Data Anonymisation Service that evaluates residual disclosure risks in harmonised pilot datasets. The service implements *k-anonymity*, *ℓ-diversity*, and *t-closeness* to quantify the likelihood of re-identification and sensitive-attribute inference under configurable quasi-identifiers and threshold parameters. It generates structured dataset-level and record-level indicators, enabling Data Controllers to assess whether additional anonymisation steps are required before analytical processing. The implementation is ready for integration into the BIO-STREAMS Node Bundles to support privacy-focused quality control within the project's data-processing workflow.
Keywords	Anonymisation, pseudonymisation, *k-anonymity*, *ℓ-diversity*, *t-closeness*, privacy risk assessment, clinical data, GDPR compliance.

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
V0.1	01/10/2025	Initial ToC	Eleni Georga
V0.2-V0.3	24/10/2025	Working versions	All authors
V0.4	28/10/2025	Version ready for internal review	All authors
V0.5	30/10/2025	Version integrating reviewers' feedback	All authors
V1.0	31/10/2025	Final version	All authors

Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the other granting authorities. Neither the European Union nor the granting authority can be held responsible for them.

Copyright notice

Project co-funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	DEM	
Dissemination Level		
PU	*Public, fully open, e.g. web*	**x**
SEN	*Sensitive, limited under the conditions of the Grant Agreement*	

Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* *R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

Executive summary

BIO-STREAMS manages harmonised clinical datasets that are pseudonymised at source by Data Controllers. To identify and reduce residual disclosure risks, this deliverable provides an operational Data Anonymisation Service encompassing three established anonymisation models, i.e., *k-anonymity*, *ℓ-diversity* and *t-closeness*, which respectively assess:

- indistinguishability within quasi-identifier groups (*k-anonymity*),
- resistance to sensitive-attribute inference (*ℓ-diversity*), and
- similarity of sensitive-attribute distributions to the global distribution (*t-closeness*).

The service accepts user-defined quasi-identifiers, binning rules and privacy thresholds, and produces structured outputs including dataset-level summaries and per-record flags. These outputs enable Data Controllers to evaluate whether additional data transformation is required prior to data utilisation. The current implementation has been validated using Clinical Data Interchange Standards Consortium (CDISC)-harmonised datasets from BIO-STREAMS Study 1. The Data Anonymisation Service is provided to authorised technical partners for integration with the BIO-STREAMS Node Bundles, where it will function as part of the data-processing workflow alongside curation and harmonisation components.

Co-funded by the European Union

Table of contents

List of figures

List of tables

Abbreviations

CDISC	Clinical Data Interchange Standards Consortium
GDPR	General Data Protection Regulation
QI	Quasi-Identifier

1 Introduction

1.1 Anonymisation Methods Applied in BIO-STREAMS

The increasing availability of clinical and biomedical datasets continues to support advances in personalised medicine, disease modelling and data-driven health research. Within BIO-STREAMS, pseudonymisation is performed entirely by the Data Controllers before data is transferred to the consortium, and direct identifiers are not accessible within the project. However, combinations of quasi-identifiers (QIs) such as age, sex and geographic information may still enable linkage with external sources, introducing residual disclosure risks. Anonymisation methods therefore aim to restrict opportunities for re-identification and sensitive-attribute inference while retaining analytic utility. Formal anonymisation models establish measurable guarantees by defining acceptable levels of indistinguishability and uncertainty.

BIO-STREAMS adopts three well-established anonymisation models that address complementary aspects of disclosure risk. *k-anonymity* [1] requires each record to be indistinguishable from at least $k-1$ others that share the same QI values. By organising records into equivalence classes and enforcing minimum group size, the maximum probability of re-identification is limited to $1/k$. However, *k-anonymity* does not prevent attribute disclosure when sensitive values within a group are homogeneous or strongly correlated with QIs. To address this, ℓ-diversity [2] introduces a requirement for at least ℓ distinct sensitive-attribute values within each equivalence class, reducing the likelihood of accurately inferring attribute information even when the group is identified. Nonetheless, ℓ-diversity may offer limited protection when sensitive-attribute distributions are substantially skewed, allowing adversaries to derive high-confidence inferences despite nominal diversity. *t-closeness* [3] further strengthens protection by constraining the distribution of sensitive values within each equivalence class to remain within a specified threshold t of the global distribution. Using the Wasserstein distance for continuous variables and Total Variation Distance for categorical variables, this model restricts semantic disclosure by limiting distributional divergence. Its effectiveness, however, depends on threshold selection and robust estimation of the overall distribution.

In BIO-STREAMS, these three models are integrated into the BIO-STREAMS Data Anonymisation Service, an operational software service evaluating Clinical Data Interchange Standards Consortium (CDISC)-standardised multi-domain datasets collected retrospectively in Study 1. The service performs structured analyses, forming QI-based equivalence classes, calculating sensitive-attribute characteristics, and quantifying both identity- and attribute-based disclosure risks. It produces dataset-level statistics and record-level flags, enabling Data Controllers to determine whether additional anonymisation is needed before the data are shared for analysis, thereby supporting GDPR-aligned, privacy-preserving processing throughout the project. This document constitutes the accompanying report for D4.2, whose primary output is the operational Data Anonymisation Service (DEM).

2 Methods

2.1 Overview of the Data Anonymisation Service

To quantify residual re-identification and attribute-disclosure risk in harmonised clinical datasets, the BIO-STREAMS operational Data Anonymisation Service implements three complementary anonymisation models (*k-anonymity*, *ℓ-diversity*, and *t-closeness*). These models evaluate disclosure risk under user-defined privacy parameters. Each model is applied independently to a pilot-specific dataset that integrates relevant CDISC domains into a single harmonised structure. During execution, the service assesses the dataset using configurable QIs, sensitive-attribute definitions and threshold values, and generates structured outputs consisting of dataset-level summaries and record-level flags that identify potentially disclosive equivalence classes. These outputs identify areas where privacy-preserving modification of the data (e.g., aggregation, suppression, generalisation) may be necessary before analytical processing, thereby supporting proactive privacy protection within the BIO-STREAMS workflow.

2.2 Modularity and User-Specified Configuration

A key feature of the BIO-STREAMS Data Anonymisation Service is its modular design, which separates configuration, processing and reporting into distinct units. Users define a central configuration structure specifying all parameters required for disclosure-risk evaluation. These parameters include the selection of QIs, the designation of a sensitive attribute, and the definition of binning rules and privacy thresholds. Table 1 summarises the configurable elements of the service.

Table 1: Configuration Parameters of the Data Anonymisation Service

Parameter	Description
Quasi-identifiers (QIs)	Defined as categorical (e.g., sex, region) or continuous variables (e.g., age, body mass index).
Sensitive Attribute	A clinical variable (e.g., a laboratory measurement) used to assess disclosure risk under *ℓ-diversity* and *t-closeness*.
Binning Schemes	Optional binning rules applied to continuous QIs, and when required to the sensitive attribute for *ℓ-diversity* evaluation.
Privacy Thresholds	User-defined numeric criteria for each anonymisation model: • K: minimum equivalence-class size for *k-anonymity* • L: minimum number of distinct sensitive values for *ℓ-diversity* • T: maximum permitted deviation from the global sensitive-value distribution for *t-closeness*

2.3 Implementation of *k*-Anonymity

The *k-anonymity* evaluation quantifies the minimum group size associated with each unique combination of QIs. In this implementation, records sharing identical QI values form an equivalence class with size k. The BIO-STREAMS Data Anonymisation Service applies the following procedure:

1. **Equivalence-class construction:** Records are grouped according to all specified QIs, applying binning to continuous variables if configured.

2. **Group-size calculation:** The frequency of each QI-defined equivalence class is computed, generating a per-record field (k_count).

3. **Risk identification:** Records belonging to equivalence classes with $k\_count < K$ are flagged as at risk of re-identification.

The *k-anonymity* threshold K is a user-defined parameter that specifies the minimum acceptable anonymity level (e.g., $K = 5$ requires each equivalence class to contain at least five records). The service generates a structured output listing all records that do not meet this criterion, along with their QI values and computed k_count values (Figure 1).

	A	B	C	D	E	F	G
1	USUBJID	DM__SEX	SC__Country of Permanent Address	SC__District of Permanent Address	DM__AGE	VS__BMI	k_count
2	[redacted]	M	Greece	Rural	6--12	20--30	1
3	[redacted]	M	Greece	Rural	12--18	20--30	2
4	[redacted]	F	Greece	Rural	12--18	20--30	2
5	[redacted]	F	Greece	Rural	6--12	0--20	2
6	[redacted]	M	Greece	Rural	12--18	30--40	2
7	[redacted]	F	Greece	Urban	6--12	30--40	4
8	[redacted]	M	Greece	Urban	12--18	0--20	4
9	[redacted]	F	Greece	Urban	6--12	30--40	4
10	[redacted]	F	Greece	Urban	0--6	0--20	2
11	[redacted]	F	Greece	Urban	6--12	30--40	4
12	[redacted]	F	Greece	Rural	6--12	0--20	2
13	[redacted]	F	Greece	Urban	0--6	0--20	2
14	[redacted]	M	Greece	Urban	12--18	0--20	4
15	[redacted]	F	Greece	Urban	6--12	30--40	4
16	[redacted]	M	Greece	Urban	12--18	0--20	4
17	[redacted]	M	Greece	Urban	12--18	0--20	4
18	[redacted]	M	Greece	Rural	12--18	20--30	2
19	[redacted]	F	Greece	Rural	12--18	20--30	2
20	[redacted]	M	Greece	Rural	12--18	30--40	2

Figure 1: k-anonymity evaluation output for the PENTELI pilot. Each row represents an individual record with the specified QIs (e.g., sex, country and district of residence, age group, and body mass index range). The final column reports the equivalence-class size (k_count). Records with k_count values below the configured threshold ($K = 5$) are flagged as at risk of re-identification.

2.4 Implementation of *ℓ*-Diversity

The *ℓ-diversity* evaluation addresses the risk of sensitive-attribute inference within QI-defined groups. While *k-anonymity* increases indistinguishability, it does not prevent an adversary from accurately inferring sensitive information when an equivalence class contains homogeneous sensitive values. *ℓ-diversity* extends this protection by requiring that each equivalence class includes at least ℓ distinct values of the designated sensitive attribute.

The BIO-STREAMS Data Anonymisation Service applies the following procedure:

1. **Sensitive-attribute assignment:** The selected sensitive attribute is included for each record, with optional discretisation applied for continuous variables using a user-defined binning scheme (e.g., cholesterol values grouped into specified ranges).

2. **Diversity calculation:** For each equivalence class, the number of distinct sensitive values is computed, generating a per-record field (ℓ_count).

3. **Risk identification:** Records belonging to equivalence classes with $\ell\_count < L$ are flagged as being at risk of attribute disclosure.

The diversity threshold L is user-defined, allowing configuration based on dataset characteristics and privacy requirements across BIO-STREAMS pilots. Discretisation is optional but may improve the robustness of *ℓ-diversity* outcomes for continuous measurements. Figure 2 illustrates a subset of the equivalence-class records and their sensitive attribute values for the PENTELI pilot, indicating those with ℓ_count values below the configured threshold, e.g., $L = 5$.

	A	B	C	D	E	F	G	H
1	USUBJID	DM__SEX	SC__Country of Permanent Address	SC__District of Permanent Address	DM__AGE	VS__BMI	LB__Cholesterol	l_count
2	[redacted]	M	Greece	Urban	6--12	0--20	150--200	1
3	[redacted]	M	Greece	Urban	6--12	0--20		1
4	[redacted]	M	Greece	Urban	12--18	30--40	150--200	1
5	[redacted]	M	Greece	Urban	12--18	30--40	150--200	1
6	[redacted]	F	Greece	Urban	0--6	0--20		1
7	[redacted]	M	Greece	Urban	12--18	30--40		1
8	[redacted]	F	Greece	Urban	12--18	20--30		1
9	[redacted]	M	Greece	Urban	12--18	30--40	150--200	1
10	[redacted]	F	Greece	Urban	12--18	0--20	100--150	1
77	[redacted]	F	Greece	Urban	12--18	30--40	100--150	3
78	[redacted]	M	Greece	Urban	6--12	30--40	200--250	3
79	[redacted]	F	Greece	Urban	6--12	0--20		3
80	[redacted]	F	Greece	Urban	6--12	0--20		3
81	[redacted]	F	Greece	Urban	6--12	0--20	200--250	3
82	[redacted]	M	Greece	Urban	6--12	30--40	100--150	3
83	[redacted]	F	Greece	Urban	12--18	30--40	150--200	3
84	[redacted]	F	Greece	Urban	12--18	30--40	200--250	3
85	[redacted]	F	Greece	Urban	6--12	0--20	100--150	3
86	[redacted]	F	Greece	Urban	12--18	30--40	150--200	3
87	[redacted]	F	Greece	Urban	6--12	0--20		3

Figure 2: ℓ-diversity evaluation output for the PENTELI pilot using cholesterol as the sensitive attribute. Records are grouped into equivalence classes defined by sex, country and district of residence, age group, and body mass index range. The final column reports the number of distinct sensitive-attribute values (ℓ_count). Records with ℓ_count values below the configured threshold ($L = 5$) are flagged as being at risk of attribute disclosure.

2.5 Implementation of *t*-Closeness

While *ℓ-diversity* improves protection against attribute inference, it does not control the semantic similarity of sensitive-attribute values within an equivalence class. *t-closeness* addresses this limitation by ensuring that the distribution of the sensitive attribute in each equivalence class does not differ markedly from the global distribution, thereby restricting semantic disclosure. The BIO-STREAMS Data Anonymisation Service applies the following procedure:

1. **Sensitive-attribute assignment:** The selected sensitive attribute is included for each record without discretisation.

2. **Global distribution estimation:** The full-dataset distribution of the sensitive attribute is estimated (continuous distributions for numerical variables and normalised frequency distributions for categorical variables).

3. **Group-wise divergence measurement:** For each equivalence class, the sensitive-attribute distribution is compared with the global distribution. The Wasserstein distance is used for numerical variables and Total Variation Distance for categorical variables, producing a per-record distance metric ($t\_distance$).

4. **Risk identification:** Records belonging to equivalence classes where $t\_distance$ exceeds the threshold T are flagged as being at risk of attribute disclosure.

The threshold T is user-defined and reflects the maximum acceptable divergence between local and global distributions, allowing flexibility in privacy requirements across datasets. Figure 3 presents the results for the PENTELI pilot, showing data records whose sensitive attribute exhibits a $t\_distance$ greater than the configured threshold (e.g., $T = 30$).

	A	B	C	D	E	F	G	H
1	USUBJID	DM__SEX	SC__Country of Permanent Address	SC__District of Permanent Address	DM__AGE	VS__BMI	LB__Cholesterol	t_distance
2	[redacted]	F	Greece	Urban	0--6	0--20	202	52.15
3	[redacted]	M	Greece	Urban	6--12	0--20	168	40.56
4	[redacted]	M	Greece	Urban	6--12	0--20	24	40.56
5	[redacted]	M	Greece	Urban	6--12	0--20	158	40.56
6	[redacted]	M	Greece	Rural	6--12	20--30	185	38.74
7	[redacted]	M	Greece	Urban	12--18	30--40	187	33.02
8	[redacted]	M	Greece	Urban	12--18	30--40	173	33.02
9	[redacted]	M	Greece	Urban	12--18	30--40	189	33.02
10	[redacted]	M	Greece	Urban	12--18	30--40	173	33.02
11	[redacted]	M	Greece	Rural	12--18	30--40	160	30.44
12	[redacted]	M	Greece	Rural	12--18	30--40	200	30.44

Figure 3: t-closeness evaluation output for the PENTELI pilot using cholesterol as the sensitive attribute. Records are grouped into equivalence classes defined by sex, country and district of residence, age group and body mass index (BMI) range. The final column reports the distributional deviation ($t\_distance$) from the global cholesterol distribution. Records with $t\_distance$ values greater than the configured threshold ($T = 30$) are flagged as being at risk of attribute disclosure.

3 Results

3.1 Automated Privacy Risk Reporting

To support a standardised interpretation of disclosure-risk results across pilots, the BIO-STREAMS Data Anonymisation Service includes an automated reporting function that generates a structured privacy summary for each dataset. The report consolidates the outputs of the *k-anonymity*, *ℓ-diversity* and *t-closeness* evaluations and documents:

- the QIs used and their categorisation,
- the selected sensitive attribute,
- any applied binning schemes,
- the count and percentage of records flagged as being at risk, and
- per-metric disclosure-risk indicators.

Figure 4 shows an example of the automatically generated summary report for the PENTELI pilot. The report indicates the configured input parameters and displays the proportion of records identified as being at risk of re-identification or attribute disclosure under each anonymisation model. The configuration elements included within the report are summarised in Table 2.

```
# 🔒 Privacy Report for PILOT: PENTELI
=============================================================================================================
## 🔎 Quasi-Identifiers Used
- Continuous QIs: DM__AGE, VS__BMI
- Categorical QIs: DM__SEX, SC__Country of Permanent Address, SC__District of Permanent Address

## 🧱 Binning Configuration of continuous QIs
- DM__AGE: [0, 6, 12, 18]
- VS__BMI: [0, 20, 30, 40]

## 🧱 Binning Configuration (Sensitive Feature)
- LB__Cholesterol: [100, 150, 200, 250]

## 📊 K-Anonymity Summary
| Pilot   |   Total |   Risky (k<5) | % Risky   |
|---------|---------|---------------|-----------|
| PENTELI |     633 |            19 | 3.0%      |
- Risky USUBJIDs: Please refer to the saved CSV for full details.

## 📊 ℓ-Diversity Summary
| Pilot   | Sensitive Attribute   |   Total |   Risky (ℓ<5) | % Risky   |
|---------|-----------------------|---------|---------------|-----------|
| PENTELI | LB__Cholesterol       |     633 |            86 | 13.6%     |
- Risky USUBJIDs: Please refer to the saved CSV for full details.

## 📊 t-Closeness Summary
| Pilot   | Sensitive Attribute   |   Total |   Risky (t>30) | % Risky   |
|---------|-----------------------|---------|----------------|-----------|
| PENTELI | LB__Cholesterol       |     633 |             11 | 1.7%      |
- Risky USUBJIDs: Please refer to the saved CSV for full details.
```

Figure 4: Automated privacy risk report generated for the PENTELI pilot. The report presents the configured QIs, sensitive attribute, and associated binning schemes, together with counts and percentages of records flagged as being at risk under k-anonymity, ℓ-diversity and t-closeness evaluations.

Table 2: Summary of Report Output Elements produced by the Data Anonymisation Service

Section	Description
Quasi-Identifier (QI) Summary	• List of categorical and continuous QIs used in the evaluation
Binning Configuration	• Bin edges applied to continuous QIs • Binning of the sensitive attribute when applicable
Risk Metric Summaries	• Total number of analysed records • The count and percentage of records flagged as being at risk
Output Format	• Plain-text file (.txt) generated for documentation and review

3.2 GitHub Repository and Code Availability

To enable technical deployment within BIO-STREAMS, the source code of the Data Anonymisation Service is hosted in the project's GitHub repository (https://github.com/bio-streams-eu-project/BIO-STREAMS-WP4-Anonymization/) (Figure 5). Authorised partners can access the repository to integrate the service into the Node Bundles, configure it for pilot-specific datasets, and contribute to its continued improvement within the project workflow. Further implementation details and configuration guidance, as documented in the repository README, are provided in Appendix A.

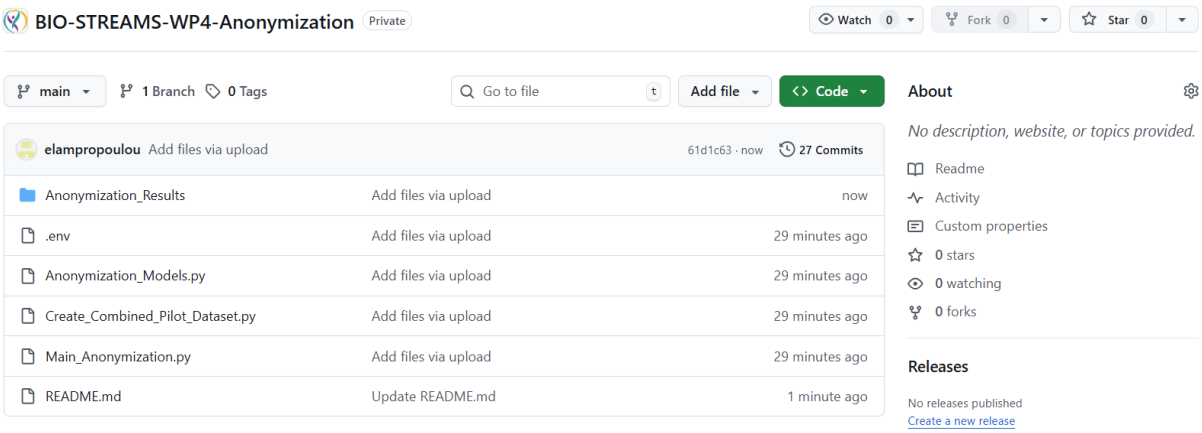

Figure 5: BIO-STREAMS GitHub Organization repository hosting the source code of the Data Anonymisation Service, accessible to authorised BIO-STREAMS technical partners for integration and configuration tasks.

4 Conclusions

This deliverable provides an operational Data Anonymisation Service for quantifying residual disclosure risks in harmonised clinical datasets within BIO-STREAMS. The service implements *k-anonymity*, *ℓ-diversity* and *t-closeness*, generating structured dataset-level and record-level outputs to support GDPR-aligned privacy-preserving data processing. The current implementation has been tested on CDISC-harmonised datasets from BIO-STREAMS Study 1, demonstrating functional readiness across multiple clinical domains. The service has been delivered to technical partners for integration into the BIO-STREAMS Node Bundles. This integration will enable auditable data anonymisation during data handling and analysis. As part of this deployment, the service will be refined and extended to align with the CDISC-based data model of BIO-STREAMS Study 2, ensuring consistency with the harmonisation procedures applied across BIO-STREAMS datasets.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, pp. 3–es, 2007.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and ℓ-diversity", in IEEE 23rd International Conference on Data Engineering, pp. 106–115, 2007.

Appendix A: BIO-STREAMS GitHub Repository – Implementation and Configuration Overview

How to Run the Full Privacy Pipeline

1. First, set the target pilot by editing the `.env` file and updating the `BIOSTREAMS_PILOT` variable (e.g., `PENTELI`).
2. Then, run `Create_Combined_Pilot_Dataset.py`. This will merge all domain-level CSVs under `pilot_report/<PILOT>/` into a single combined dataset. The output will be saved to: `pilot_report/<PILOT>/processed_for_anonym/combined_dataset.csv`.
3. Next, open `Main_Anonymization.py` and define the `config` dictionary. This includes:
   - Quasi-identifiers (categorical and continuous)
   - Binning schemes
   - The sensitive feature (with the domain-specific prefix and double underscore, e.g. LB__Cholesterol)
   - Values for K (k-anonymity), ℓ (l-diversity), and T (t-closeness)
4. Finally, run `Main_Anonymization.py`. This will:
   - Execute all privacy checks
   - Save risky rows to CSV files
   - Generate a summary report called `<PILOT>_privacy_summary.txt`

...

📁 PROJECT STRUCTURE

This project performs **privacy risk analysis** on harmonized, per-pilot datasets using:

- ✅ **k-Anonymity** — checks if each equivalence class contains at least *K* patients
- ✅ **ℓ-Diversity** — checks if sensitive values vary across at least *ℓ* categories within each QI group
- ✅ **t-Closeness** — checks if sensitive value distributions are within distance *T* of the global distribution

Each pilot is expected to have its own folder under `pilot_report/`, containing harmonized, per-domain CSVs.

🌍 `.env` File: Pilot Selection

Before running the pipeline, specify the target pilot in a `.env` file at the project root:

```
BIOSTREAMS_PILOT=PENTELI
```

This value determines:

- Which pilot folder to use inside `pilot_report/`
- Where the domain CSVs will be loaded from
- Where outputs (combined dataset, risky rows, summary report) will be saved

The Python scripts automatically load this variable using `dotenv`.

> ✅ Tip: You can switch between pilots by simply editing `.env` — no code changes needed.

📄 `Create_Combined_Pilot_Dataset.py`

🔧 **Purpose:** Merges all domain-level data for a single pilot into a unified .csv dataset.

Key Features:

- Prefixes all column names with their domain (e.g., `DM__AGE`, `VS__BMI`)
- Handles special ID logic for the `PENTELI` pilot
- Joins all domain CSVs by patient index (`USUBJID`)
- Drops predefined non-informative columns

Output:

```
pilot_report/<PILOT>/processed_for_anonym/combined_dataset.csv
```

`Anonymization_Models.py`

🧠 **Purpose:** Core logic module containing all privacy metric and utility functions.

Main Functions

Function	Description
`run_k_anonymity()`	Flags records belonging to equivalence classes with fewer than **K** members.
`run_l_diversity()`	Flags QI groups where the sensitive attribute has fewer than ℓ distinct values.
`run_t_closeness()`	Flags QI groups where the sensitive value distribution diverges by more than **T** from the global distribution.
`generate_full_privacy_report()`	Compiles all summaries and parameters into a single `.txt` report.

Helper Functions

Function	Description
`validate_config()`	Ensures the configuration dictionary contains all required keys and valid data types.
`bin_feature_column()`	Converts continuous values into binned intervals (e.g., `0-6`, `6-12`, `12-18`).
`apply_feature_binning()`	Applies user-defined binning rules to multiple DataFrame columns simultaneously.
`categorical_distance()`	Computes the total variation distance between two categorical distributions.
`numerical_distance()`	Computes Wasserstein (Earth Mover's) distance between two numerical distributions.

`Main_Anonymization.py`

🚀 **Purpose:** Orchestrates the full privacy analysis pipeline for a single pilot.

What it does:

- Loads the pilot name from `.env` (`BIOSTREAMS_PILOT`)
- Defines and validates all configuration parameters (QIs, sensitive feature, thresholds, binning)
- Executes:
  - `run_k_anonymity()`
  - `run_l_diversity()`
  - `run_t_closeness()`
  - `generate_full_privacy_report()`
- Produces:
  - Per-check risky row `.csv` files
  - A consolidated `.txt` privacy summary report

Output Example:

```
pilot_report/<PILOT>/processed_for_anonym/
│
├── combined_dataset.csv
├── risky_rows_k5_anonymity.csv
├── risky_rows_l5_<sensitive>.csv
├── risky_rows_t30_<sensitive>.csv
└── <PILOT>_privacy_summary.txt
```

CONFIGURATION

Main_Anonymization.py uses a central `config` dictionary to define quasi-identifiers (QIs), binning schemes, and privacy parameters for **k-anonymity**, **ℓ-diversity**, and **t-closeness**. Below is a breakdown of each field (EXAMPLES):

Quasi-Identifiers

These columns define the "identity group" of each individual:

```
"categorical_qis": ["DM__SEX", "SC__Country of Permanent Address", "SC__District of Permanent Address"],
"continuous_qis": ["DM__AGE", "VS__BMI"],
```

- `categorical_qis`: Treated as-is during group construction.
- `continuous_qis`: Binned using intervals to reduce granularity before grouping.

QI Binning

```
"binning_config": {"DM__AGE": [0, 6, 12, 18], "VS__BMI": [0, 20, 30, 40]}
```

Each continuous QI is binned into discrete intervals using the given cutoffs. If not specified, the variable will be left unbinned.

Sensitive Feature

```
"sensitive_feature": "LB__Cholesterol"
```

This is the column whose distribution is analyzed within each QI group for ℓ-diversity and t-closeness. The naming schema contains the domain-specific prefix and double underscore (as is in the combined dataset .csv).

Sensitive Feature Binning (ℓ-diversity only)

```
"sensitive_binning_config_for_L": {"LB__Cholesterol": [100, 150, 200, 250]}
```

This optional binning is **only used for ℓ-diversity**, to reduce the granularity of the sensitive feature. **t-closeness always uses raw (unbinned) values**.

Privacy Parameters

```
"K": 5,
"L": 5,
"T": 50
```

- **K**: Minimum group size for **k-anonymity**.
- **L**: Minimum number of distinct sensitive values (after binning) for **ℓ-diversity**.
- **T**: Maximum allowed distance from global distribution for **t-closeness**.

K-ANONYMITY

Definition: A dataset satisfies **k-anonymity** if each combination of quasi-identifiers (QIs) appears in at least **k** records. This ensures that each individual is indistinguishable from at least `k - 1` others based on QIs.

Implementation Details

- **QIs used:** Defined in the config (`categorical_qis + continuous_qis`) and optionally binned.
- **Equivalence Classes:** Each row is grouped based on its QI combination. The group size is called `k_count`.
- **Risky Records:** Any row belonging to a group where `k_count < K` is flagged as risky.
- **Output:** A file named:

```
risky_rows_k{K}_anonymity.csv
```

containing:

  - `USUBJID`: Patient identifier
  - QI values
  - `k_count`: Number of records with the same QI combination

Example

If `K = 5` and a group with QIs [Male, Greece, Urban, 12--18, 30--40] has only 3 members, all 3 are flagged.

L-DIVERSITY

Definition: A dataset satisfies **ℓ-diversity** if every group of records sharing the same quasi-identifiers (QI group) contains **at least ℓ "well-represented" values** for the sensitive attribute (SA). This protects against homogeneity attacks, where attackers can infer sensitive values even if `k-anonymity` is satisfied.

Implementation Details

- **Sensitive Feature:** Defined in config as `sensitive_feature`. Can be numeric or categorical.
- **QI Groups:** The dataset is grouped by quasi-identifiers (after binning), and for each group, we count the number of *distinct* sensitive values.
- **Diversity Metric:**
  - We compute the number of *unique* sensitive values in each QI group: `l_count = nunique(sensitive_feature)`
  - Rows in groups with `l_count < L` are considered **risky**.
- **Optional Binning of SA (for L only):** A separate binning config can be provided (`sensitive_binning_config_for_L`) to discretize continuous sensitive features for this diversity check only.
- **Output:** A file named:

```
risky_rows_l{L}_{sensitive_feature}.csv
```

containing:

  - `USUBJID`: Patient identifier
  - QI values
  - Sensitive feature values
  - `l_count`: Number of unique sensitive values in the group

Example

If `L = 5` and a QI group [Male, Greece, Urban, 12--18, 30--40] has only 2 distinct values of (binned) Cholesterol, it is flagged as not ℓ-diverse.

T-CLOSENESS

Definition: A dataset satisfies **t-closeness** if the distribution of a sensitive attribute (SA) within each quasi-identifier (QI) group is **no more than** `t` **distance** away from the global distribution of the SA. This prevents attackers from inferring sensitive information based on subtle distributional shifts.

Implementation Details

- **Sensitive Feature:** Specified in `config["sensitive_feature"]`. It can be **categorical or numerical**.
- **t-Distance Calculation:**
  - For **categorical SAs**: The distribution in each QI group is compared to the global distribution using the **Total Variation Distance** (a.k.a. L1 distance divided by 2).
  - For **numerical SAs**: Uses **Wasserstein Distance** (Earth Mover's Distance) between the global and group-level value distributions.
- **Optional SA Binning:** If `sensitive_binning_config` is provided, the sensitive attribute is discretized *only for this check*.
- **Risky Group Detection:** Any record in a QI group with `t_distance > T` is considered **risky** and saved.
- **Output:** A file named:

```
risky_rows_t{T}_{sensitive_feature}.csv
```

containing:

  - `USUBJID`: Unique subject ID
  - All quasi-identifiers
  - Sensitive feature value
  - `t_distance`: Distance to global SA distribution

Example:

If `T = 50` and a QI group [Male, Greece, Urban, 12--18, 30--40] has a (non-binned) Cholesterol distribution highly skewed from the global (non-binned) Cholesterol distribution, it will be flagged as **t-insecure**.