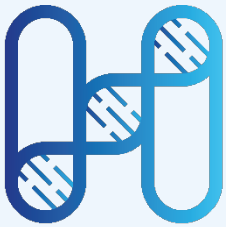


# HEREDITARY

HetERogeneous sEmantic Data integratIon for the guT-bRain interplaY

## Deliverable 3.8

# Privacy-preserving analytics: first release

Version 3.00, 26 June 2026

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

## EXECUTIVE SUMMARY

This deliverable presents the first release of the privacy-preserving analytics workflow developed within the HEREDITARY project. It builds on the *Hereditary Data Network (HDN)* and addresses the need to perform analytics over sensitive medical data while preserving institutional control and reducing disclosure risks. The deliverable revises the privacy setting of the HDN by introducing PRISM, a query-level privacy-risk assessment framework designed in collaboration with WP7 and grounded in the legal and ethical requirements for federated processing of sensitive health data. PRISM enables calibrated and explainable assessment based on the query, ontology annotations, and, where required, data-level statistics.

The deliverable also positions *Privacy-Preserving Synthetic Data Generation (PP-SDG)* as a controlled fallback mechanism for queries and analytics tasks that cannot be safely answered over real data. We focus on tabular synthesis methods and survey privacy metrics for estimating risks associated to synthetic data releases. We organise these metrics in a taxonomy based on their computational structure and threat model, and we operationalise them in PrivEval, an interactive framework supporting data owners in assessing residual disclosure risk and metric applicability.

The experimental evaluation focuses on the tabular ALS data, at the core of many analyses of the clinical partners. Among the PP-SDG methods considered (DPGAN, PATE-GAN, and PrivBayes), PrivBayes provides the strongest observed utility, preserving selected marginal distributions and achieving higher utility than DPGAN and PATE-GAN. At the same time the empirical privacy metrics do not indicate a higher residual disclosure risk for PrivBayes, suggesting that PrivBayes can achieve *Differential Privacy (DP)* without significantly degrading the ALS data. At the same time, we identified limitations of current PP-SDG methods, including scalability constraints, weakened dependencies between attributes, and possible generation of clinically implausible records.

Finally, the deliverable evaluates k-means as a representative downstream federated analytics task. The results show that federated k-means over real ALS data approximates the corresponding centralised clustering with limited distortion, while the main loss of utility is introduced by the synthetic data generation step. The deliverable also introduces a SPARQL-based implementation path for federated k-means, aimed at closer integration with the HDN execution model. Together, these results establish the foundations of the HEREDITARY privacy-preserving analytics pipeline: PRISM provides the governance layer, PP-SDG provides a candidate fallback mechanism, PrivEval provides the empirical assessment layer, and the ALS experiments identify the main limitations to address in subsequent iterations.

## DOCUMENT INFORMATION

| | |
|---|---|
| **Deliverable ID** | **3.8** |
| **Deliverable Title** | **Privacy-preserving analytics: first release** |
| **Work Package** | **WP 3** |
| **Lead Partner** | **Aalborg University** |
| **Due Date** | **30 June 2026** |
| **Date of submission** | **26 June 2026** |
| **Deliverable Type** | **R – Report** |
| **Dissemination level** | **PU – Public** |

## AUTHORS


| Name | Organization |
|---|---|
| Daniele Dell'Aglio | AAU |
| Luca Fabbian | AAU |
| Juan Manuel Rodriguez | AAU |
| Anna Romanovych (internal reviewer) | UNIPD |
| Tobias Schreck (internal reviewer) | TUGRAZ |
| Gianmaria Silvello | UNIPD |
| Frederik Marinus Trudslev | AAU |
| Peter Waldert (internal reviewer) | TUGRAZ |

## REVISION HISTORY

| Version | Date | Author | Description |
|---|---|---|---|
| 0.01 | 14.04.2026 | Daniele | Initial table of content. |
| 0.02 | 30.04.2026 | Daniele | First draft of assessment of sensitive data protection. |
| 0.03 | 11.05.2026 | Luca | Revised and edited section on on privacy assessment. |
| 0.04 | 12.05.2026 | Daniele | Draft of differential privacy, neural network models for PP-SDG, and ALS data description. |
| 0.05 | 12.05.2026 | Frederik | First draft of ALS data preparation for SDG and federated k-means evaluation. |
| 0.06 | 12.05.2026 | Juan | Draft of the ALS data preparation for k-means. |
| 0.07 | 18.05.2026 | Frederik | Extended the description of variables in the ALS data. |
| 0.08 | 18.05.2026 | Daniele | Draft of the introduction. |
| 0.09 | 19.05.2026 | Frederik | Draft of tabular data synthesis with PrivBayes and data preparation for SDG. |
| 0.10 | 19.05.2026 | Gianmaria | Added PRISM section. |
| 0.11 | 20.05.2026 | Luca | First draft of SPARQL-based federated k-means. |
| 0.12 | 20.05.2026 | Frederik | Initial draft of Synthetisation of the HEREDITARY ALS tabular data. |
| 0.13 | 21.05.2026 | Luca | Reviewed and edited PRISM and background sections. |
| 0.14 | 21.05.2026 | Daniele | Introduction of the privacy metrics section and refinement of the deliverable introduction and privacy metric taxonomy. |
| 0.15 | 22.05.2026 | Frederik | Draft of Synthetisation of the HEREDITARY ALS tabular data. |
| 0.16 | 24.05.2026 | Juan | Cleaned and improved data preparation section. |
| 0.17 | 24.05.2026 | Juan | Added first version of section 5.3.4. |
| 0.18 | 26.05.2026 | Frederik | Updated figures and text for Synthetisation of the HEREDITARY ALS tabular data. |
| 0.19 | 27.05.2026 | Frederik | Updated section structure and text regarding Experiments with private synthetic data on federated k-means with Flower compared to centralised k-means. |
| 0.20 | 28.05.2026 | Luca | Draft of the *SPARQL Protocol And Query Language (SPARQL)*-based implementation of k-means. |
| 0.21 | 29.05.2026 | Daniele | Restructuring and editing of Experiment section; draft of the milestone section. |
| 0.22 | 29.05.2026 | Luca | Grammar check and edits. |
| 1.00 | 29.05.2026 | Daniele | Executive summary and conclusions, minor edits. Ready for review |
| 1.01 | 13.06.2026 | Tobias and Peter | Review. |
| 1.02 | 17.06.2026 | Daniele | Addressed reviewers' comments: fixed typos and broken references, removed duplicate citation, added punctuation to the equations, fixed the product and exponent notation in the equations. |
| 1.03 | 18.06.2026 | Frederik | Fixed math notation. |
| 2.00 | 19.06.2026 | Juan and Luca | Grammar checking and typo corrections. Ready for final reviewer check. |
| 2.01 | 22.06.2026 | Tobias | Approved the revision, document ready for finalisation and submission. |
| 2.02 | 23.06.2026 | Anna | Final check of the deliverable; comments shared. |
| 3.00 | 24.06.2026 | Luca | Addressed comments and finalised the deliverable. |

# Contents

## List of Tables



## List of Figures

## List of Acronyms

**AIA** Attribute Inference Attack
**ALS** Amyotrophical Lateral Sclerosis
**AIR** Attribute Inference Risk
**AMI** Adjusted Mutual Information
**ARI** Adjusted Rand Index
**AUC** Area Under the Curve
**AUCROC** *Area Under the Curve (AUC)* of the *Receiver Operating Characteristic (ROC)*
**Auth** Authenticity
**BGP** Basic Graph Pattern
**BN** Bayesian Network
**CQE** Controlled Query Evaluation
**CRP** Common Rows Proportion
**CVP** Close Value Probability
**DAG** Directed Acyclic Graph
**DCR** Distance to Closest Record
**D-MLP** Detection MLP
**DLMLP-AIA** Data Leakage *Multi Layer Perceptron (MLP) Attribute Inference Attack (AIA)*
**DOMIAS** Detecting Overfitting for *Membership Inference Attacks (MIAs)* against Synthetic data
**DOMIAS-MIA** *Detecting Overfitting for MIAs against Synthetic data (DOMIAS)* MIA
**DP** Differential Privacy
**DP-SGD** Differentially Private Stochastic Gradient Descent
**DPIA** Data Protection Impact Assessments
**DPGAN** Differentially-Private *Generative Adversarial Network (GAN)*
**DPO** Data Protection Officer
**DPV-PD** Data Privacy Vocabulary
**DVP** Distant Value Probability
**FM** Fowlkes–Mallows score
**GAN** Generative Adversarial Network
**GCAP** Categorical Generalized CAP
**GDPR** General Data Protection Regulation
**HDN** Hereditary Data Network
**HEREDITARY** HetERogeneous sEmantic Data integratIon for the guT-brAin inteRplaY
**HERO** Hereditary Ontology
**HiddR** Hidden Rate
**HitR** Hitting Rate
**IDS** IDentifiability Score
**MDCR** Median Distance to Closest Record
**MIA** Membership Inference Attack
**MIR** Membership Inference Risk
**MLP** Multi Layer Perceptron
**MRF** Markov Random Field
**NMI** Normalised Mutual Information
**NNAA** Nearest Neighbour Adversarial Accuracy
**NNDR** Nearest Neighbour Distance Ratio
**NSND** Nearest Synthetic Neighbour Distance
**OBDA** Ontology–Based Data Access

**OBDF** Ontology–Based Data Federation
**OWL** Ontology Web Language
**OWL 2 QL** *Ontology Web Language (OWL)* 2 Query Language
**PATE** Private Aggregation of Teacher Ensembles
**PGM** Probabilistic Graphical Model
**PP-SDG** Privacy-Preserving Synthetic Data Generation
**PPOBDA** Policy-Protected *Ontology–Based Data Access (OBDA)*
**PRISM** Privacy Risk Index for SPARQL-based Mediation in OBDA
**ROC** Receiver Operating Characteristic
**SDG** Synthetic Data Generation
**SPARQL** SPARQL Protocol And Query Language
**SRA** Synthetic Ranking Agreement
**VAE** Variational Auto Encoder
**W3C** World Wide Web Consortium
**ZCAP** Categorical Zero CAP

# 1 Introduction

One of the main objectives of HEREDITARY is to enable federated analytics and learning across medical datasets while strictly adhering to data privacy regulations. This objective requires both a regulatory grounding and a technical infrastructure that jointly ensure sensitive clinical data remains under the control of the data holders and that any analytical workflow is continuously assessed and governed with respect to the disclosure risk.

The foundations of data privacy in HEREDITARY have been established in previous deliverables. From a legal and ethical perspective, Deliverables D2.1 [9] and D7.1 [29] establish the ethical principles and the regulatory framework guiding data protection in the processing of sensitive health data, including compliance with *General Data Protection Regulation (GDPR)* and other related European legislation, and outlining the ethical requirements for handling sensitive health data and AI-based methods throughout the project. From a technical perspective, Deliverables D2.11 [43], D2.14 [38], and D2.15 [42] describe the federated learning infrastructure, including secure and scalable computing environments, communication protocols, and privacy-preserving mechanisms such as secure aggregation, differential privacy, and encrypted communication. At the analytics level, Deliverable D3.2 [4] introduced the HDN and the first approach to privacy-aware query processing, based on a classification of queries into predefined risk levels.

This deliverable builds on these foundations and revises the privacy setting of the HDN. The key observation motivating this revision is that a static classification of queries based solely on their syntactic structure is insufficient to capture their actual disclosure risk. Instead, privacy must be assessed at the level of individual queries, taking into account their semantic content, their interaction with the ontology, and, when necessary, the characteristics of the underlying data.

To address these challenges, we introduce *Privacy Risk Index for SPARQL-based Mediation in OBDA (PRISM)*, a framework for calibrated and explainable query-level risk assessment in federated analytics. PRISM enables the HDN to evaluate queries before execution and to enforce privacy controls in a principled manner. At the same time, PRISM defines a boundary: queries whose risk exceeds a policy-defined threshold cannot be safely answered over real data.

To preserve the answerability of such queries, we adopt PP-SDG as a controlled fallback mechanism. Synthetic datasets approximate the statistical properties required to answer a query while reducing the risk of disclosing sensitive information. This introduces a complementary challenge: assessing the privacy guarantees of synthetic datasets themselves in a way that is operational, interpretable, and aligned with practical use cases.

This deliverable addresses these challenges through four main contributions.

Firstly, we revisit and refine the privacy-aware query analysis mechanism introduced in Deliverable D3.2, replacing the static classification of queries with PRISM, a risk assessment framework grounded in ontological semantics and aligned with GDPR requirements.

Secondly, we investigate PP-SDG in the context of HEREDITARY medical data. We analyse state-of-the-art methods for tabular data and evaluate their applicability to the HEREDITARY ALS dataset, highlighting both their strengths and limitations in a real setting.

Thirdly, we introduce a unified framework for the assessment of privacy risks in synthetic data. We organise existing privacy metrics into a coherent taxonomy, based on their underlying computational structure, and reformulate them using a shared mathematical notation. This framework clarifies the relationships between metrics and supports their systematic comparison. We operationalise this contribution in PrivEval, a tool implementing 17 privacy metrics and supporting data owners in evaluating the privacy risk of releasing or analysing a synthetic datasets derived from the real data.

Fourthly, we integrate these components into a privacy-aware analytics workflow that combines federated query processing, synthetic data generation, and analytics tasks. We demonstrate this workflow through a

federated K-means scenario on the HEREDITARY ALS dataset, comparing centralised and federated implementations under privacy constraints.

Taken together, these contributions define a coherent privacy-preserving analytics pipeline for the HDN. Disclosure risk is assessed at the query level through PRISM, mitigated through synthetic data when necessary, and validated through quantitative privacy metrics, enabling advanced analytics to be performed without compromising data protection requirements.

The remainder of this deliverable is organised as follows. Section 2 introduces the revised privacy setting of HDN and presents PRISM. Section 3 reviews methods for PP-SDG. Section 4 presents the framework for privacy assessment in synthetic data, including the taxonomy of metrics and its implementation in PrivEval. Section 5 reports on the experimental evaluation of privacy-preserving mechanisms on the HEREDITARY ALS dataset, including synthetic data generation and federated k-means. Section 6 provides verification of the Milestone 8, and Section Section 7 concludes with remarks and future directions.

# 2 Revising the Privacy Setting of the Hereditary Data Network

Deliverables 3.1 [8] and 3.2 [4] established the ontology layer, *Hereditary Ontology (HERO)* and the OBDA/*Ontology–Based Data Federation (OBDF)* execution stack of the HDN. For privacy governance, D3.1 adopted a working assumption that is natural at the design stage: a SPARQL query can be assigned to a privacy *level* on the basis of its syntactic construct, with the intuition that `ASK` returns a boolean and is therefore "safe", `SELECT` returns tuples and is therefore "risky", and aggregations (`COUNT`, `AVG`) sit in between. This view was a useful working hypothesis: it allowed the consortium to scope the privacy problem early and to identify the ingredients required to address it (an annotated ontology, a federated execution layer, mappings, and governance). In the present deliverable, we build on that scaffolding and revise the underlying assumption, which, on closer analysis, is too coarse to discern the actual disclosure profile of a query. The revision rests on two observations that hold independently of the SPARQL fragment used.

The output cardinality of a query is not, in general, an upper bound on its disclosure capacity. A boolean answer to "Does patient $x$ have ALS?" is already a complete disclosure of a special category of personal data under Art. 9 of the GDPR [15]: what matters is the semantic content of the answer, not its width. By the same argument, an aggregate (`COUNT`, `AVG`) computed over a strongly identifying or selective context can be more disclosive than a wide `SELECT` over neutral properties.

In a federated OBDA/OBDF setting, even a logically aggregate query may require the planner to ship raw matching records from a site to the orchestrator before performing the aggregation. The same logical expression may therefore result either in micro-level data leaving a site or in aggregate counts only, depending on how the federation unfolds and pushes down the query [23]. The SPARQL construct constrains the *output shape* of the answer; it does not, by itself, constrain the *data movement* that produced it. Together, these two points imply that a privacy classification keyed to the SPARQL construct alone systematically miscalibrates risk: it over-permits aggregates that should be flagged and over-denies selections that are in fact innocuous. The revision pursued in this deliverable replaces the a priori level assignment with a per-query risk assessment that takes as input (i) the query graph pattern, (ii) the ontology annotations attached to the classes and properties it touches, and, where applicable, (iii) data statistics at the endpoints. The construct is retained as one signal among others, not as the sole classifier.

## 2.1 Regulatory Anchoring

The revision was developed in a joint collaboration between WP3 and WP7, where the KU Leuven team contributed to the legal and ethical analysis of the HDN. The joint analysis establishes that a per-query

risk assessment is not only a technical improvement; it is the construct that best aligns the HDN with the obligations that the GDPR [15] places on a federated infrastructure processing special categories of data.

Four provisions are load-bearing for this argument:

- **Recital 26 ("all means reasonably likely to be used").** The recital fixes the operational notion of identifiability: a datum is personal data when an adversary, using means reasonably likely to be used, can re-identify the data subject. This is intrinsically a risk-based criterion. It cannot be answered by inspecting the SPARQL construct in isolation; it requires reasoning about which properties are identifying or quasi-identifying, which paths through the ontology graph the query induces, and what data populations support those paths.
- **Art. 25 (data protection by design and by default).** Art. 25 requires that technical and organisational measures be *calibrated to the likelihood and severity of risks* arising from processing. A binary, construct-based gate cannot calibrate; a graduated, query-level assessment can, and is therefore the appropriate vehicle for compliance by design in the HDN.
- **Art. 26 (joint controllers).** The HDN is a multi-institution federation: clinical sites jointly determine purposes and means of processing for federated analytics. Art. 26 requires that joint controllers fix, in a transparent arrangement, the apportionment of responsibilities. A per-query risk score that is explainable, reproducible, and auditable provides the technical substrate on which such an arrangement can rest: a decision at the gateway is traceable to the ontology annotations and to the query structure that produced it.
- **Art. 35 (Data Protection Impact Assessment).** Art. 35 mandates a *Data Protection Impact Assessments (DPIA)* whenever the processing is likely to result in a high risk to the rights and freedoms of natural persons, explicitly including large-scale processing of Art. 9 categories (health, genetic). Art. 35(7) requires the assessment to describe the foreseen risks and the measures envisaged to address them. A query-level risk score with tailored explanations is precisely a machine-actionable artifact that feeds, and is auditable inside, the DPIA process.

Across Art. 13–15, Art. 22, Art. 25 and Art. 35, the GDPR repeatedly asks the controller to be able to *explain* the choices made when processing personal data. In the HDN, this translates into a hard requirement: a query that is rejected, downgraded, or modified at the gateway must come with a reason that can be communicated to a *Data Protection Officer (DPO)*, to a researcher, and ultimately to a data subject. The revised setting elevates explainability from a desirable property to a structural one: every risk decision is decomposable into the ontology annotations and the query features that produced it. This is the second axis along which the construct-based view of HDN v1 falls short: a level keyed to the syntactic shape of the query carries no explanation of *why* that shape is risky in the present ontology and data.

## 2.2 Two Positions in the Literature: Formalists and Pragmatists

The literature on privacy-aware ontology-mediated query answering falls into two broad positions; the design of the revised HDN setting is best understood by placing it explicitly within that landscape.

The **formalist** line, codified in the body of work on *Controlled Query Evaluation (CQE)* and *Policy-Protected OBDA (PPOBDA)* [7, 11], treats privacy as a logical property of the query–policy pair. Given a set of denial assertions or confidential views, a query is either certifiably safe under certain-answer semantics, or it is censored. The decision is binary, correctness-preserving, and exact. The cost of these guarantees is twofold. First, the decision procedure is opaque to the issuer: a denied query carries no risk-magnitude information and, often, no human-readable rationale; a routine epidemiological query that incidentally activates a denial assertion is over-denied. Second, the policy space scales poorly with user roles and contexts: covering $n$ contexts requires, in the worst case, $n$ separately maintained policy sets.

The **pragmatist** line, which the HDN revision adopts, accepts that absolute certification is neither always achievable nor always desirable in real federated infrastructures, and instead asks: given the ontology, the query, and (when needed) the data, what is the *magnitude* of the disclosure risk, expressed on a calibrated scale, and what are its drivers? Decisions remain reproducible and auditable, but they are graduated, explainable, and parameterised by governance policy rather than hardwired into denial assertions.

The two positions are not in opposition. A formalist hard floor (CQE) catches definite violations cheaply; a pragmatist graduated layer classifies what passes the floor. The revised HDN setting deploys the two in series: PPOBDA-style assertions, when present, act as a hard gate; the revised assessment, called PRISM, then scores the residual queries (see Section 2.3).

## 2.3 PRISM: Goal, Inputs, and Outputs

PRISM is the risk assessment component that operationalises the revised HDN setting. Its goal is stated narrowly: given a SPARQL query $Q$ over the HERO ontology $\mathcal{O}$, produce, before any sensitive data crosses an institutional boundary, a calibrated risk score $R(Q, \mathcal{O}) \in [0, 1]$, a discrete risk bucket $\{\text{LOW}, \text{MODERATE}, \text{HIGH}, \text{CRITICAL}\}$, and a structured explanation in which every contribution to the score is attributable to (i) an ontological aspect or (ii) a feature of $Q$.

PRISM receives three **inputs**:

1. The ontology $\mathcal{O} = (\mathcal{C}, \mathcal{P}, \mathcal{T})$, expressed in *OWL 2 Query Language (OWL 2 QL)*, has three governance annotations attached to its signature: a property sensitivity capacity $\text{sens} : \mathcal{P} \to (0, 1]$, an area sensitivity $\sigma : \mathcal{C} \to [0, 1]$, and an identification role $\text{idrole} : \mathcal{C} \cup \mathcal{P} \to (0, 1]$ distinguishing direct identifiers, quasi-identifiers, and neutral predicates. The annotation methodology is aligned with the *World Wide Web Consortium (W3C) Data Privacy Vocabulary (DPV-PD)* [14] so that the categories used in HERO map onto the personal-data categories used in legal instruments.
2. The query $Q$, normalised to a canonical form $Q^*$ by a bounded set of rewriting rules (deduplication, type subsumption, TBox-implied type elimination) that preserve the risk score and keep the analysis polynomial.
3. A policy configuration $\Theta$, set by the Data Protection Officer and the governance board, that fixes the identification thresholds, the area sensitivity scale, the bucket cut-offs, and the data-level parameters (e.g. the minimum cohort size $k$).

Given $(Q, \mathcal{O}, \Theta)$, PRISM proceeds through three sequential levels of increasing data dependence: an intensional schema-level analysis ($L_1$), a query-construct adjustment ($L_2$), and an extensional, data-aware check ($L_3$). The structure of the pipeline is summarized in Figure 1.

## 2.4 The Three Levels

The pipeline reads the query in three sequential passes. Each pass uses richer information than the previous one. The first two passes are intensional: they look only at the query and at the ontology, and do not contact any site of the federation. The third pass is extensional: it contacts the endpoints and validates the query against actual data statistics. The full formal model, with its scoring functions, its monotonicity proofs, and the algorithms that realise the three passes, is the object of the next deliverable. In this section, we describe only the design intent of each pass, why it is needed, and what it produces.

The three passes operate on a small set of tags placed by domain experts on the HERO ontology. Each class and each property receives two tags. The first tag describes the *sensitivity* of the concept: how much privacy-relevant content the class or property can carry. A property linking a person to a disease or to a genetic variant is tagged as highly sensitive; a property linking a clinical trial to a hospital site is tagged as low

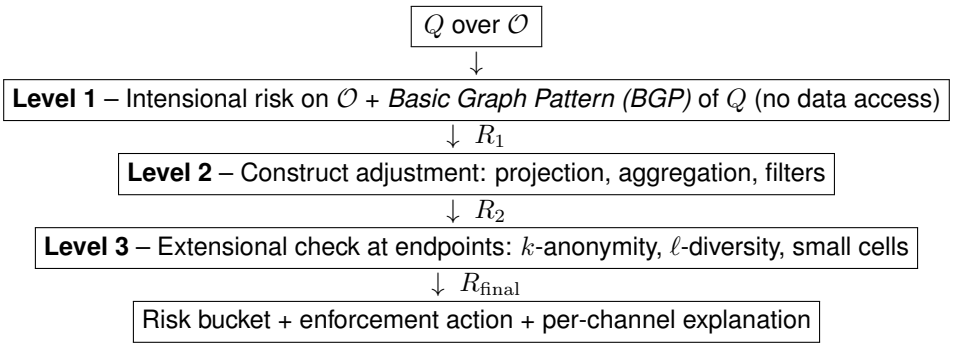



**Figure 1**: The three-level PRISM pipeline. The intensional layer (Levels 1–2) is data-free and acts as a configurable gate for the extensional layer (Level 3).

sensitivity. The tags are inherited along the class hierarchy, so a more specific class is at least as sensitive as the class above it.

The second tag describes the *identification role* of a class or property: whether its values, alone or in combination with others, can single out an individual. Properties such as the fiscal code or a national identifier are tagged as *direct identifiers*. Properties such as the year of birth, gender, or postal code are tagged as *quasi-identifiers*, because their values narrow the population substantially without identifying anyone uniquely. The remaining properties are tagged as *neutral*.

The tags are placed once, in coordination with the clinical and legal partners, through the same governance process that maintains HERO. The categories used to assign tags are aligned with the W3C DPV-PD, so that the labels used inside HERO map onto the personal-data categories used in legal instruments. These two tags are the only domain-expert input PRISM needs in addition to the query and the ontology.

### Level 1 – Schema-Level Risk

The first pass looks at the query only against the ontology. No site is contacted, and no row is read. It asks two questions about the query's graph pattern.

The first question is what sensitive content the query traverses. The pipeline reads the chain of classes and properties that the query stitches together and aggregates the sensitivity tags along that chain. A query that connects a patient to a disease and to a clinical exam traverses more sensitive ground than a query that relates a disease label to a trial code, and the aggregate reflects that.

The second question is whether the query touches any identifying properties at all. The pipeline checks whether any of the properties involved in the query is a direct identifier or a quasi-identifier, irrespective of whether their values are returned in the answer. Touching an identifier in the graph pattern is, by itself, a precondition for re-identification.

The two answers are combined into a single Level 1 score. The combination is conjunctive on purpose: high schema-level risk requires *both* meaningful sensitive content *and* a way to attach that content to an individual. A query that traverses very sensitive properties but no identifier (for instance, an epidemiological aggregate over a population of millions) and a query that touches an identifier but carries no sensitive content (for instance, the fiscal codes of clinical-trial principal investigators) both end up below the threshold that the model considers concerning. The score becomes high only when both ingredients are present, which is the regime in which re-identification is plausible.

Level 1 requires no data access, runs in time polynomial in the size of the query, and produces the same result at every site of the federation: the same query over the same annotated HERO yields the same Level 1

score regardless of which endpoint ultimately executes it.

### Level 2 – Adjustment by the Query Construct

The second pass takes the Level 1 score and refines it by examining what the SPARQL construct actually does with the variables identified in Level 1. Three distinct effects are considered.

The first effect concerns what the query *returns*. A query that traverses a sensitive or identifying property in its graph pattern but does not return it is qualitatively less risky than a query that does return it. Aggregations enter through the same channel: `COUNT` or `AVG` reduces exposure because the output is a summary rather than a list of records; very selective `FILTER` clauses increase exposure because they restrict the answer to a small subgroup; `GROUP BY` on a quasi-identifier effectively re-exposes that property through the row labels of the result.

The second effect concerns the *direct exposure* of identifiers. A direct identifier in the `SELECT` clause, such as a fiscal code, is the most explicit form of re-identification risk: this effect alone raises the Level 2 score significantly, regardless of the rest of the query.

The third effect concerns the *indirect exposure* through quasi-identifiers. A year of birth alone is rarely identifying; a year of birth, gender, and postal code together can be. The pipeline combines the Level 1 score with the projected quasi-identifiers, so that a query is flagged when a combination of properties is jointly identifying, even if no single property would be.

The three effects are not averaged. The Level 2 score is the *largest* of the three. The three effects describe qualitatively different kinds of disclosure (output exposure, direct exposure, indirect exposure), and the gateway must respond to the worst case scenario among them. A query that projects a fiscal code is risky because of that fact, regardless of how innocuous the rest of the pattern is. Level 2, like Level 1, is a schema-only pass: it inspects the query and the ontology, not the data.

### Level 3 – Data-Aware Validation

Whether the third pass runs at all is a policy decision. A governance-set threshold separates queries that the institution is willing to validate against actual data from queries it wants to block on the basis of the schema alone. Queries that exceed the threshold after Level 2 are rejected at the gateway without contacting any endpoint, which is the regime in which the most dangerous requests are kept away from the data by design. All remaining queries proceed to Level 3.

Level 3 contacts the endpoints involved in the query and validates the query against three requirements that are standard in the data-anonymisation literature.

The first requirement is that the population matched by the query be *large enough*: each combination of quasi-identifier values that the query would expose must be shared by at least $k$ records, with $k$ set by policy. This is the standard $k$-anonymity criterion. A query that, on the present data, would isolate a cohort of a handful of individuals is downgraded or refused, even if its schema-level score is moderate.

The second requirement is that the sensitive values within the matched cohort be *diverse enough*: each group must contain at least a configurable number of distinct values of the sensitive attribute that the query exposes. This is the $\ell$-diversity criterion, and it captures situations in which a group is large, but the sensitive attribute is constant across it (a large cohort, all of whom have the same rare disease, still discloses the disease for every member of the group).

The third requirement applies to aggregate outputs and rules out *small cells*: counts below a policy-set minimum are returned redacted rather than as exact numbers so that aggregate queries cannot be used to single out individuals through small-count cells.

The output of Level 3 is the final risk score, the bucket the query falls into, and the structured explanation that lists, for each contribution, which ontology annotation or which query feature drove it. This explanation is what makes PRISM's decisions auditable in the sense required by the GDPR articles discussed in Section 2.1: a rejected, downgraded, or redacted query comes with a reason that can be reported to the DPO and, where applicable, to the data subject. Endpoint contact is incurred only by Level 3, and only after Levels 1 and 2 have established that the query is worth executing; this design preserves data minimisation at the gateway itself.

## 2.5 Why the Three-Level Pipeline Is Adequate for HEREDITARY

The pipeline is adequate for the HDN along three axes that map onto the constraints set in Deliverable D3.1.

Because Levels 1–2 operate on the schema, they are independent of which sites are involved in a given query: the same intensional risk is computed regardless of whether the query unfolds over one endpoint or many. Level 3 then specialises the assessment to the actual data populations at the endpoints that the federation contacts. The separation matches the OBDF architecture documented in D3.1 and D3.2, in which sites retain autonomy and the orchestrator does not centralise data.

The score is decomposable by construction: every contribution is traceable to either an ontology annotation or a query feature. This makes the gateway directly auditable, gives the DPO a substrate on which to argue Art. 35 assessments, and supports the role-based adaptation that the HDN requires: namespace-partitioned annotations let the same mathematical machinery produce different scores for different user categories without altering the model.

The pipeline is designed so that refining a query, by adding predicates, widening the projection, or tightening filters, cannot lower the assessed risk. This monotonicity is the technical analogue of the regulatory principle that requests for more data cannot be treated as less risky than requests for less, and is what makes PRISM's decisions defensible in front of a controller, a DPO, and an auditor.

## 2.6 Scope and Roadmap: From Single-Query Assessment to the HDN Gateway

In this deliverable, we introduce PRISM at the granularity at which the framework is already well-defined and validated: *the single-query setting*. For a query $Q$ submitted in isolation against the HDN, the pipeline produces $R_{\text{final}}(Q)$, the bucket, and the explanation. The HDN gateway can already enforce per-query policy on top of this output.

Two extensions are scoped for subsequent project deliverables.

PRISM has been validated on the BrainTeaser Ontology [16]; its instantiation on HERO requires the annotation campaign described in Deliverable D3.1 to be extended with the three governance functions ($\mathrm{sens}$, $\sigma$, $\mathrm{idrole}$). This activity will be reported in the next WP3 deliverable, jointly with the policy configuration agreed with the clinical partners.

The single-query assumption is a deliberate restriction. Two phenomena that PRISM does not yet model are: (i) the *compositional* risk that arises when a sequence of individually acceptable queries jointly reconstructs a sensitive attribute through differencing or intersection; and (ii) the *session-level* exposure that arises when the same user issues many queries within a window and adversarial recombination is feasible. Addressing these phenomena requires a session-level model of cumulative disclosure, which we plan to develop in the final part of the project on top of the per-query foundation introduced here. The single-query pipeline defined in this deliverable is a deliberate subroutine of that extension; the multi-query layer composes per-query scores rather than recomputing them.

The HDN privacy setting evolves from a static, construct-based level assignment to a per-query, ontology-aware risk assessment grounded in the GDPR risk-based posture and developed jointly with WP7. The mech-

anism, PRISM, produces a calibrated and explainable score in three levels, with the intensional core acting as a data-minimising gate and the extensional layer attached only when the data is actually needed. The full formal model, the HERO annotation effort, and the multi-query extension are scheduled for the subsequent WP3 deliverables.

### 2.7 Synthetic Data as a Fallback

The HDN pursues two primary objectives that are jointly load-bearing for the project: to answer as many queries as the governance posture allows and to enable federated learning across the participating sites. PRISM, as described so far, serves the first objective by classifying queries on a calibrated risk scale and gating them through the three levels of the pipeline. The pipeline, however, also defines a frontier: queries that exceed the policy-set threshold after Level 2, and queries that fail the data-level requirements at Level 3, are not answerable from the real data without violating that posture. A binary outcome at the frontier (answer or refuse) would sacrifice answerability for protection by construction. HEREDITARY addresses this trade-off by investing in *synthetic data* as a controlled fallback for the queries that the pipeline cannot clear.

A synthetic dataset reproduces the statistical structure that the query needs (joint distributions of the variables involved, cohort sizes, dependence patterns) while severing the link to identifiable individuals. When PRISM determines that a query cannot be answered against the real data, or against a differentiated view of the real data (role-restricted, redacted, or otherwise perturbed), the gateway re-routes the query to a synthetic substitute, answers it there, and marks the answer in the explanation as synthetic-derived. The same mechanism feeds federated-learning workflows: model training and auditing can be performed on the synthetic substitute when the underlying records cannot leave the sites that hold them, which is the second of HEREDITARY's primary objectives.

PRISM is the component that decides when and how this substitution happens. The risk bucket is the trigger: queries in the LOW bucket are answered from the real data; queries in the MODERATE bucket may be answered from a differentiated view of it; queries in the HIGH bucket that carry a legitimate analytical purpose are redirected to synthetic data rather than refused outright. The per-channel explanation produced by the pipeline is informative beyond the trigger: it tells the synthetic-data generator *which* statistical properties the substitute must preserve to answer the query usefully (for example, the marginal of a sensitive attribute, or a specific joint that the query aggregates over), and it tells the governance team *which* features of the real data must *not* be carried over to the substitute (typically the direct identifiers and the quasi-identifier combinations that drove the high score in the first place).

The net effect of coupling PRISM with the synthetic-data layer is a gateway that maximises answerability without weakening the auditability of the decision: every answer the gateway returns is either backed by real data that the pipeline cleared, or backed by a synthetic substitute whose substitution decision is itself recorded and explainable in the same terms.

## 3 Background in Privacy-Preserving Synthetic Data Generation for tabular data

This section reports on our current activities to analyse how techniques from the state of the art in privacy-preserving synthetic data generation perform when applied to the ALS data of HEREDITARY (related to use cases 1 and 2).

## 3.1 Differential privacy

*Differential Privacy (DP)* is a mathematical framework for quantifying the privacy risk of algorithms that operate on sensitive data. The intuition of DP is that an algorithm would not significantly change the output when the input changes by one individual, who can either be added or removed.

**Definition 1** (*Differential Privacy (DP)*)**:** Let $\mathcal{M}$ be a randomised mechanism that maps datasets to an output space $\mathcal{O}$. Let $D$ and $D'$ be two datasets that differ in one record. The mechanism $\mathcal{M}$ is $(\varepsilon, \delta)$-differentially private if, for each subset $S \subseteq \mathcal{O}$:

$$P(\mathcal{M}(D) \in S) \leq e^{\varepsilon} P(\mathcal{M}(D') \in S) + \delta$$

The parameter $\varepsilon$, usually referred to as *privacy budget*, controls the trade-off between privacy and utility: the smaller the value of $\varepsilon$, the stricter the privacy. The parameter $\delta$ allows for the possibility that the guarantee does not hold. When $\delta = 0$, the mechanism enjoys *pure differential privacy* (or is $\varepsilon$-differentially private); when $\delta > 0$, the mechanism enjoys *approximate differential privacy* (or is $(\varepsilon, \delta)$-differentially private).

## 3.2 Methods for tabular data synthesis

We can broadly distinguish two families of methods for synthesising tabular data under privacy constraints: those based on neural networks and those based on *Probabilistic Graphical Models (PGMs)*.

The first family builds on modern deep generative model architectures, such as GANs and *Variational Auto Encoders (VAEs)*. The former are models that use two neural networks, a generator and a discriminator, which compete against each other in a game-like process to produce realistic synthetic data; VAEs are probabilistic models that learn to encode input data in a compressed latent space, from which new, smooth variations of the original data can be sampled and decoded. Examples of methods in this family include DPGAN [45] and PATE-GAN [28], which enhance GANs with differential privacy to ensure that the generated data is privacy-preserving.

The second family uses explicit structural models that define dependencies between variables, such as *Bayesian Networks (BNs)* and *Markov Random Fields (MRFs)*. The former are PGMs that represent variables and their conditional dependencies using a *Directed Acyclic Graph (DAG)*, allowing for efficient representation and calculation of probabilities over observed and unobserved variables. The latter are undirected graphical models that represent the statistical dependencies between random variables, where the probability distribution is defined by functions on cliques in an undirected graph. Examples of methods in this family include PrivBayes [52] and PrivMRF [6].

In the remainder of the section, we introduce the three methods we consider in this deliverable: DPGAN, PATE-GAN and PrivBayes.

### 3.2.1 Neural Network-based approaches

Neural network-based approaches for privacy-preserving synthetic data generation rely on generative AI models, such as GAN and VAE. Many of the existing methods in this category build on top of the GAN architecture.

A GAN [22] is an architecture designed for generating synthetic data that closely mimics the distribution of real-world samples. At its core, a GAN comprises two competing neural networks: the generator ($G$) and the discriminator ($D$). The generator's objective is to learn the underlying probability distribution $P_{data}$ by mapping random noise variables $z \sim P_z(z)$ into synthetic data samples. The discriminator, conversely, acts as a classifier that attempts to distinguish between genuine real data and the fake data produced by the generator. The training process is formalized as a minimax game: the generator constantly improves

its realism to fool the discriminator, while the discriminator continuously refines its ability to spot artifacts, resulting in an equilibrium where the generated samples are realistic.

**Definition 2** (*Generative Adversarial Network (GAN)*)**:** A generator is a mapping $G : \mathcal{Z} \to \mathcal{X}$ that transforms a random noise $z \sim P_z(z)$ into a synthetic sample, i.e. $G(z) = x$. A discriminator is a function $D : \mathcal{X} \to [0, 1]$ that outputs the probability that a given sample $x$ comes from the real data distribution $P_{data}$ rather than the generator $G$, i.e. $D(x) = P(x \sim P_{data})$. A GAN trains the generator and the discriminator as a minimax game:

$$\min_G \max_D \; V(D, G) = \mathbb{E}_{x \sim P_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim P_z(z)} \left[\log \left(1 - D(G(z))\right)\right],$$

where $V(D, G)$ indicates the value function.

*Differentially-Private GAN (DPGAN)* extends the standard GAN training by incorporating differential privacy into discriminator training, since the learning process accesses the sensitive data only in the context of the discriminator training. The key idea is to bound and perturb the gradients used to update the discriminator parameters by employing the *Differentially Private Stochastic Gradient Descent (DP-SGD)* [1]: gradients are first clipped to limit the impact that individual data points can have on the learning, and then Gaussian noise is added to the gradients before performing the parameter update. As a result, DPGAN is a $(\epsilon, \delta)$-differential privacy mechanism for privacy-preserving synthetic data generation.

PATE-GAN is another $(\epsilon, \delta)-$differential privacy mechanism for privacy-preserving synthetic data generation that takes a different approach. It builds on *Private Aggregation of Teacher Ensembles (PATE)* [36], a paradigm that wraps a machine learning algorithm in a differential privacy framework. PATE-GAN uses multiple teacher discriminators, each trained on a disjoint subset of the real data that contains sensitive information.

**Definition 3** (PATE-GAN's teacher)**:** Let $X$ be the real dataset, let $(X_1, \ldots, X_k)$ be a partition of $X$, and let $G$ be the generator network. The $i$-th teacher $T_i$ learns from the real data from the corresponding dataset $X_i$ and the synthetic data from the generator $G(z), z \sim P_z(z)$. The objective function is:

$$\max_{T_i} \left(\mathbb{E}_{x \sim P_{data,i}} \log T_i(x) + \mathbb{E}_{z \sim P_z(z)} \log \left(1 - T_i \left(G(z)\right)\right)\right)$$

During training, the teachers vote on whether the samples are real or fake, and their aggregated output is perturbed with noise according to a $(\epsilon, \delta)-$differential privacy mechanism. A student discriminator is then trained on these noisy, aggregated labels, and the generator learns from the student as in the standard GAN setup.

**Definition 4** (PATE-GAN's student)**:** Let $\hat{y}(x)$ be the noisy aggregated label obtained from the teacher ensemble for a sample $x$, and let $G$ be the generator. The student $S : \mathcal{X} \to [0, 1]$ is trained to approximate these labels. The objective function is:

$$\min_S \left(\mathbb{E}_{x \sim P_{data}} \ell \left(S(x), \hat{y}(x)\right) + \mathbb{E}_{z \sim P_z(z)} \ell \left(S(G(z)), \hat{y}(G(z))\right)\right),$$

where $\ell$ is a suitable loss function.

We used the OpenDP Smartnoise Synthesizer[1] implementations of DPGAN and PATE-GAN. In the case of PATE-GAN, we patched the implementation with the bug fix proposed by Ganev et al. [18].

[1] Cf. https://smartnoise.org/

### 3.2.2 PGM-based approaches

*Probabilistic Graphical Models (PGMs)* use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space [30]. *Bayesian Networks (BNs)* are PGMs that represent the probabilistic relations between variables as a *Directed Acyclic Graph (DAG)* ($\mathcal{G}$), where variables are represented as vertices ($V$), and the relations among them are modelled as directed edges ($E$).

**Definition 5** (*Bayesian Network (BN)*)**:** Let $X = (X_1, \ldots, X_m)$ be a collection of $m$ random variables, each taking on a value within a finite domain. A BN $\mathcal{B}$ over $X$ is a pair $\mathcal{B} = (\mathcal{G}, \mathcal{P})$, where $\mathcal{G} = (V, E)$ is a DAG with the set of vertices $V = \{X_1, \ldots, X_m\}$, and for each variable $X_i \in V$, $\Pi_i \subseteq V \setminus \{X_i\}$ denotes the set of parents for $X_i$ in $\mathcal{G}$. $\mathcal{P}$ is the set of conditional probability distributions, such that $\mathcal{P} = \{P(X_i|\Pi_i)\}_{i=1}^{m}$. The joint distribution of $\mathcal{B}$ is thereby defined as:

$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i|\Pi_i) \tag{1}$$

The graph structure $\mathcal{G}$ can be obtained either through structure learning algorithms or by incorporating domain knowledge to define the relations among variables in $\mathcal{G}$.

Given a BN, a synthetic dataset $Z = \{z^{(1)}, \ldots, z^{(N)}\}$ can be generated by sampling from the joint distribution encoded by the network, where each record $z^{(i)} = (z_1^{(i)}, \ldots, z_m^{(i)})$ is independently drawn according to this distribution.

PrivBayes [52] utilises a noisy local Bayesian score based on information gain:

$$I(X_i \mid \Pi_i) \tag{2}$$

This score is used to construct a BN structure $\mathcal{G}$ over the $m$ attributes in $D$, while making the process differentially private by adding noise to the sampling probability of attribute-parent pairs, that have a prior sampling probability given by the mutual information gain. Once a noisy $\mathcal{G}$ is constructed, the conditional probability distributions are learned. PrivBayes uses a parameter learning algorithm with noise for privacy-preservation. For private parameter learning, PrivBayes allocates a privacy budget to estimate noisy conditional distributions. For each attribute-parent pair $(X_i, \Pi_i)$, Laplace noise is injected to make the process differentially private, yielding a noisy probability distribution:

$$P^*(X_i, \Pi_i) = P(X_i, \Pi_i) + Lap(\frac{2(d-k)}{n\varepsilon_2}) \tag{3}$$

where $d$ is the number of attributes, $k$ is the maximum number of parents in $\mathcal{G}$, and $n$ is the number of records in the dataset.

We used the DataSynthesizer implementation of PrivBayes from DataResponsibly[2].

# 4 Privacy assessment of synthetic tabular data

We introduced the idea of synthetic data in HEREDITARY as fallback mechanism to preserve query answerability under privacy constraints, as described in Section 2.7. While PRISM governs the disclosure risk of queries executed over real data, synthetic data is used when these queries cannot be safely resolved within the federation. This shift introduces a complementary problem: how to assess, in a principled and operational manner, the privacy risk associated with the synthetic datasets that replace the original data.

[2]Cf. https://dataresponsibly.github.io

Discussions with use case partners and data owners have made this limitation explicit. In practice, it is not sufficient to rely on theoretical guarantees of the PP-SDG mechanism alone. Mechanisms based on DP, for example, provide formal bounds on disclosure risk, but these guarantees are expressed through abstract parameters ($\varepsilon$ and $\delta$) that do not directly translate into an intuitive and interpretable notion of risk for domain experts. Furthermore, these guarantees assume the correct implementation and configuration of the PP-SDG mechanisms, and even relying on existing libraries does not guarantee correctness, as highlighted in recent studies [18]. Therefore, in a realistic setting, deviations in modelling choices, parameter tuning, or software implementations may affect the actual level of protection that one can achieve.

As a consequence, the assessment of privacy in synthetic data cannot be reduced to the specification of the PP-SDG process. Instead, it requires evaluation methods that quantify the extent to which a synthetic dataset may expose information about the original data. The recent literature proposes a broad range of such metrics, each capturing specific threat models (e.g., re-identification, attribute inference, membership inference) and relying on different computational assumptions. However, these metrics have different definitions; they often overlap in intent but differ in formulation, and there is not a standardised set of metrics that one can safely rely on.

In this section, we address this gap with a twofold contribution. Firstly, we survey privacy metrics for synthetic tabular data and organise them in a coherent taxonomy, reformulating their definitions using a shared mathematical notation. This unifying step is intended to make the relationships between metrics explicit and to provide a foundation for future work on comparing and contrasting privacy metrics. Secondly, we operationalise the survey results in PrivEval, a tool that implements 17 of these metrics and supports data owners in evaluating the privacy properties of synthetic data in practice.

## 4.1 Privacy Metrics

We start the survey by specifying the key concepts of our setting.

**Definition 6** (Dataset)**:** Let $A = \{a_1, a_2, \ldots, a_m\}$ be a set of attributes, and let $t$ map each attribute to its datatype, i.e. $t : A \to \{\text{continuous}, \text{categorical}\}$.

A dataset is a multiset $D$ containing pairs $(\hat{D}, f_D)$, where $\hat{D}$ is the set of tuples in the multiset:

$$\hat{D} \subset rng(a_1) \times rng(a_2) \times \cdots \times rng(a_m), \tag{4}$$

where $rng$ denotes the range of an attribute, and $f_D : D \to \mathbb{N}$ is a multiplicity function such that, for each $d \in \hat{D}$, $f_D(d)$ gives its frequency in $D$.

When referring to multisets, we use:

- $|D|$ for their size, i.e. $|D| = \sum_{d \in \hat{D}} f_D(d)$,
- $d \in D$ to denote membership, i.e. $d \in D \iff d \in \hat{D}$, and
- $D \setminus \{d\}$ for removing all occurrences of $d$ from $D$, i.e. $D \setminus \{d\} = (\hat{D}', f_{D'})$ such that $d \notin \hat{D}'$.

Let $A' \subseteq A$ be a subset of attributes. For $d_j \in D$, we write $d_j[A']$ for the projection of $d_j$ onto $A'$ (it follows that $d_j = d_j[A]$). Abusing the notation, $D[A']$ denotes the projection of all tuples in $D$ over $A'$, i.e. $\{d_1[A'], \ldots, d_n[A']\}$, and $d_j[a_i]$ is the value of attribute $a_i$ in the $j$-th instance.

Let $\mathcal{D}$ be the set of all datasets, with $Y \in \mathcal{D}$, the real sensitive dataset and $Z \in \mathcal{D}$, the synthetic dataset generated by a PP-SDG mechanism. A privacy metric is defined as follows.

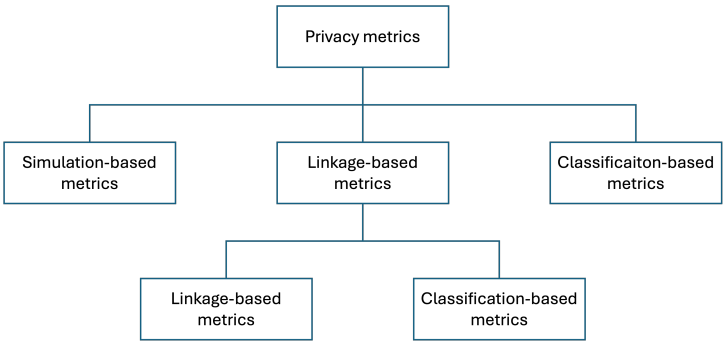



**Figure 2**: Taxonomy of privacy metrics.

**Definition 7** (Privacy Metric)**:** A privacy metric $p$ measures the threat introduced by the synthesis mechanism by mapping a real dataset and a synthetic dataset to a score in $[0, 1]$:

$$p : \mathcal{D} \times \mathcal{D} \to [0, 1],$$

where a score of $0$ means full privacy and a score of $1$ means no privacy.

When a metric $p_1$ is well-designed, one expects $p_1(Y, Z) = 1$ when $Z = Y$. Likewise, if $Y$ and $Z$ are independent, $p_1(Y, Z)$ should be close to $0$.

### 4.1.1 Metric Categorisation

The privacy metrics proposed in the literature differ in how they operationalise disclosure risk. While they often target similar threat models, their computational structures vary, making direct comparison difficult. To clarify these differences, we organise the metrics into a hierarchical taxonomy (in Figure 2) that reflects the underlying mechanisms through which risk is estimated.

At the highest level, we categorise the privacy metrics in three families: simulation-based, linkage-based, and classification-based. These families correspond to different ways of modelling an adversary and a privacy attack, and quantifying the success in exploiting the synthetic data to learn about the real data.

**Simulation-based metrics** Simulation-based metrics model an adversary attempting to infer information about the real dataset. Given access to the synthetic data, and optionally to partial knowledge about the individuals in the real dataset, the adversary produces a set of guesses linking real and synthetic data records. An oracle then determines which guesses are correct, allowing the metric to quantify the success of the attack. These metrics, therefore, relate to a concrete threat model and directly measure its outcome.

**Linkage-based metrics** Linkage-based metrics rely on constructing associations between real and synthetic records, typically through nearest-neighbour matching or distance-based similarity measures. The privacy-risk is then derived from the structure of these associations. Metrics in this family share a common computational pipeline: a set of links between records is first established, and the metric is computed by analysing these links. We can further define two subcategories of metrics:

- **Threshold-based metrics** filter the set of links according to a predefined condition (e.g., a distance threshold) and compute the score from the proportion of links satisfying that condition.

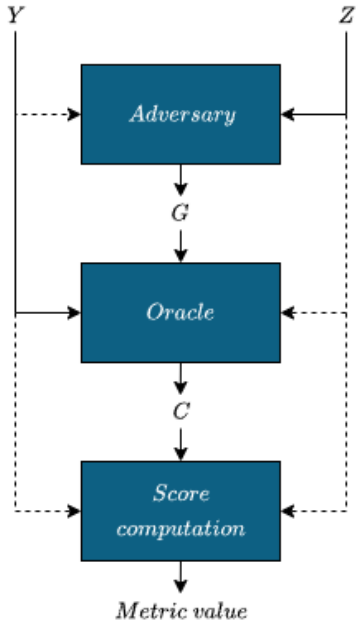



**Figure 3**: Metric computation in simulation-based privacy metrics.

- **Distance–based metrics** use all the links to compute the final score, defining the risk as a continuous function of the distances between matched records.

**Classification-based metrics** Classification-based metrics frame privacy as a prediction problem. Instead of explicitly constructing links or simulating specific attacks, these metrics train a model to distinguish between real and synthetic records, or to infer properties such as membership or sensitive attributes. The performance of the classifier is then used as a proxy for the information leakage captured by the synthetic dataset.

This taxonomy highlights that privacy metrics differ not only in their outputs, but also in the assumptions they make about the adversary and in the computational artifacts they use (guesses, links, or learned models). In the following sections, we present the metrics we identified in each category.

### 4.1.2 Simulation-based privacy metrics

The simulation-based privacy metrics model an adversary attempting a privacy attack, and their typical computation is illustrated in Figure 3. These metrics model an adversary that uses $Z$ and, optionally, part of $Y$ as background knowledge, producing a set $G \subset Y \times Z$ of *guesses*, i.e., associations between elements of $Y$ and $Z$. An oracle then identifies the correct guesses $C \subseteq G$, and optionally the wrong and missing ones, as shown in Figure 3. The score is computed from $C$ and $G$.

***Categorical Zero CAP (ZCAP)* [26]** ZCAP estimates the risk of inferring a categorical sensitive attribute ($a_S \in A$) from key attributes $A_K \subseteq A \setminus \{a_S\}$, which may be continuous or categorical. It assumes the adversary knows $y[A_K]$ and all attributes of the synthetic dataset $Z$.

The attack is simulated by matching real and synthetic records on $A_K$ and checking how many of these also agree on $a_S$. We define the candidate matches as:

$$G_{\text{ZCAP}} = [\![(y,z)|y \in Y, z \in Z, y[A_K] = z[A_K]]\!],$$

and the exact matches as:

$$C_{\text{ZCAP}} = [\![(y,z)|(y,z) \in G_{ZCAP}, y[a_S] = z[a_S]]\!].$$

For a real record $\hat{y} \in Y$:

$$\text{zcap}(\hat{y}, Z) = \frac{|\,[\![z|(\hat{y},z) \in C_{\text{ZCAP}}]\!]\,|}{\max\{1, |\,[\![z|(\hat{y},z) \in G_{\text{ZCAP}}]\!]\,|\}}, \tag{5}$$

where the denominator is the number of matches, or $1$ if none exist. For the full dataset:

$$\text{ZCAP}(Y,Z) = \frac{1}{|Y|}\sum_{y \in Y} \text{zcap}(y,Z). \tag{6}$$

Thus, ZCAP ranges from $0$ to $1$, with higher values indicating a greater reconstruction risk.

***Categorical Generalized CAP (GCAP)* [26]** GCAP relaxes ZCAP by replacing exact matching with similarity. If no synthetic record shares the same key values, the closest synthetic record(s) are used instead. As in ZCAP, it assumes that the adversary knows $y[A_K]$ and $Z$, and that $a_S$ is categorical.

The matching multiset is defined as:

$$G_{\text{GCAP}} = \left[\!\!\left[(y,z)|y \in Y, z = \arg\min_{z' \in Z} d_H(y[A_K], z'[A_K])\right]\!\!\right],$$

where $d_H$ indicates the Hamming distance. $G_{\text{GCAP}}$ contains each real record and its closest synthetic match in the key-attribute space.

The definition of $C_{\text{GCAP}}$ mirrors that of $C_{\text{GCAP}}$:

$$C_{\text{GCAP}} = [\![(y,z)|(y,z) \in G_{\text{GCAP}}, y[a_S] = z[a_S]]\!].$$

As for ZCAP, we first define GCAP for a real individual $\hat{y} \in Y$:

$$\text{gcap}(\hat{y}, Z) = \frac{|\,[\![z|(\hat{y},z) \in C_{\text{GCAP}}]\!]\,|}{|\,[\![z|(\hat{y},z) \in G_{\text{GCAP}}]\!]\,|}, \tag{7}$$

and for the whole real dataset:

$$\text{GCAP}(Y,Z) = \frac{1}{|Y|}\sum_{y \in Y} \text{gcap}(y,Z). \tag{8}$$

A score of $0$ means no individual is reconstructed, while $1$ means full reconstruction.

***Attribute Inference Risk (AIR)* [49]** AIR extends GCAP to categorical and continuous attributes. It matches real records with the closest synthetic ones in the key-attribute space, using Euclidean distance. Categorical attributes are encoded as numeric vectors with one hot encoding. The matching multiset is:

$$G_{\text{AIR}} = \left[\!\!\left[(y,z)|y \in Y, z = \arg\min_{z' \in Z} d_E(y[A_K], z'[A_K])\right]\!\!\right],$$

where $d_E$ is the Euclidean distance.

The sensitive attribute $a_S$ may be categorical or continuous. For categorical $a_S$, correct matches are defined as in ZCAP and GCAP:

$$C_{\text{AIR}} = [\![(y,z)|(y,z) \in G_{\text{AIR}}, y[a_S] = z[a_S]]\!] .$$

For continuous $a_S$, a match is correct if the values differ by at most $10\%$:

$$C_{\text{AIR}} = [\![(y,z)|(y,z) \in G_{\text{AIR}}, |y[a_S] - z[a_S]| \leq y[a_S] \cdot 0.1)]\!] .$$

AIR measures attack success with the F1 score. Here, $C_{\text{AIR}}$ gives true positives, $F_{\text{AIR}} = G_{\text{AIR}} \setminus C_{\text{AIR}}$ gives false positives, and false negatives are:

$$M_{\text{AIR}} = [\![(y,z)|\exists \bar{z} : (y,\bar{z}) \in F_{\text{AIR}}, z \in Z, y[a_S] = z[a_S]]\!] ,$$

for categorical $a_S$, and:

$$M_{\text{AIR}} = [\![(y,z)|\exists \bar{z} : (y,\bar{z}) \in F_{\text{AIR}}, z \in Z, |y[a_S] - z[a_S]| \leq z[a_S] \cdot 0.1)]\!] ,$$

for continuous $a_S$.

Let $TP = |C_{\text{AIR}}|$, $FP = |F_{\text{AIR}}|$ and $FN = |M_{\text{AIR}}|$, the F1 score is:

$$F_1(y[a], Z) = \frac{2 \cdot \frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}}. \tag{9}$$

Finally, AIR for the synthetic dataset is defined as:

$$\text{AIR}(Y,Z) = \sum_{y \in Y} w(y) \cdot F_1(y[a_s], Z), \tag{10}$$

where $w(y)$ is:

$$w(y) = \frac{-P(y)\ log(P(y))}{H(Y)}, \tag{11}$$

and $P(y)$ is the probability of observing the attribute values of $y \in Y$.

The entropy of $Y$ is:

$$H(Y) = -\sum_{y \in Y} P(y)\ log(P(y)). \tag{12}$$

***Common Rows Proportion (CRP)*** **[39]** CRP estimates re-identification risk by checking whether a synthetic individual exactly matches a real one. It assumes access to $Z$ and uses equality for re-identification. The matching set is:

$$G_{\text{CRP}} = [\![(z,z)|z \in Z]\!] ,$$

from which the correct matches are:

$$C_{\text{CRP}} = [\![(z,z)|(z,z) \in G_{\text{CRP}}, z \in Y]\!] .$$

The CRP score is:

$$\text{CRP}(Y,Z) = \frac{|C_{\text{CRP}}|}{|Y| + 10^{-8}}, \tag{13}$$

where $10^{-8}$ ensures numerical stability. A score of $0$ means no identical real and synthetic individuals, while $\text{CRP} \approx 1$ means all are identical.

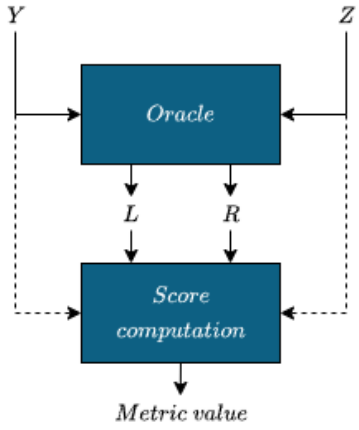



**Figure 4**: Metric computation in threshold-based privacy metrics.

### 4.1.3 Threshold-based privacy metrics

Threshold-based privacy metrics directly compare the real and synthetic datasets $Y$ and $Z$, as shown in Figure 4. The process to compute these metrics directly analyses $Y$ and $Z$. A nearest-neighbour linker builds a set $L$ of pairs, where each real record is matched to its closest synthetic one, or vise versa. The set $L$ is then filtered into a subset $R$ that satisfies a given condition. The metric is computed from the size of $R$ relative to $Y$, $Z$, or $L$.

***Close Value Probability (CVP)*) [39]** CVP estimates the chance that a synthetic record can be linked back to the real individual it comes from, and thus re-identify that person. It treats re-identification as a distance problem: if the normalised distance between a real record $y$ and its nearest synthetic record $z$ is below $0.2$, the record is considered identifiable. We first define the link set:

$$L_{\text{CVP}} = \left[\!\left[(y,z)|y \in Y, z = \arg\min_{z' \in Z} d_E(y,z')\right]\!\right], \tag{14}$$

where $d_E$ is the Euclidean distance. Unlike ZCAP, GCAP, and AIR, this distance is computed over all attributes, not only the key ones.

Let also:

$$d_{max} = \max_{y \in Y, z \in Z} d_E(y,z) \tag{15}$$

$$d_{min} = \min_{y \in Y, z \in Z} d_E(y,z) \tag{16}$$

be the maximum and minimum distance values between elements of $Y$ and $Z$. The risk set is then:

$$R_{\text{CVP}} = \left[\!\left[(y,z)|(y,z) \in L_{CVP}, \frac{d_E(y,z) - d_{min}}{d_{max} - d_{min}} \leq 0.2\right]\!\right],$$

and the metric is:

$$\text{CVP}(Y,Z) = \frac{|R_{\text{CVP}}|}{|Y|}. \tag{17}$$

A score of $1$ means that all real records are re-identifiable, while $0$ means that none are.

***Distant Value Probability (DVP)* [39]** DVP looks at the opposite case: it measures how often a real record is far from its nearest synthetic neighbour. It uses the same nearest-neighbour links as CVP, so $L_{\text{DVP}} = L_{\text{CVP}}$, together with the same $d_{max}$ and $d_{min}$ from Equation 15 and Equation 16. The far-distance set is:

$$R_{\text{DVP}} = \left[\!\!\left[(y,z)|(y,z) \in L_{\text{DVP}}, \frac{d_E(y,z) - d_{min}}{d_{max} - d_{min}} \geq 0.8\right]\!\!\right],$$

and:

$$\text{DVP}(Y,Z) = \frac{|R_{\text{DVP}}|}{|Y|}. \tag{18}$$

To match the convention in 7, we use:

$$\text{DVP}'(Y,Z) = 1 - \text{DVP}(Y,Z).$$

A score of $1$ means no synthetic record is far from the real data, while $0$ means they all are.

***Authenticity (Auth)* [39]** Auth checks whether real records are closer to other real records than to synthetic ones. For each $y \in Y$, we find its nearest synthetic neighbour:

$$L_{\text{Auth}} = \left[\!\!\left[(y,z)|y \in Y, z = \arg\min_{z' \in Z} d_E(y,z')\right]\!\!\right], \tag{19}$$

and define:

$$R_{\text{Auth}} = [\![(y,z)|(y,z) \in L_{\text{Auth}}, \exists y' \in Y \setminus \{y\} : d_E(y,y') < d_E(y,z)]\!]. \tag{20}$$

The score is the fraction of real individuals whose nearest real neighbour is closer than their nearest synthetic one:

$$\text{Auth}(Y,Z) = \frac{|R_{\text{Auth}}|}{|Y|}. \tag{21}$$

Again, we use:

$$\text{Auth}'(Y,Z) = 1 - \text{Auth}(Y,Z),$$

so higher values indicate more non-authentic synthetic records, while lower values indicate the opposite.

***IDentifiability Score (IDS)* [39]** IDS estimates re-identification risk while weighting attributes by their informational value. For each attribute $a \in A$, we compute:

$$H(a,v) = -P(Y[a] = v)\log(P(Y[a] = v)), \tag{22}$$

where

$$P(Y[a] = v) = \frac{|\,[\![y|y \in Y, y[a] = v]\!]\,|}{|Y|}.$$

High entropy means the attribute is less useful for identification; low entropy means it is more useful. We therefore assign higher weight to low-entropy attributes:

$$w(a,v) = \frac{1}{H(a,v) + 10^{-16}}, \tag{23}$$

and build weighted datasets $\hat{Y}$ and $\hat{Z}$ by setting:

$$\hat{y}[a] = \frac{y[a]}{w(a, y[a]) + 10^{-16}}, \qquad \forall a \in A. \tag{24}$$

This yields a representation where more informative attributes contribute more to distance. We then define the nearest-neighbour pairs:

$$L_{\text{IDS}} = \left[\!\!\left[ (\hat{y}, \hat{z}) | \hat{y} \in \hat{Y}, \hat{z} = \underset{\hat{z}' \in \hat{Z}}{\arg\min}\, d_E(\hat{y}, \hat{z}') \right]\!\!\right],$$

and the subset:

$$R_{\text{IDS}} = \left[\!\!\left[ (\hat{y}, \hat{z}) | (\hat{y}, \hat{z}) \in L_{\text{IDS}}, \nexists \hat{y}' \in \hat{Y} \setminus \{\hat{y}\} : d_E(\hat{y}, \hat{y}') < d_E(\hat{y}, \hat{z}') \right]\!\!\right].$$

The final score is:

$$\text{IDS}(Y, Z) = \frac{|R_{\text{IDS}}|}{|\hat{Y}|}. \tag{25}$$

A value close to $1$ means many individuals are re-identifiable, while a value close to $0$ means most are not.

***Nearest Neighbour Adversarial Accuracy (NNAA)*** **[31]** NNAA captures whether a PP-SDG has memorised the training data, and thus how well an adversary can distinguish in- and out-of-distribution records from nearest neighbours.

The score is the gap between adversarial accuracy on the train set ($Y_{train}$) and the evaluation set ($Y_{eval}$). Adversarial accuracy combines two probabilities: (i) a real record is closer to a synthetic record than to another real one, and (ii) a synthetic record is closer to a synthetic record than to a real one. To estimate the first term, we define $L^Y_{\text{NNAA}}$ as in Equation 19 and $R^Y_{\text{NNAA}}$ as in Equation 20, then split them into train ($T$) and evaluation ($E$) parts:

$$\begin{aligned}
L^{T\to Z}_{\text{NNAA}} &= \left[\!\!\left[ (y, z) | (y, z) \in L^Y_{\text{NNAA}}, y \in Y_{train} \right]\!\!\right], \\
R^{T\to Z}_{\text{NNAA}} &= \left[\!\!\left[ (y, z) | (y, z) \in R^Y_{\text{NNAA}}, y \in Y_{train} \right]\!\!\right], \\
L^{E\to Z}_{\text{NNAA}} &= \left[\!\!\left[ (y, z) | (y, z) \in L^Y_{\text{NNAA}}, y \in Y_{eval} \right]\!\!\right], \\
R^{E\to Z}_{\text{NNAA}} &= \left[\!\!\left[ (y, z) | (y, z) \in R^Y_{\text{NNAA}}, y \in Y_{eval} \right]\!\!\right].
\end{aligned}$$

For the second term, we define:

$$\begin{aligned}
L^{Z\to T}_{\text{NNAA}} &= \left[\!\!\left[ (z, y) | z \in Z, y = \underset{y' \in Y_{train}}{\arg\min}\, d_E(z, y') \right]\!\!\right], \\
R^{Z\to T}_{\text{NNAA}} &= \left[\!\!\left[ (z, y) | (z, y) \in L^{ZT}_{\text{NNAA}}, \exists z' \in Z \setminus \{z\} : z' = \underset{z'' \in Z\setminus\{z\}}{\arg\min}, d_E(z, z'') < d_E(y, z) \right]\!\!\right], \\
L^{Z\to E}_{\text{NNAA}} &= \left[\!\!\left[ (z, y) | y \in Y_{eval}, z = \underset{z' \in Z}{\arg\min}\, d_E(z', y) \right]\!\!\right], \\
R^{Z\to E}_{\text{NNAA}} &= \left[\!\!\left[ (z, y) | (z, y) \in L^{ZE}_{\text{NNAA}}, \exists z' \in Z \setminus \{z\} : z' = \underset{z'' \in Z\setminus\{z\}}{\arg\min}, d_E(z, z') < d_E(y, z) \right]\!\!\right].
\end{aligned}$$

The adversarial accuracy on the training and evaluation sets is then:

$$\text{AA}_{train}(Y, Z) = 0.5 \cdot \left( \frac{|R^{TZ}_{\text{NNAA}}|}{|L^{TZ}_{\text{NNAA}}|} + \frac{|R^{ZT}_{\text{NNAA}}|}{|L^{ZT}_{\text{NNAA}}|} \right) \tag{26}$$

$$\mathrm{AA}_{eval}(Y,Z) = 0.5 \cdot \left( \frac{|R^{EZ}_{\mathrm{NNAA}}|}{|L^{EZ}_{\mathrm{NNAA}}|} + \frac{|R^{ZE}_{\mathrm{NNAA}}|}{|L^{ZE}_{\mathrm{NNAA}}|} \right). \tag{27}$$

Finally:

$$\mathrm{NNAA}(Y,Z) = \mathrm{AA}_{eval}(Y,Z) - \mathrm{AA}_{train}(Y,Z). \tag{28}$$

This score lies in $[-1,1]$: values near $1$ suggest strong overfitting, values near $-1$ suggest underfitting, and values near $0$ indicate little difference between train and evaluation. To align with our privacy convention, we use:

$$\mathrm{NNAA}'(Y,Z) = \max\{0, \mathrm{NNAA}(Y,Z).\} \tag{29}$$

So, $0$ means no advantage for the adversary, while $1$ means reliable re-identification.

***Hitting Rate (HitR)*** **[31]** HitR measures the chance of inferring whether a person contributed to the real dataset. It counts real-synthetic pairs that agree on categorical attributes and are close on continuous ones.

Let $A_c \subseteq A$ be the categorical attributes and $A_r \subseteq A$ the continuous ones. We define:

$$L_{\mathrm{HitR}} = [\![(y,z)|y \in Y, z \in Z, y[A_c] = z[A_c]]\!].$$

From these pairs, we keep the closest synthetic record to each real record under the Manhattan distance:

$$L^{'}_{\mathrm{HitR}} = \left[\!\!\left[(y,z)|(y,z) \in L_{\mathrm{HitR}}, z = \underset{z':(y,z')\in L_{\mathrm{HitR}}}{\arg\min} d_M(y[A_r], z'[A_r])\right]\!\!\right],$$

where $A_r$ contains the continuous attributes.

We define the threshold:

$$h(a) = \frac{\max\limits_{y\in Y} y[a] - \min\limits_{y\in Y} y[a]}{30}, \tag{30}$$

and then:

$$R_{\mathrm{HitR}} = \left[\!\!\left[(y,z)|(y,z) \in L^{'}_{\mathrm{HitR}}, \forall a \in A_r, |y[a] - z[a]| \leq h(a))\right]\!\!\right],$$

with:

$$\mathrm{HitR}(Y,Z) = \frac{|R_{\mathrm{HitR}}|}{|Y|}. \tag{31}$$

A value near $1$ means that many individuals may be inferred, while a value near $0$ means that the synthetic data stays well separated.

***Hidden Rate (HiddR)*** **[24]** HiddR also targets membership risk, but assumes each synthetic record comes from a real one.

For every synthetic individual $z$, it checks whether its nearest real neighbour is the same one that generated it. Let $g$ map each real record $y$ to its synthetic offspring, and let $\Phi^k$ be the dimensionality reduction function from Equation 35. We define:

$$L_{\mathrm{HiddR}} = \left[\!\!\left[(y,z)|y \in Y, z = \underset{z'\in Z}{\arg\min}\, d_M(\Phi^k(y), \Phi^k(z'))\right]\!\!\right],$$

where $d_M$ is the Minkowski distance. Then:

$$R_{\mathrm{HiddR}}\, [\![(y,z)|(y,z) \in L_{\mathrm{HiddR}}, z = g(y)]\!],$$

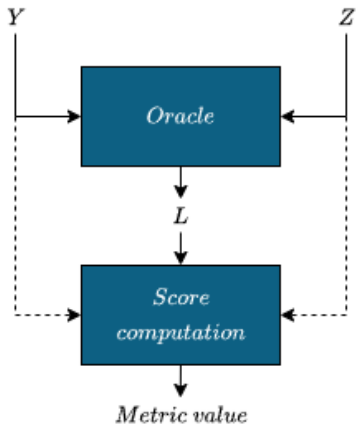



**Figure 5**: Metric computation in distance-based privacy metrics.

and:

$$\mathrm{HiddR}(Y, Z) = \frac{|R_{\mathrm{HiddR}}|}{|Z|}. \tag{32}$$

A score near $1$ means synthetic records stay very close to their sources, while a score near $0$ means they are more decoupled and varied.

### 4.1.4 Distance-based privacy metrics

The distance-based privacy metrics define risk through the distance between real and synthetic records and their nearest opposite-type neighbours, as shown in Figure 5. As in the threshold-based case, a neighbour set $L$ is first computed, but no filtering step is applied. All pairs contribute to the final score.

***Nearest Synthetic Neighbour Distance (NSND)*** **[39]** NSND measures how close each real record is to its nearest synthetic counterpart, and thus how strongly a synthetic record can be traced back to its source. We reuse the match set from Equation 14, so $L_{\mathrm{NSND}} = L_{\mathrm{CVP}}$. Let $d_{max}$ and $d_{min}$ be the maximum and minimum distances between real and synthetic records, as in Equation 15 and Equation 16. Then:

$$\mathrm{NSND}(Y, Z) = \frac{1}{|Y|} \sum_{(y,z)\in L_{\mathrm{NSND}}} \frac{d_E(y, z) - d_{min}}{d_{max} - d_{min}}. \tag{33}$$

Small distances make the score low, while large distances push it upward. To follow our privacy convention, we invert it:

$$\mathrm{NSND}'(Y, Z) = 1 - \mathrm{NSND}(Y, Z). \tag{34}$$

A value near $1$ means real records are very close to synthetic ones, while a value near $0$ means they are far apart.

***Nearest Neighbour Distance Ratio (NNDR)* [24]** NNDR measures re-identification risk through a distance-based comparison between real and synthetic records. It first embeds both sets in a lower-dimensional space, then compares nearest-neighbour distances. Let

$$\Phi^k : \mathbb{R}^{|A|} \to \mathbb{R}^k \tag{35}$$

be a mapping from a $|A|$-dimensional space to a $k$-dimensional Euclidean space, e.g. by using PCA, MCA or FAMD. Let also the set of matching elements $L_{\text{NNDR}}$ be the set of matches between synthetic individuals $z \in Z$ and their top-2 closest neighbours from $Y$ in the $k$-dimensional space, i.e.:

$$L_{\text{NNDR}} = \left[\!\!\left[ (y', y'', z) | z \in Z, y' = \underset{y \in Y}{\arg\min}\, d_E(\Phi^k(y), \Phi^k(z))\}, y'' = \underset{y \in Y \setminus \{y'\}}{\arg\min}\, d_E(\Phi^k(y), \Phi^k(z)) \right]\!\!\right].$$

We define $L'_{\text{NNDR}}$ as the set of triples in $L_{\text{NNDR}}$ where the closest real and synthetic records coincide in the embedded space:

$$L'_{\text{NNDR}} = \left[\!\!\left[ (y', y'', z) | (y', y'', z) \in L_{\text{NNDR}}, d_E(\Phi^k(y'), \Phi^k(z)) = 0 \right]\!\!\right].$$

We can now define NNDR. For each synthetic record $z$, the score is $1$ if the nearest neighbour in $Y$ is at distance $0$, and otherwise the ratio between the distances to the first and second nearest real neighbours. NNDR is the average of these scores, i.e.:

$$\text{NNDR}(Y, Z) = \frac{\left( |L'_{\text{NNDR}}| + \sum\limits_{(y', y'', z) \in L_{\text{NNDR}} \setminus L'_{\text{NNDR}}} \frac{d_E(y', z)}{d_E(y'', z)} \right)}{|Z|}. \tag{36}$$

A smaller ratio yields a larger score, whereas a larger ratio lowers it. For consistency with our privacy convention, we invert the score:

$$\text{NNDR}'(Y, Z) = 1 - \text{NNDR}(Y, Z). \tag{37}$$

Thus, $\text{NNDR}' = 1$ indicates that all real individuals are re-identifiable, while $\text{NNDR}' \approx 0$ indicates that none are.

***Distance to Closest Record (DCR)* [24]** DCR captures re-identification risk as a distance-based aggregate score. First, we build the set of real records and their nearest synthetic neighbours. As in NNDR, distances are computed in the same reduced space defined by $\Phi^k$ in Equation 35. Therefore, we define $C_{\text{DCR}}$ as:

$$L_{\text{DCR}} = \left[\!\!\left[ (y, z) | y \in Y, z = \underset{z' \in Z}{\arg\min}\, d_E(\Phi^k(y), \Phi^k(z')) \right]\!\!\right].$$

Next, DCR is the mean distance over these pairs:

$$\text{DCR}(Y, Z) = \frac{1}{|Y|} \sum_{(y,z) \in L_{\text{DCR}}} d_E(\Phi^k(y), \Phi^k(z)). \tag{38}$$

The original DCR [24] is unbounded and depends on the embedding used. To map it to $[0, 1]$, we can rescale it as follows:

$$\text{DCR}'(Y, Z) = 1 - \sigma\left(log\left(\text{DCR}(Y, Z)\right)\right), \tag{39}$$

where $\sigma$ is the sigmoid function. Here, values near $0$ indicate low re-identification risk, while values near $1$ indicate high risk.

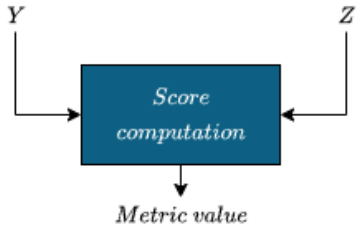



**Figure 6**: Metric computation in classification-based privacy metrics.

***Median Distance to Closest Record (MDCR)* [31]** MDCR estimates re-identification risk using a median-based distance ratio. First, we define the synthetic match set as in Equation 14, i.e. $L^Z_{\text{MDCR}} = L_{\text{CVP}}$. We then define the nearest-neighbour pairs within the real data:

$$L^Y_{\text{MDCR}} = \left[\!\!\left[ (y, y') | y \in Y, y' = \underset{y'' \in Y \setminus \{y\}}{\arg\min}\ d_E(y, y'') \right]\!\!\right].$$

MDCR is the ratio between the median real-to-synthetic distance and the median real-to-real distance:

$$\text{MDCR}(Y, Z) = \frac{med\left(\left[\!\!\left[ d_E(y, z) | (y, z) \in L^Z_{\text{MDCR}} \right]\!\!\right]\right)}{med\left(\left[\!\!\left[ d_E(y, y') | (y, y') \in L^Y_{\text{MDCR}} \right]\!\!\right]\right)}, \tag{40}$$

where $med(\cdot)$ denotes the median.

Like DCR, MDCR is scale-dependent and may take any value. We can therefore transform it into $[0, 1]$ with the proper direction:

$$\text{MDCR}'(Y, Z) = 1 - \sigma\left(\text{MDCR}(Y, Z)\right), \tag{41}$$

where $\sigma$ is the sigmoid function. A score close to $0$ means low re-identification risk, whereas a score close to $1$ means high risk.

### 4.1.5 Classification-based metrics

The classification-based privacy metrics cast privacy assessment as a classification problem, as illustrated in Figure 6. Unlike the previous categories, they do not rely on intermediate sets and compute the score directly from $Y$ and $Z$.

***Detection MLP (D-MLP)* [37, 39]** D-MLP measures how well a binary classifier separates synthetic from real data. In other words, it simulates an attacker training a shallow neural network to decide whether a record is real or synthetic. We first construct a labeled dataset $D$, assigning label $1$ to records in $Y$ and label $0$ to records in $Z$. AMLP is then trained on $D$ with $k$-fold cross-validation, and each split $D^i$ is divided into training and test sets ($D^i_{train}$ and $D^i_{test}$, respectively). The D-MLP score is then:

$$\text{D} - \text{MLP}(Y, Z) = 1 - ((\max\{0.5, \frac{1}{k}\sum_{i=1}^{k} AUC_i(D^i_{test}, D^i_{train})\} \cdot 2) - 1).$$

A score near $1$ means real and synthetic records are hard to tell apart, which is undesirable from a privacy standpoint because it reflects a similarity between the two sets. A score near $0$ means the classifier is close to random guessing ($AUC \approx 0.5$), suggesting better privacy.

***Membership Inference Risk (MIR)*** **[31]** MIR quantifies the risk of inferring whether an individual contributed to the real dataset. It simulates an attacker using a LightGBM classifier that sees synthetic data and tries to predict the membership of real records. To compute it, the real dataset $Y$ is split into $Y_{train}$ and $Y_{test}$, with $Y_{train}$ used to generate $Z$. For training, a dataset $D_{train}$ is built so that $Z, Y_{test} \in D_{train}$, with records in $Z$ labeled $1$ and records in $Y_{test}$ labeled $0$. The classifier is trained on this distinction. For evaluation, a dataset $D_{eval}$ is built so that $Z, Y_{train} \in D_{eval}$, with records in $Z$ labeled $0$ and records in $Y_{train}$ labeled $1$. MIR is then computed as the mean recall, i.e., the share of real individuals correctly identified as members. A score of $0$ means no real individuals were correctly classified as members, while a score of $1$ means all were, corresponding to a high privacy risk.

***DOMIAS MIA (DOMIAS-MIA)*** **[41]** DOMIAS estimates the risk that an adversary can determine whether an individual contributed to the real dataset. It does so by checking whether the PP-SDG overfits to specific records. Let the real dataset $Y$ be split into $Y_{train}$ and $Y_{test}$, where $Y_{test}$ serves as a reference set and $Y_{train}$ is used to generate $Z$. The adversary constructs a labeled set of individuals as follows:

$$\mathcal{L}_{\text{Domias}} = [\![(y, \ell(y)) | y \in Y]\!],$$

where the labels are defined by:

$$\ell(y) = \begin{cases} 1 & \text{if } y \in Y_{train} \quad \text{(member)} \\ 0 & \text{if } y \in Y_{test} \quad \text{(non-member).} \end{cases} \tag{42}$$

Let $X$ be the data space and $p_Y(X)$ the real distribution. The metric assumes that $Y_{train}$ is sampled by the probability distribution $p_Y(X)$ that denotes the training data used to generate $Z$, such that individuals in the synthetic dataset are sampled from a distribution ($p_Z$), i.e., $x \sim p_Z$. Two density estimates are then produced for the synthetic and real distributions. Let $\hat{p}_Z(\cdot)$ be an estimate of $p_Z(\cdot)$ fitted on $Z$, and let $\hat{p}_Y(\cdot)$ be an estimate of $p_Y(\cdot)$ fitted on $Y_{test}$

We then construct a score set in which each real record $y \in Y$ is paired with a label and a score based on the synthetic-to-real likelihood ratio $r(y)$:

$$S_{\text{Domias}} = \left[\!\!\left[ (r(y), \ell(y)) | (y, \ell(y)) \in \mathcal{L}, r(y) = \frac{\hat{p}_Z(y)}{\hat{p}_Y(y) + 10^{-12}} \right]\!\!\right],$$

where $10^{-12}$ is added for numerical stability. The score is then obtained by using $r(y)$ as a classification statistic. The adversary's performance is measured by the *AUC of the ROC (AUCROC)*:

$$\text{DOMIAS-MIA}(Y, Z) = AUCROC(S_{\text{Domias}}). \tag{43}$$

The original DOMIAS-MIA may already align with our privacy convention, but the optimal privacy value here is $0.5$, which corresponds to random guessing. To match our direction, we therefore redefine it as:

$$\text{DOMIAS-MIA}'(Y, Z) = 2 \cdot \max\{0.5, \text{DOMIAS-MIA}(Y, Z)\} - 1. \tag{44}$$

A score near $0$ means that members and non-members are difficult to distinguish, indicating low membership-inference risk, while a score near $1$ means the likelihood ratio separates them well, indicating high risk.

***Data Leakage MLP AIA (DLMLP-AIA)*** **[39]** DLMLP-AIA measures the risk of inferring sensitive attributes ($a_S \in A$) from key attributes $A_K \subseteq A \setminus \{a_S\}$. It assumes the adversary knows the key values of a real individual ($y[A_K]$) and has access to all attributes in the synthetic dataset $Z$.

| Metric | Papers | Applicable Datatypes $\mathbb{R}$ | $\sum$ | $\{0,1\}$ | Additional input | Metric class | Output Handling | Attack |
|---|---|---|---|---|---|---|---|---|
| *Categorical Zero CAP (ZCAP)* | [26, 37] | / | + | + | $A_K, A_S$ | Simulation-based | – | Reconstruction |
| *Categorical Generalized CAP (GCAP)* | [26, 27, 34, 35, 37, 40] | / | + | + | $A_K, A_S$ | | Inversion | |
| *Attribute Inference Risk (AIR)* | [10, 13, 25, 31, 37, 49, 53] | + | + | + | $A_K, A_S$ | | Inversion | |
| *Common Rows Proportion (CRP)* | [39] | / | + | + | – | | – | Re-identification |
| *Nearest Neighbour Distance Ratio (NNDR)* | [24, 31, 32, 54] | + | + | + | – | Distance-based | Inversion | |
| *Nearest Synthetic Neighbour Distance (NSND)* | [33, 39] | + | + | + | – | | Inversion | |
| *Distance to Closest Record (DCR)* | [24, 39, 54] | + | + | + | A | | Inversion, Normalised | |
| *Median Distance to Closest Record (MDCR)* | [17, 31] | + | / | + | – | | – | |
| *Authenticity (Auth)* | [2, 39] | + | + | + | – | Threshold-based | Inversion | |
| *Close Value Probability (CVP)* | [39] | + | / | + | Threshold | | – | |
| *Distant Value Probability (DVP)* | [39] | + | / | + | Threshold | | – | |
| *IDentifiability Score (IDS)* | [39, 50] | + | + | + | – | | – | |
| *Nearest Neighbour Adversarial Accuracy (NNAA)* | [25, 31, 44, 46, 48, 49] | + | / | + | – | | – | Tracing |
| *Hitting Rate (HitR)* | [5, 31, 33, 44] | + | + | + | A | | – | |
| *Hidden Rate (HiddR)* | [24] | + | + | + | – | | Inversion, Normalised | |
| *Detection MLP (D-MLP)* | [37] | + | + | + | – | Classification-based | – | |
| *DOMIAS MIA (DOMIAS-MIA)* | [19, 21, 39, 41] | + | + | + | – | | | |
| *Membership Inference Risk (MIR)* | [10, 13, 25, 31, 46, 47, 49, 53] | + | + | + | – | | – | |
| *Data Leakage MLP AIA (DLMLP-AIA)* | [39] | + | + | + | $A_K, A_S$ | | | Reconstruction |

To estimate this risk, we simulate an attacker that trains a MLP on $Z$ and uses it to predict sensitive attributes in $Y$. If $a_S$ is categorical, the model is a classifier; if it is continuous, it is a regressor. For each sensitive attribute $a_S \in A_S$, a training set $D_{train}[A_K]$ is built from $Z[A_K]$ with target $Z[a_S]$, and the MLP is trained to predict that attribute.

For evaluation, a test set $D_{test}[A_K]$ is built from $Y[A_K]$, using the same key attributes as input and $Y[a_S]$ as the target. The trained model then predicts $y[a_S]$ from $y[A_K]$, where $\hat{p}(y[A_K])$ denotes the predicted value for $y[a_S]$. A successful inference for one sensitive attribute is computed as:

$$S_{\text{DLMLP-AIA}}(y[a_S]) = \begin{cases} \frac{1}{|Y|} \sum_{y \in Y} \mathbf{1}\{\hat{p}(y[A_K]) = y[a_S]\}, & \text{if } a_S \text{ is categorical,} \\ \max(0, R^2(\hat{p}(y[A_S]), y[a_S])), & \text{if } a_S \text{ is continuous.} \end{cases}$$

DLMLP-AIA aggregates the scores over all sensitive attributes:

$$\text{DLMLP-AIA}(Y, Z) = \frac{1}{|A_S|} \sum_{a_S \in A_S} S_{\text{DLMLP-AIA}}(y[a_S]).$$

A score of $0$ means the adversary cannot infer sensitive attributes, i.e., low privacy risk. A score of $1$ means all sensitive attributes are accurately inferred, i.e., high privacy risk.

## 4.2 PrivEval

PrivEval is an interactive framework that helps data curators evaluate and explore privacy metrics. It is independent of the SDG method, requiring only a confidential dataset and a synthetic dataset as input. Currently, PrivEval supports CSV datasets, since this format is widely adopted. The framework consists of three main stages, shown in Figure 7.

**Generate dataset statistics.** First, PrivEval computes statistics and visualisations to help users understand the datasets. For categorical attributes, it reports distinct values and the mode. For numerical attributes, it computes the mean, standard deviation, minimum, maximum, and quartiles. In addition, both datasets are visualised using t-distributed stochastic neighbour embedding (t-SNE).

**Estimate privacy of the synthetic dataset.** PrivEval then computes privacy metrics that allow users to assess the privacy of the synthetic dataset. Because some metrics depend on adversarial background knowledge, the user must specify which attributes are sensitive.

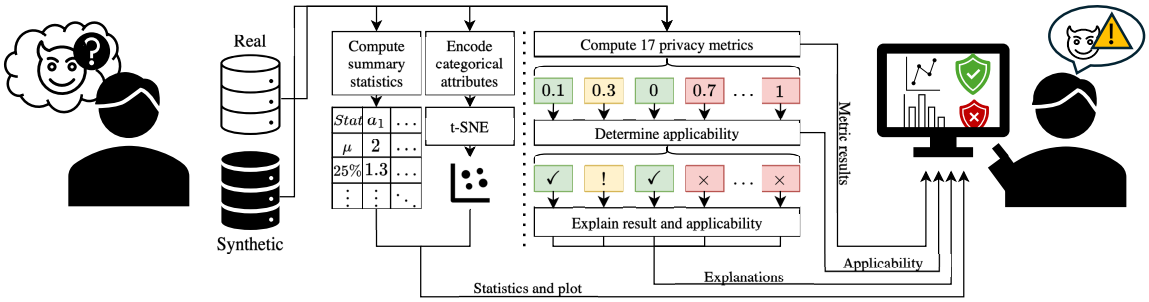



**Figure 7**: General structure of PrivEval.

PrivEval also introduces the notion of *shareability*, i.e., whether a synthetic dataset should be considered suitable for sharing based on a metric outcome. Three categories are defined: (1) shareable, (2) conflicts, and (3) unshareable. The conflict category denotes inconclusive results that require further investigation of specific records. Each metric is associated with rules that determine its shareability outcome.

To facilitate interpretation, PrivEval provides explanations of each metric and its implications. These include intuitive descriptions, dataset-based examples, visual illustrations, and technical details for expert users.

**Determine metric applicability.** PrivEval evaluates whether a privacy metric is applicable using dataset statistics and assumptions underlying the metric computation. Applicability reflects whether a metric can estimate risk reliably without being distorted by dataset characteristics. Examples include checking compatibility with continuous attributes or robustness to high-dimensional data.

PrivEval also detects confidential records that may be easily singled out through the synthetic dataset. For each confidential record, the framework identifies its three nearest synthetic neighbours and compares their distances to flag potentially revealing cases.

Based on these analyses, applicability is classified into three categories: *No assumption is compromised*, *Potential insufficient measurement of risk*, and *Unreliable privacy estimation*. For each metric, users can inspect why a category was assigned and how the privacy estimation may be improved.

# 5 Experiments

This section evaluates the use of privacy-preserving data in a federated medical analytics setting. Our experimental study is centered on the HEREDITARY ALS dataset and follows two objectives. Firstly, we assess whether state-of-the-art PP-SDG methods can produce tabular ALS data that retains useful statistical structure while reducing the disclosure risk. Secondly, we investigate how the synthetic dataset behaves in a downstream federated analytics task using k-means.

The purpose of our experiments is not to derive clinically validated ALS patient subgroups, nor to optimise the clustering model for medical interpretation. Instead, we use k-means as a task to study the effect of privacy-preserving technologies in the analytics workflow. We compare clustering results obtained from real and synthetic data, and from centralised and federated execution settings, in order to determine to what extent utility degrades.

We organise the section as follows: Section 5.1 introduces the HEREDITARY ALS dataset and the experimental data views used in the evaluation. Section 5.2 describes the generation and assessment of PP-SDG,

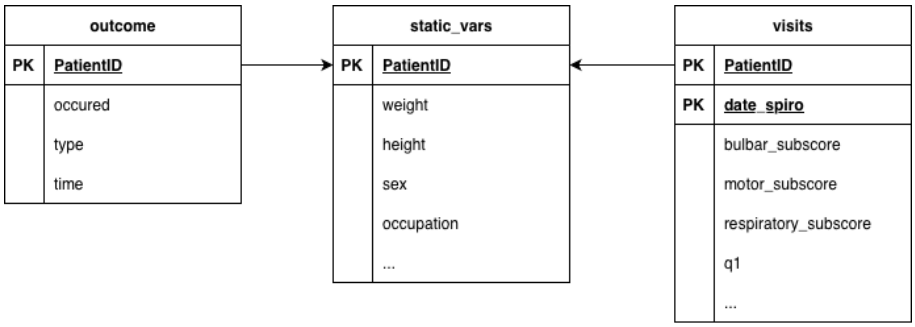



**Figure 8**: Schema of the relevant tables in the HEREDITARY ALS dataset.

including utility and privacy analyses. Section 5.3 presents the federated k-means implementation strategies we consider in the project. Section 5.4 reports the clustering results on real and synthetic data, comparing centralised and federated settings. Section 5.5 summarises the main findings and limitations of the current experimental evaluation.

## 5.1 The HEREDITARY ALS dataset and experimental data views

The experiments are based on the HEREDITARY ALS dataset, a clinical dataset derived from patient records and used as a representative medical scenario for privacy-preserving analytics. The dataset contains both static patient information and longitudinal clinical observations. Its schema, part of which is depicted in Figure 8, includes separate tables for patient attributes, disease outcomes, and clinical visits.

The *static-vars* table describes demographic and anamnesis information about each patient. It includes attributes such as weight and height, smoking habits, and information about previous surgeries. The *outcome* table describes the disease outcome, including whether death occurred, the type of outcome, and the corresponding time. The longitudinal data is described in the *visits* table, which records information collected during patient visits, including symptom-specific variables such as salivation and dyspnea, and summary scores such as motor and bulbar subscores. The dataset includes information from multiple sites, including Lisbon and Turin.

For the purpose of the experimental evaluation, the ALS data is used through two distinct data views: the synthetic data generation view and the clustering view. The distinction is necessary because PP-SDG and k-means have different requirements for the input representation.

The *synthetic data generation* view is used for PP-SDG. It is obtained by joining the *static-vars* and *outcome* tables on the shared primary key, *PatientID*. The patient identifier is then removed since it is not required for synthetic data generation and should not be reproduced in the synthetic output. Columns with more than 25% missing values are removed, and records with remaining null values are discarded. This preprocessing produces a compact tabular dataset with 29 columns and 1504 individuals, which is then split into training and test sets. This lower-dimensional representation is used to evaluate the PP-SDG considered methods: DPGAN, PATE-GAN, and PrivBayes.

The *clustering view* is used for the downstream k-means clustering evaluation. k-means requires a clean numerical matrix with a consistent feature schema across all datasets and execution settings. For this reason, the ALS data are transformed into a feature matrix that combines static patient information with information from the patients' last visit. Sparse or non-informative features are removed, continuous features are imputed and normalised, and categorical variables are expanded into binary indicator variables. In the federated setting, this preprocessing is applied locally at each site where possible, while global normalisation statistics

for continuous variables are computed through a federated aggregation of local sufficient statistics. The resulting clustering representation contains 105 features, including 23 continuous features and 82 binary features.

The two views are therefore complementary. The synthetic data generation view is kept compact in order to remain compatible with the current state of the art in privacy-preserving tabular synthesis, especially BN-based methods such as PrivBayes, whose scalability decreases as the number of attributes grows. The clustering view, instead, is designed to satisfy the requirements of k-means and to ensure that centralised, federated, real, and synthetic variants can be compared under the same feature schema.

When synthetic data is used in the downstream clustering experiments, we generate it from the synthetic data generation view and then post-process it into the clustering view. In particular, categorical encoding is applied after synthesis so that the resulting synthetic datasets match the numerical feature representation required by k-means. This organisation allows us to evaluate the synthetic data both directly, through utility and privacy metrics, and indirectly, through their effect on a downstream federated analytics task. The operational details of this transformation are described in Section 5.3.2.

## 5.2 Privacy-Preserving Synthetic Data Generation and assessment

In this section, we describe the generation and assessment of privacy-preserving synthetic data for the HEREDITARY ALS dataset. The goal is twofold: firstly, we assess whether state-of-the-art PP-SDG methods preserve the statistical properties of the original ALS data; secondly, we evaluate the privacy risk of the generated datasets using the privacy metrics introduced in Section 4.1.

### 5.2.1 Synthesis methods

We consider the three PP-SDG methods for tabular data described in Section 3.2: DPGAN, PATE-GAN, and PrivBayes. The experiments consider multiple values of the privacy budget $\varepsilon$, including values between $0.11$ and $5$ for the initial hyperparameter search. For DPGAN and PATE-GAN, the search also varied the batch size between $16$ and $128$ and the latent dimension between $8$ and $128$. For PATE-GAN, the number of teachers varied between $8$ and $64$.

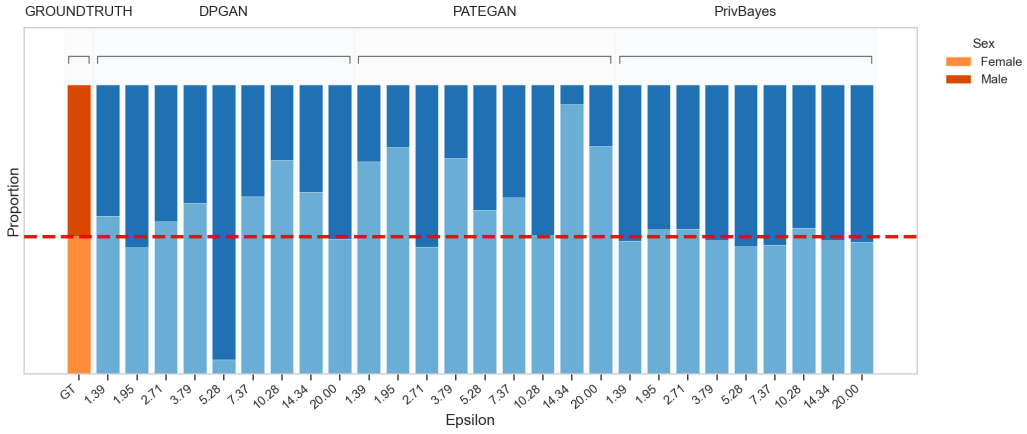
The methods are applied to the synthetic data generation view introduced in Section 5.1, obtained by joining the *static-vars* and *outcome* tables and applying the missing-value filtering described there.

### 5.2.2 Utility Assessment

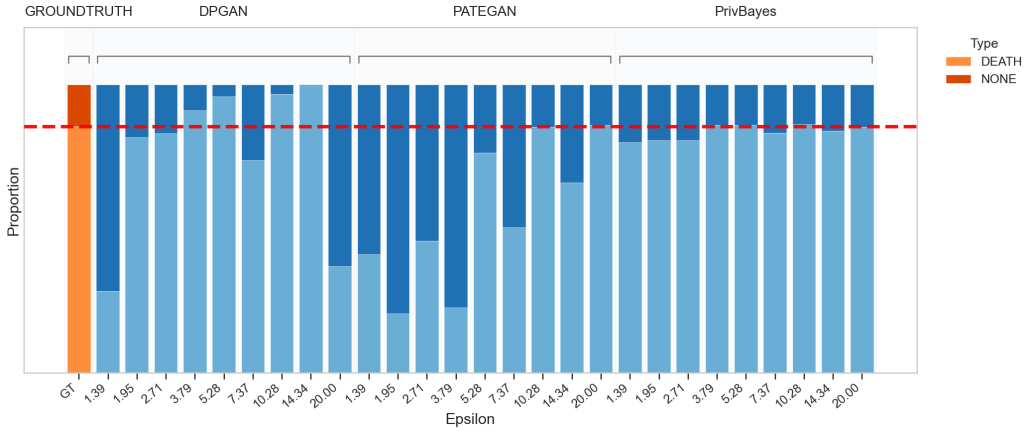
We assess the utility of the generated synthetic datasets by comparing their statistical properties with those of the original ALS data, through the preservation of attribute distributions and the *Synthetic Ranking Agreement (SRA)* score. SRA, initially introduced by Yoon et al. [51], measures whether the relative ranking of machine learning models trained on synthetic data is consistent with the ranking obtained when training on real data.

The comparison of binary attribute distributions showed that DPGAN and PATE-GAN often produced synthetic datasets with substantial deviations from the ground truth. The distributions of the data synthesized by PrivBayes are the closest to the distributions of original data across multiple values of $\varepsilon$, suggesting that, for the tabular ALS view considered here, PrivBayes is more reliable than the GAN-based methods in preserving marginal distributions. This behaviour is exemplified in Figure 9, which reports the distributions of the *sex* and *type* attributes for synthetic datasets generated with different methods and privacy budgets. We also observed similar behaviours for the other attributes.

The SRA results confirm this observation. As shown in Figure 10, PrivBayes achieved higher utility than DPGAN and PATE-GAN across multiple privacy-budget values. This indicates that, in addition to better




(a)




(b)

**Figure 9**: The distribution of the $sex$ and $type$ attribute in the synthetic datasets generated by DPGAN, PATE-GAN and PrivBayes at various levels of privacy ($\varepsilon$), compared to the ground truth (dashed line).

preserving selected attribute distributions, PrivBayes also better supports downstream predictive behaviour as measured by the relative ranking of models.

### 5.2.3 Privacy assessment

We then assess the privacy properties of the synthetic datasets using the privacy metrics described in Section 4.1. The purpose of this analysis is not to replace the formal guarantees provided by the underlying PP-SDG mechanisms, but to provide an empirical view of the disclosure risks that may remain in the generated datasets. The assumption at the basis of the following analysis is that the considered implementations are correctly configured and correctly implement their intended DP mechanisms. Under this assumption, DPGAN, PATE-GAN, and PrivBayes provide the formal privacy guarantees associated with their respective synthesis procedures. The empirical privacy metrics computed in this section should therefore not be read as replacing these guarantees, but as complementing them with an attack-oriented assessment of residual

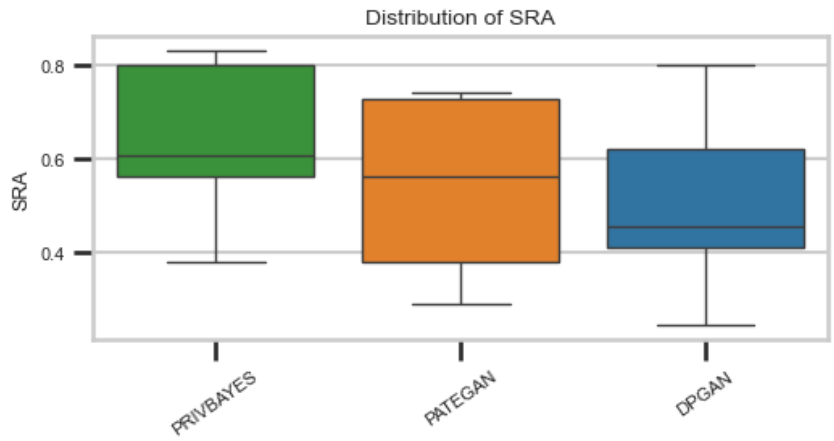



**Figure 10**: The distribution of the SRA metric score, across multiple $\varepsilon$-values (higher $\rightarrow$ better).

disclosure risk.

Overall, PrivBayes shows a slight decrease in measured privacy risk when compared with DPGAN and PATE-GAN. At first sight, the lower measured risk of PrivBayes may appear counter-intuitive, since PrivBayes also achieved higher utility. However, higher utility on the aspects measured in this experiment does not necessarily imply higher empirical privacy risk. A possible explanation is that the noise injected by PrivBayes has a lower impact on the marginal distributions and model-ranking behaviour considered in the utility analysis than the noise injected by the GAN-based methods, while still limiting the forms of record-level proximity and inference captured by the selected privacy metrics. In other words, the stronger utility of PrivBayes does not contradict its DP guarantee, nor does it necessarily imply a correspondingly higher measured privacy risk under the metrics considered.

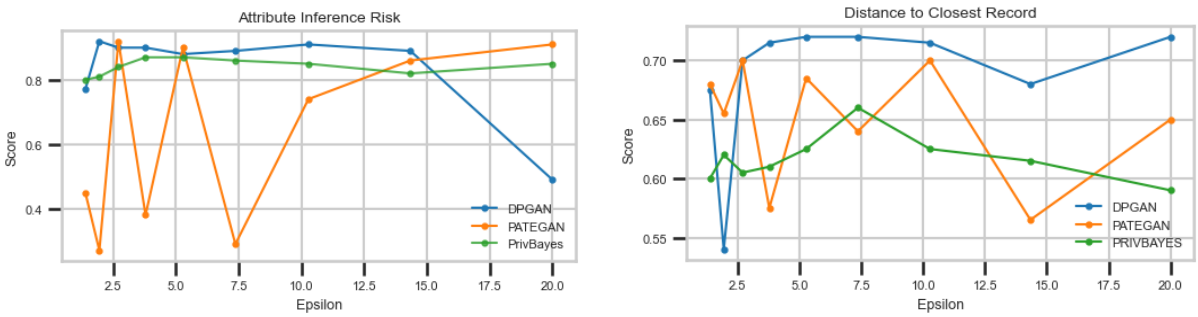
While we measured all the privacy metrics, we exemplify their behaviour through AIR and DCR, illustrated in Figure 11. For lower values of $\varepsilon$, the privacy scores vary substantially across generated datasets, reflecting the effect of noise in the synthesis process. However, PrivBayes is not significantly worse than DPGAN or PATE-GAN with respect to AIR. For DCR, PrivBayes even appears to produce datasets that are more private across multiple values of $\varepsilon$, while also improving utility. These results suggest that, for the current ALS tabular view, PrivBayes provides the best trade-off among the considered methods.

### 5.2.4 Observed Limitations

Despite its better performance in the utility and privacy analyses, PrivBayes also presents some limitations. Since PrivBayes relies on BN, continuous variables must be discretised before synthesis. In the implementation used in this deliverable, discretisation is performed through equi-width bins. This can reduce utility because variation within each bin is lost, and the resulting synthetic distributions may fail to capture relevant continuous patterns in the original data.

A second limitation concerns the learning of the BN. PrivBayes uses a score-based, privacy-preserving structure learning algorithm. The noise injected during this process can distort the learned dependencies between attributes, weakening the relationships that are important for downstream analyses. This limitation is particularly relevant for the ALS dataset, where clinically meaningful patterns may depend on combinations of demographic, genetic, and longitudinal variables.

Finally, we observe that sampling from the noisy BN can produce implausible records. In particular, some




(a) The distribution of the AIR metric score, as elaborated in Section 4.1.2 (Lower → better).

(b) The distribution of the DCR metric score, as elaborated in Section 4.1.4 (Lower → better).

**Figure 11**: Privacy metric results for the synthetic datasets generated by DPGAN, PATE-GAN and PrivBayes at various levels of privacy ($\varepsilon$).

generated patients contain genetic mutation combinations that are not clinically plausible. This observation highlights a general limitation of current PP-SDG methods when applied to complex medical data: preserving marginal distributions is not sufficient if the generator fails to preserve domain constraints and higher-order dependencies.

These limitations are important for interpreting the downstream k-means experiments. Although PrivBayes provides the strongest synthetic data utility among the methods considered, the generated data may still fail to preserve the cluster-conditional structure required by k-means. The downstream clustering evaluation, therefore, provides a complementary test of whether synthetic data are useful not only according to direct privacy and utility metrics, but also for federated analytics tasks.

## 5.3 Federated k-means as the downstream analytics task

We now introduce the downstream analytics task used to assess the practical utility of privacy-preserving synthetic data in a federated medical setting. A major challenge addressed by HEREDITARY is that clinically relevant evidence is often distributed across multiple hospitals, while local analyses performed at individual sites may be insufficient to derive robust conclusions. In our scenario, we are interested in grouping patients with similar clinical profiles through clustering across hospitals, while complying with the legal and privacy constraints discussed in Section 2.1.

### 5.3.1 Federated k-means Algorithm

We adopt a federated formulation of k-means proposed by Garst and Reinders [20] and shown in Algorithm 1, using the notation from Table 1. The method is designed to exploit distributed datasets without pooling raw records in a central location.

The algorithm proceeds in rounds. Each round alternates between a client-side phase and a server-side phase. During the client-side phase, each client runs a local k-means update on its own data partition, computes local centroids and the corresponding sample counts, removes empty clusters, and sends only the resulting aggregate information to the server. During the first round, the local centroids are initialised using k-means++ [3].

**Algorithm 1** Federated k-means [20]

**Require:** $K_g$

**Init:**
  **on each client** $i \in \mathcal{N}$ **do:**
    $K_i \leftarrow K_g$
    $S_i, C_i \leftarrow$ KMEANS++_INIT$(X_i, K_i)$ ▷ get cluster means using k-means++ initialisation
    send $S_i, C_i$ to server
**For each round** $r$ **do:**
  **On server do:**
    $C_l \leftarrow [C_1 \mid C_2 \mid \cdots \mid C_{|\mathcal{N}|}]$ ▷ concatenate local cluster means
    $S_l \leftarrow [S_1 \mid S_2 \mid \cdots \mid S_{|\mathcal{N}|}]$ ▷ collect sample counts per cluster
    $C_g \leftarrow$ KMEANS$(C_l, K_g, \textit{weights} = S_l)$ ▷ obtain new global clusters
    send $C_g$ to all clients
  **On each client** $i \in \mathcal{N}$ **do:**
    $C_i \leftarrow C_g$
    $S_i \leftarrow$ KMEANS_ASSIGN$(X_i, C_i)$ ▷ determine empty clusters
    $C_i \leftarrow C_i[s \neq 0$ for $s \in S_i]$ ▷ drop empty global clusters
    $K_i \leftarrow \text{size}(C_i)$
    $S_i, C_i \leftarrow$ KMEANS$(X_i, K_i, \textit{init} = C_i)$ ▷ run local k-means from global centroids
    send $S_i, C_i$ to server

During the server-side phase, the server reconciles the centroids received from the clients into a global configuration. The server does not access raw patient records. Instead, it performs a weighted k-means step using the local centroids as input points and the corresponding sample counts as weights. The resulting global centroids are then broadcast back to the clients and used as the starting point for the next local update. This makes the procedure suitable for distributed medical settings, where the objective is to share aggregate information rather than sensitive records.

Table 1: Notation used for federated k-means

| Symbol | Description |
|---|---|
| $X_i$ | Data on client $i$ |
| $K_g$ | Global number of clusters |
| $K_i$ | Number of clusters on client $i$ |
| $C_g$ | Global cluster means |
| $C_i$ | Cluster means of client $i$ |
| $S_i$ | Number of samples assigned to each cluster on client $i$ |
| $\mathcal{N}$ | Set of clients |

### 5.3.2 Data Preparation for k-means

Section 5.1 introduced the two experimental views used in this evaluation. We now describe the operational preprocessing steps used to construct the clustering view for k-means. This view is a numerical representation of the HEREDITARY ALS dataset with a consistent feature schema across centralised, federated, real, and synthetic variants.

Each federated client independently preprocesses its local partition of the HEREDITARY ALS dataset through a standardised pipeline before any inter-client communication takes place. The pipeline produces a feature matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times d}$ for partition $k$. When needed for auxiliary evaluation, the corresponding outcome vector $\mathbf{y}^{(k)} \in \mathbb{R}^{n_k}$ is kept separate and is not used by k-means. Throughout this section, we use $\mathcal{C} \subset \{1, \ldots, d\}$ to denote the index set of continuous features and $\mathcal{B} = \{1, \ldots, d\} \setminus \mathcal{C}$ to denote the binary features, with $|\mathcal{C}| = 23$ and $|\mathcal{B}| = 82$.

The dataset comprises records from three independent hospital sites. Each site constitutes one federated partition, and no raw data is exchanged between sites or with the server at any point.

Three genetic mutation fields store free-text allele notations and are mapped to binary indicators. The `SOD1` field is positive if and only if the variant `c.281G>T (p.(Gly94Val))` is present, which is the sole variant classified as pathogenic in the reference database. The `C9orf72` field is positive if and only if a hexanucleotide repeat expansion is documented [12]. The `TARDBP` field is positive if and only if the variant `c.1144G>A, p.(Ala382Thr)` is present. All other string values are mapped to negative, while missing values are propagated before the subsequent preprocessing steps.

Sparse and non-informative features are then removed. Features with a missing-value rate exceeding 50% are excluded. This threshold eliminates 51 variables spanning biochemical laboratory panels, smoking history, pre-onset trauma events, surgical interventions, and spirometry. The threshold and the resulting feature set were determined from a prior analysis of the centralised dataset. In a fully privacy-preserving deployment, this selection would be established through federated analytics.

Missing values are handled in a type-aware manner. Continuous features that are not removed due to missingness are imputed using the local training-set mean. Categorical features are expanded into binary indicator variables. Twenty-six categorical features, seven fully observed and nineteen partially missing, are encoded into binary columns. For a feature $f$ with category set $\mathcal{V}_f$, the encoding produces $|\mathcal{V}_f|$ columns $z_{i,f,v} = \mathbf{1}[x_{if} = v]$, with $v \in \mathcal{V}_f$. To guarantee column-set consistency across partitions, a shared category vocabulary derived from the complete multi-site training data is used when available. Using local vocabularies independently would risk producing incompatible feature spaces and would invalidate federated aggregation.

Standardisation of continuous features is intentionally deferred from local preprocessing. Applying local statistics $(\mu_j^{(k)}, \sigma_j^{(k)})$ would produce site-specific scales that are inconsistent under non-independent and identically distributed data distributions, biasing the global centroid aggregation in the federated k-means step. Instead, each client communicates its local sufficient statistics for every continuous feature $j \in \mathcal{C}$:

$$s_j^{(k)} = \sum_{i=1}^{n_k} x_{ij}^{(k)}, \qquad q_j^{(k)} = \sum_{i=1}^{n_k} \left(x_{ij}^{(k)}\right)^2.$$

The server aggregates these statistics across all $K$ partitions to obtain global statistics:

$$\bar{\mu}_j = \frac{\sum_{k=1}^{K} s_j^{(k)}}{N}, \qquad \bar{\sigma}_j = \sqrt{\frac{\sum_{k=1}^{K} q_j^{(k)}}{N} - \bar{\mu}_j^2}, \tag{45}$$

where $N = \sum_k n_k$ is the total number of training samples. Zero-variance features are handled by setting

$\bar{\sigma}_j = 1$. Each client then applies the global $z$-score transform:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}, \qquad j \in \mathcal{C}. \tag{46}$$

Binary features $j \in \mathcal{B}$ are left unchanged, preserving their indicator semantics. This federated $z$-score normalisation protocol ensures that the feature scales used during clustering reflect the global data distribution rather than any individual site.

In the non-federated runs, the same normalisation is applied centrally and does not require the two-step aggregation of sufficient statistics. However, the resulting normalisation of the federated and centralised versions of the dataset is equivalent, since the federated protocol computes the same global means and standard deviations from aggregated sufficient statistics.

When synthetic data are used for k-means, the preprocessing follows the same final feature schema. However, the preprocessing before *Synthetic Data Generation (SDG)* differs from the clustering preprocessing. PrivBayes relies on BNs, and its scalability decreases as the number of attributes grows. For this reason, categorical encoding is not performed before synthesis. Instead, the generated synthetic table is post-processed after SDG to match the clustering feature schema, including the categorical encoding. This makes synthetic training data compatible with real preprocessed test data for k-means fitting and evaluation.

### 5.3.3 Federated Implementation Strategies

We considered two strategies for implementing federated k-means. The first follows the current state of the art in federated learning and is based on Flower. The second is a SPARQL-based design that builds on the architecture presented in Deliverable D3.2 [4] and aims to align federated analytics directly with the HDN.

**Implementation through Flower.** Flower is a federated learning framework built around a separation between server-side and client-side logic. In Flower, the server is represented by a `ServerApp`, which orchestrates client selection, configuration, and aggregation, while each client is represented by a `ClientApp`, which executes the local computation close to the data holder. Flower also supports dynamic configuration and custom messages, and provides built-in DP mechanisms, including central and local DP through clipping and noise addition.

This implementation is well suited for evaluating the federated k-means workflow in a controlled experimental setting. It allows us to compare centralised and federated execution on real data, and to evaluate how private synthetic datasets affect the resulting clustering.

**SPARQL-based implementation design.** The Flower-based implementation introduces practical limitations in the context of the HDN, as described in Deliverable D3.2. Adopting Flower requires deploying and maintaining an additional software stack on top of the existing data infrastructure, which increases operational complexity and expands the trusted runtime environment. Moreover, the local data access pattern in Flower is imperative and runtime-driven: client-side code typically loads the local partition and performs the local computation in memory. This is natural for many machine-learning workloads, but it is less directly aligned with the SPARQL-centred architecture of the HDN.

Therefore, we investigated an alternative implementation strategy that uses SPARQL engines as the execution substrate for client-side computation. Instead of deploying a separate federated-learning runtime, local data holders expose their data through SPARQL endpoints in the HDN, and the client-side k-means operations are expressed as declarative queries. In this design, local data remain under the control of the respective data holders, while the server receives only aggregate outputs such as centroid coordinates and cluster sizes.

This architecture is more closely aligned with the HDN because it keeps local computation within the data store and avoids moving raw patient data across system boundaries. It also provides a clearer separation of responsibilities: clients compute local aggregates, the server combines them, and the surrounding utility layer supports simulation, query execution, and visualisation. At the current stage, demonstrations have been tested on the Oxigraph local engine, but it should work on any SPARQL-compliant query engine, facilitating future integration with the HDN.

The SPARQL-based design preserves the logic of federated k-means while replacing imperative client code with declarative query execution. The server acts as the orchestrator and stores intermediate states, such as the current global centroids. This assumption is necessary because SPARQL endpoints may be read-only and may not support `INSERT` or `UPDATE`. Client-side routines are therefore rewritten as SPARQL-based computations and sent to the endpoint.

Concretely, the prototype identifies three client-side functions that must be expressed in SPARQL: initialising local centroids, assigning local points to received centroids, and running one local k-means update from a given set of centroids. The assignment and update steps can be expressed by injecting the current centroids into the query using `VALUES`, computing distances between local points and centroids, and aggregating the results by the nearest centroid. Since SPARQL does not provide a native `argmin` operator, the nearest-centroid computation is implemented through subqueries: an inner query computes the minimum distance for each point, and an outer query retrieves the centroid corresponding to that minimum distance.

The most challenging operation is k-means++ initialisation. k-means++ requires randomised sampling with probabilities proportional to the squared distance from the already selected centroids. This is difficult to express efficiently in SPARQL, because SPARQL does not provide cumulative-sum aggregates. Moreover, centroids have to be injected multiple times (one time for each subquery) and thus cannot be generated on the fly, as repeated calls to `RAND()` would result in different values across query fragments. Therefore, a SPARQL-based implementation requires workarounds such as self-joins, server-generated random thresholds, or alternative initialisation strategies.

Several options can mitigate these limitations. A first possibility is to replace k-means++ with simpler random initialisation, which is compatible with the query model but may slow convergence or increase sensitivity to local minima. A second option is to perturb selected centroids by adding noise, for instance, Laplacian noise, or using a barycenter, namely the mean of the first $X$ points closest to that centroid. This may improve privacy but can also alter the geometry of the centroids and affect clustering quality. The parameter $X$ should be chosen according to privacy considerations and in accordance with the policies enforced by PRISM. Finally, random values could be generated on the server and injected into the query, making it possible to reuse the same centroids across query fragments, but this would require longer queries and would shift responsibility for randomness to the server.

At this stage, the SPARQL-based implementation should be understood as a design and prototyping activity rather than as a complete experimental evaluation. The available quantitative clustering results in this deliverable are based on the Flower implementation. The SPARQL-based direction remains important for the next release because it provides a path towards tighter integration between federated analytics, ontology-mediated data access, and privacy-aware query governance in the HDN.

## 5.4 Clustering results on real and synthetic data

This section reports the clustering results obtained with the Flower-based federated k-means implementation described in Section 5.3. The goal is to assess how federation and PP-SDG affect the clustering structure of the HEREDITARY ALS dataset. The focus is not on evaluating the clinical quality of the resulting clusters, but on measuring how closely the cluster assignments obtained under privacy-preserving settings approximate those obtained from the original data.

Table 2: Pairwise clustering agreement between synthetic datasets and the two reference clusterings. *als*: federated real data, representing the plausible real-data scenario; *u_als*: unfederated real data, representing the ideal centralised reference; $\varepsilon$: differential-privacy budget. Best value among synthetic datasets per column is shown in bold.

| | | vs. Federated ALS *als* | | | | vs. Unfederated ALS *u_als* | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | $\varepsilon$ | **ARI** | **AMI** | **NMI** | **FM** | **ARI** | **AMI** | **NMI** | **FM** |
| Fed. ALS | – | – | – | – | – | 0.798 | 0.722 | 0.725 | 0.858 |
| Federated synthetic | 1.0 | 0.160 | 0.239 | **0.252** | 0.548 | 0.205 | **0.301** | **0.313** | 0.582 |
| Federated synthetic | 5.0 | 0.204 | 0.212 | 0.224 | 0.550 | 0.251 | 0.299 | 0.310 | **0.586** |
| Federated synthetic | 10.0 | 0.213 | 0.209 | 0.219 | 0.549 | 0.252 | 0.270 | 0.279 | 0.578 |
| Unfederated synthetic | 1.0 | 0.215 | 0.240 | 0.249 | **0.562** | 0.237 | 0.275 | 0.283 | 0.581 |
| Unfederated synthetic | 5.0 | 0.144 | 0.170 | 0.180 | 0.413 | 0.144 | 0.188 | 0.198 | 0.417 |
| Unfederated synthetic | 10.0 | **0.238** | **0.241** | 0.249 | 0.475 | **0.253** | 0.258 | 0.265 | 0.489 |

### 5.4.1 Comparison setting

We compare four clustering settings. The first setting, denoted as *u_als*, corresponds to k-means on the unfederated real ALS data and is used as the ideal centralised reference. The second setting, denoted as *als*, corresponds to federated k-means on the real ALS data and represents the plausible real-data scenario in which patient records remain distributed across sites. The third setting, denoted as *u_pals*, applies k-means to private synthetic data in an unfederated setting. The fourth setting, denoted as *pals*, applies federated k-means to private synthetic data.

For the synthetic-data settings, we consider multiple values of the privacy budget $\varepsilon$, namely $\varepsilon \in \{1, 5, 10\}$. The comparison is based on pairwise agreement between cluster assignments, using standard clustering agreement metrics: *Adjusted Rand Index (ARI)*, *Adjusted Mutual Information (AMI)*, *Normalised Mutual Information (NMI)*, and *Fowlkes–Mallows score (FM)*. These metrics allow us to quantify how close each privacy-preserving variant is to the two real-data references: the federated real-data clustering *als* and the unfederated real-data clustering *u_als*.

### 5.4.2 Quantitative agreement results

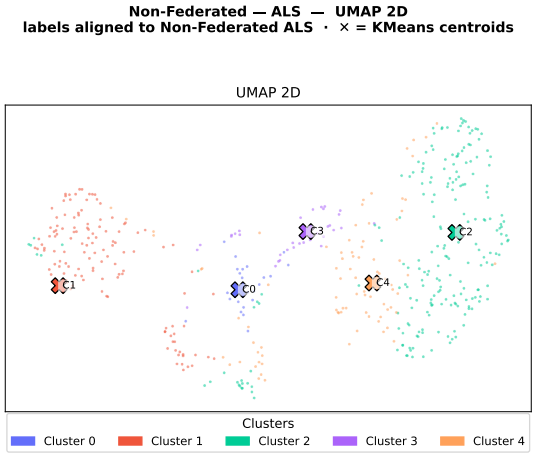
Table 2 reports the pairwise agreement between the synthetic-data clusterings and the two real-data references. The agreement between *als* and *u_als* is high, with an ARI of $0.798$ and an FM score of $0.858$. This indicates that federated execution over real data preserves most of the clustering structure obtained in the corresponding centralised real-data setting.

The results are substantially different for the private synthetic datasets. Across both federated and unfederated synthetic variants, ARI remains much lower, ranging approximately between $0.14$ and $0.25$. This indicates that the dominant source of degradation is not the federated execution itself, but the use of privacy-preserving synthetic data.

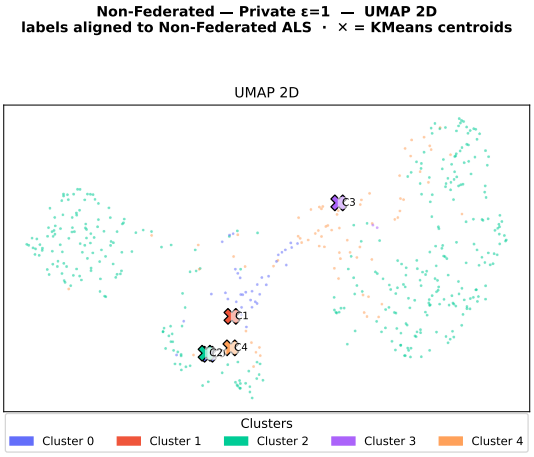
### 5.4.3 Visual analysis

Figure 12 and Figure 13 complement the quantitative results reported in Table 2. They show the cluster centroids obtained under the different real and synthetic data settings, projected onto lower-dimensional spaces

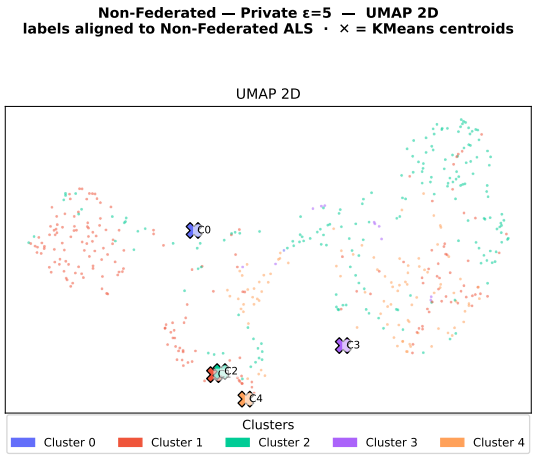
for visual inspection. In the real-data settings, the centralised and federated centroids occupy broadly consistent regions of the projected feature space. This supports the quantitative observation that federation introduces limited distortion when compared with the centralised real-data reference.



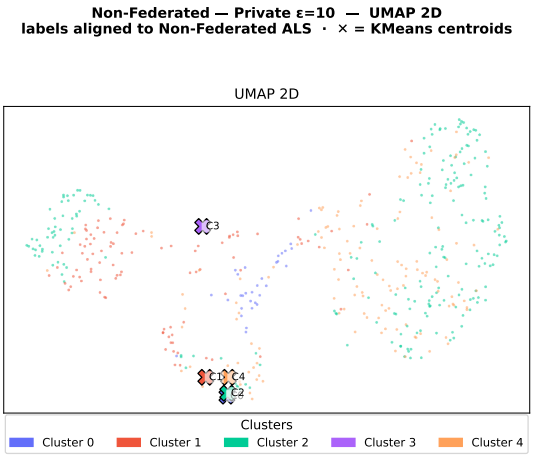

(a) HEREDITARY ALS




(b) Private data $\epsilon = 1.0$




(c) Private data $\epsilon = 5.0$



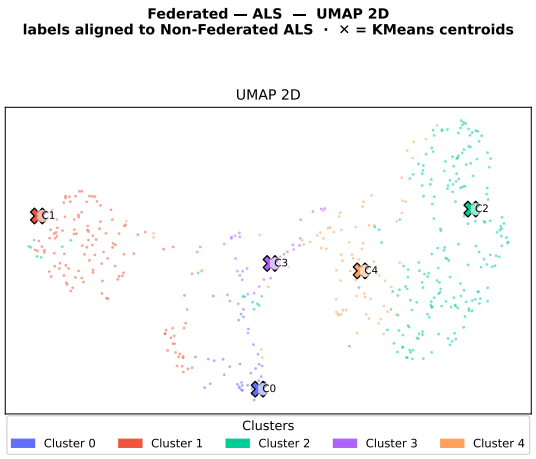

(d) Private data $\epsilon = 10.0$

**Figure 12**: Unfederated k-means

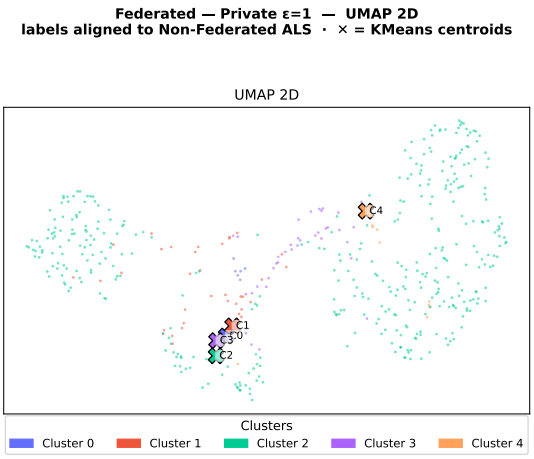
By contrast, the centroids obtained from synthetic data tend to collapse towards a more central region of the projection. This behaviour is visible in both the unfederated and federated synthetic settings, and is largely consistent across the considered privacy budgets. The visual evidence therefore supports the interpretation that the main loss of utility is caused by the PP-SDG step rather than by the federated execution of k-means.

### 5.4.4 Interpretation

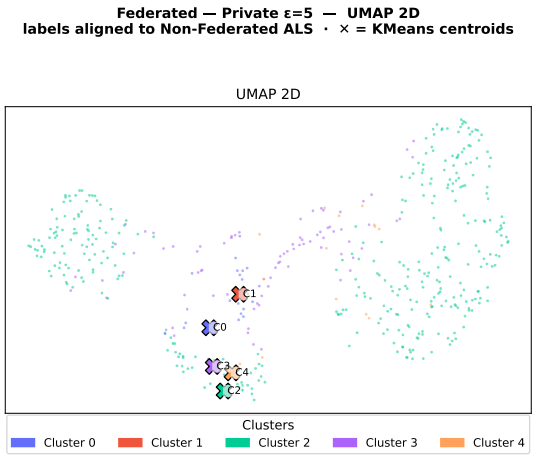
The results reveal two distinct effects: the cost of federation and the cost of PP-SDG.



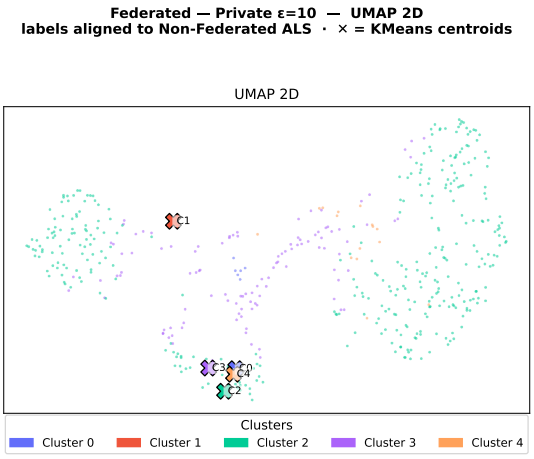

(a) HEREDITARY ALS




(b) Private data $\epsilon = 1.0$




(c) Private data $\epsilon = 5.0$.




(d) Private data $\epsilon = 10.0$.

**Figure 13**: Federated k-means

First, federation itself appears to be largely benign in this experiment. The high agreement between *als* and *u_als* suggests that the federated k-means protocol can approximate the clustering obtained from centralised real data. This is consistent with the design of the algorithm, where clients share centroids and cluster sizes rather than raw patient records.

Second, PP-SDG introduces a much stronger degradation in the downstream clustering structure. All private synthetic variants, both federated and unfederated, show substantially lower agreement with the real-data references. This is also visible in Figure 12 and Figure 13: whereas the real-data centroids remain distributed across distinct regions of the projected space, the synthetic-data centroids concentrate in a smaller central area. This suggests that the synthetic data generation step does not preserve the cluster-conditional structure required by k-means, even when it preserves some marginal distributions and achieves favourable results under direct utility and privacy metrics.

Third, increasing the privacy budget $\varepsilon$ from $1$ to $10$ does not lead to a consistent improvement in clustering agreement. For example, the ARI against the unfederated real-data reference remains in a narrow range for the federated synthetic variants, moving from $0.205$ at $\varepsilon = 1$ to $0.251$ at $\varepsilon = 5$ and $0.252$ at $\varepsilon = 10$. A similar absence of monotonic improvement can be observed for the unfederated synthetic variants. This indicates that the bottleneck is not only the amount of noise introduced by the privacy mechanism, but also the ability of the synthetic data model to preserve the multivariate structure relevant for clustering.

Finally, federation does not appear to introduce systematic additional distortion on top of synthetic data generation. The federated and unfederated synthetic variants achieve comparable agreement scores across privacy budgets. This reinforces the main conclusion of the downstream evaluation: in this task, the main loss of utility is caused by the synthetic data generation step rather than by the federated execution of k-means.

Overall, Table 2, Figure 12 and Figure 13 lead to the same conclusion. Federated k-means over real data approximates the centralised real-data clustering reasonably well, while current privacy-preserving synthetic data do not preserve the multivariate cluster structure needed by the downstream task.

These results qualify the findings of Section 5.2. PrivBayes provides the strongest utility among the considered PP-SDG methods when assessed through marginal distributions and SRA, and it does not show a higher empirical privacy risk under the selected privacy metrics. However, the downstream clustering evaluation shows that these favourable direct metrics do not necessarily imply the preservation of the cluster structure required by k-means. This motivates future work on task-aware synthetic data generation and on tighter integration between privacy-preserving synthesis, downstream analytics requirements, and the HDN governance mechanisms.

## 5.5 Summary of findings

The experiments reported in this section provide an initial validation of the privacy-preserving analytics workflow presented in this deliverable. The evaluation adopts two complementary perspectives. First, we assessed whether PP-SDG methods can generate useful synthetic variants of the HEREDITARY ALS data. Second, we evaluated whether such synthetic data can support a downstream federated analytics task, using k-means clustering as a representative example.

The direct assessment of synthetic data shows that PrivBayes provides the most promising results among the methods considered in this release. Compared with DPGAN and PATE-GAN, PrivBayes better preserves selected marginal distributions and achieves stronger utility according to the SRA score. At the same time, under the assumption that all implementations correctly realise their intended DP mechanisms, the empirical privacy metrics do not indicate a higher residual disclosure risk associated with PrivBayes. This suggests that, for the compact tabular ALS view used in the synthesis experiments, PrivBayes offers the best observed trade-off between utility and measured privacy risk.

However, the downstream clustering evaluation qualifies this conclusion. Although PrivBayes performs well under direct utility and privacy metrics, the synthetic datasets do not preserve the multivariate cluster

structure required by k-means. This is reflected both in the low agreement scores between real-data and synthetic-data clusterings and in the visual analysis of the projected centroids. The synthetic-data centroids tend to concentrate in a central region of the projected space, rather than reproducing the more dispersed structure observed in the real-data clusterings.

By contrast, the federated execution of k-means over real data introduces only limited distortion. The agreement between the federated real-data clustering and the corresponding centralised real-data clustering is high, indicating that federation itself is not the main source of utility loss in this experiment. The dominant bottleneck is instead the PP-SDG step, which does not yet preserve the cluster-conditional dependencies required by the downstream task.

These findings lead to three main conclusions. First, federated analytics can approximate centralised analytics on the real ALS data while avoiding the pooling of raw patient records. Second, current PP-SDG methods can preserve some statistical properties of the ALS data, but this preservation is not sufficient to guarantee utility for downstream clustering. Third, direct synthetic-data utility metrics and downstream task utility should be treated as complementary: good performance on marginal distributions or model-ranking agreement does not necessarily imply preservation of the structures required by a specific analytics task.

The results therefore support the role of synthetic data as a privacy-preserving fallback mechanism, but also show that this fallback must be used with task-aware validation. In future work, the synthetic data generation process should be more tightly coupled with the requirements of downstream analytics, for example, by preserving task-relevant dependencies or by using the risk explanations produced by PRISM to guide which statistical properties must be retained and which identifying structures must be suppressed. The SPARQL-based k-means prototype further points to a tighter integration between federated analytics, ontology-mediated data access, and privacy-aware governance in the HDN.

# 6 Milestone Verification

This deliverable verifies Milestone 10 - privacy-preserving methods, described as: “federated workflow execution engine enhanced with privacy-preserving algorithms”. We present the first release of a privacy-preserving analytics workflow for the HDN. The deliverable extends the federated workflow with three complementary components: PRISM, which provides query-level privacy-risk assessment and governance; PP-SDG methods, including DPGAN, PATE-GAN, and PrivBayes, which provide privacy-preserving synthetic substitutes for sensitive tabular data; and an empirical privacy-assessment framework, implemented through PrivEval, which supports the evaluation of residual disclosure risks in synthetic datasets.

The milestone is further substantiated by experimental validation on the HEREDITARY ALS dataset. The deliverable evaluates privacy-preserving synthetic data in terms of both utility and privacy, and uses k-means clustering as a representative downstream federated analytics task. The results show that federated execution over real data introduces limited distortion with respect to the corresponding centralised setting, while the main loss of downstream utility is caused by the current PP-SDG step, which does not yet fully preserve the multivariate cluster structure required by k-means. This shows that the federated workflow has been enhanced with privacy-preserving mechanisms and has been evaluated in a concrete medical analytics scenario.

In addition, the deliverable investigates a SPARQL-based implementation path for federated k-means, aimed at tighter integration with the HDN and its ontology-mediated query layer. This component is included as a design and prototyping activity for the next release, whereas the quantitative validation reported in this deliverable is based on the Flower implementation. Together, the governance layer, the synthetic-data fallback, the privacy-assessment framework, and the federated k-means evaluation provide evidence that the milestone has been achieved within the scope of this first release. The remaining work towards the final version concerns the quantitative validation of the SPARQL-based execution path, tighter end-to-end integration with PRISM, and improved task-aware PP-SDG.

# 7 Conclusions and Remarks

This deliverable presented the first release of the HEREDITARY privacy-preserving analytics workflow. We build on the federated infrastructure and privacy foundations established in previous deliverables, and focus on the mechanisms needed to assess, mitigate, and validate disclosure risk when analytics are performed over sensitive medical data. The deliverable contributes to this objective through the revision of the privacy setting of the HDN, the analysis of privacy-preserving synthetic data generation methods, the organisation of privacy metrics for synthetic data, and an experimental evaluation on the HEREDITARY ALS dataset.

A first contribution concerns the privacy governance of the HDN. The deliverable revises the previous construct-based view of query risk with PRISM, a query-level risk-assessment framework designed in collaboration with WP7 and grounded in the legal and ethical requirements governing the federated processing of sensitive health data. Rather than relying only on the syntactic form of a query, PRISM considers the query, the ontology annotations, and, where required, data-level statistics at the endpoints. It produces a calibrated risk score, a risk bucket, and a structured explanation of the factors driving the decision. This provides the basis for a more fine-grained, explainable, and auditable privacy control mechanism, aligned with the spirit of GDPR and with the need to support accountable decisions at the HDN gateway. In this first release, PRISM is introduced at the single-query level and establishes the governance layer on which further integration with the HDN gateway can build.

A second contribution concerns the role of synthetic data in the privacy-preserving analytics workflow. The deliverable positions PP-SDG as a controlled fallback mechanism for cases in which real data cannot be safely queried or analysed under the applicable governance policy. In this setting, synthetic data are not treated as a replacement for privacy assessment, but as a mitigation mechanism whose own privacy and utility properties must be evaluated. This links the query-level assessment performed by PRISM with the generation and validation of synthetic substitutes.

To support this validation, the deliverable surveyed privacy metrics for synthetic tabular data and organised them into a common taxonomy. The metrics were grouped according to their computational structure and underlying threat model, including simulation-based, threshold-based, distance-based, and classification-based approaches. The taxonomy clarifies the assumptions behind different metrics and supports a more systematic interpretation of residual disclosure risk. The survey is operationalised in PrivEval, which implements a broad set of metrics and supports data owners in evaluating both privacy risk and metric applicability in practical settings.

The experimental part of the deliverable evaluated PP-SDG methods on the HEREDITARY ALS dataset. The evaluation focused on the tabular view currently most relevant to the clinical partners, obtained from static patient variables and outcome information. Among the methods considered in this release, PrivBayes provides the strongest observed utility on the ALS tabular data. It better preserved selected marginal distributions and achieved higher SRA scores than DPGAN and PATE-GAN. Under the assumption that the considered implementations correctly realise their intended DP mechanisms, the empirical privacy metrics did not indicate a higher residual disclosure risk for PrivBayes. Within the scope of this first release, PrivBayes therefore emerged as the most promising PP-SDG method among those evaluated.

The experiments also highlighted limitations that are relevant for the next iterations of the workflow. PrivBayes relies on BNs, which requires discretisation of continuous variables and may face scalability issues as the number of attributes grows. The noise introduced during structure and parameter learning may weaken dependencies between attributes, and the generated records may violate domain constraints, as observed in the case of implausible genetic mutation combinations. Moreover, the downstream k-means evaluation showed that good performance on direct utility and privacy metrics does not necessarily imply preservation of the multivariate structure required by a specific analytics task. In particular, the synthetic datasets did not preserve the cluster-conditional structure needed by k-means.

The federated K-means experiments provided an additional validation of the analytics workflow. The

results showed that federated execution over real ALS data approximates the corresponding centralised real-data clustering with limited distortion. By contrast, the main degradation in clustering agreement was introduced by the synthetic data generation step. For the task considered in this deliverable, federation itself is therefore not the main source of utility loss; the current limitation lies in the ability of PP-SDG methods to preserve the task-relevant structure of the data. This result indicates that synthetic data should be assessed not only through general-purpose utility and privacy metrics, but also through validation against the downstream analytics tasks for which they are intended.

Finally, the deliverable also introduces a SPARQL-based implementation path for federated K-means. This activity is included as a design and prototyping contribution towards closer integration with the HDN execution model. Compared with a framework-based implementation such as Flower, the SPARQL-based approach keeps computation closer to the local data stores, expresses client-side operations declaratively, and aligns more naturally with ontology-mediated access and privacy-aware query governance. At the same time, the work identified technical challenges related to nearest-centroid assignment, k-means++ initialisation, state management, and the lack of native query constructs such as cumulative sums and stable random sampling. These results provide a concrete basis for assessing how federated analytics can be moved closer to the HDN query layer in future iterations.

In the current release, we focus on the synthetisation of tabular data. The longitudinal data in clinical datasets remains relevant for richer modelling of disease progression, but it requires additional methodological choices, both for synthetic data generation and for privacy assessment. The tabular view addressed in this deliverable was therefore the appropriate target for the first release, as it is currently the most immediate representation used by the clinical partners and is compatible with the PP-SDG methods evaluated here.

Overall, the deliverable establishes the foundations of the HEREDITARY privacy-preserving analytics pipeline. PRISM provides the governance mechanism for assessing when real data can be used and when privacy-preserving alternatives are required. PP-SDG provides a candidate fallback mechanism for maintaining answerability under privacy constraints. The privacy-metric taxonomy and PrivEval provide the empirical assessment layer needed to interpret residual risks in synthetic data. The ALS experiments demonstrate how these components can be combined in a federated analytics scenario, while also identifying the main limitations of the current release.

The results provide input for the next activities of WP3, and especially Tasks T3.1, T3.2 and T3.5. Further work may refine the integration of PRISM with the HDN gateway, extend the assessment of privacy risk beyond isolated queries, improve task-aware and domain-aware PP-SDG, and evaluate the SPARQL-based execution path in the HDN. The observed generation of implausible records also indicates the need to investigate how clinical and biological expert knowledge could be used to constrain or validate synthetic data generation. Ontology-guided mechanisms, for example based on HERO or related domain resources, may provide a way to encode such constraints while remaining aligned with the semantic layer of the HDN. These directions will inform the consolidation of the privacy-preserving analytics workflow towards the final release of HDN.

## References

[1] Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep Learning with Differential Privacy. In *ACM Conference on Computer and Communications Security*, pages 308–318. ACM.

[2] Alaa, A. M., van Breugel, B., Saveliev, E., and van der Schaar, M. (2021). How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. *CoRR*, abs/2102.08921.

[3] Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *SODA*, pages 1027–1035. SIAM.

[4] Barouh, M., Boytcheva, S., Cazzaro, M., Dell'Aglio, D., Fabbian, L., Gyurov, P., Rodríguez, J. M., and Silvello, G. (2025). Deliverable 3.2: Federated workflow execution methods. first release.

[5] Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., Ueyama, K., and Williams, A. R. (2020). A synthetic supplemental public use file of low-income information return data: Methodology, utility, and privacy implications. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings*, page 257–270, Berlin, Heidelberg. Springer-Verlag.

[6] Cai, K., Xiao, X., and Cormode, G. (2023). Privlava: Synthesizing relational data with foreign keys under differential privacy.

[7] Calvanese, D., Giacomo, G. D., Lembo, D., Lenzerini, M., Poggi, A., and Rosati, R. (2007). Ontology-based database access. In Ceci, M., Malerba, D., and Tanca, L., editors, *Proceedings of the Fifteenth Italian Symposium on Advanced Database Systems, SEBD 2007*, pages 324–331, Torre Canne, Fasano (BR), Italy.

[8] Cazzaro, M., Menotti, L., Primov, T., Rueda, M., and Silvello, G. (2024). Deliverable 3.1: Ontology and federated execution methods. EU Project Deliverable D3.1, Hereditary Project.

[9] Chiò, A., Grassano, M., and Kamenjasevic, E. (2024). Deliverable 2.1: Ethical guidelines, data collection and sharing.

[10] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2018). Generating multi-label discrete patient records using generative adversarial networks.

[11] Cima, G., Lembo, D., Marconi, L., Rosati, R., and Savo, D. F. (2024). A gentle introduction to controlled query evaluation in DL-Lite ontologies. *SN Computer Science*, 5(4):335. Survey article, topical collection "Advances on Web Information Systems and Technologies".

[12] DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. A., Flynn, H., Adamson, J., et al. (2011). Expanded ggggcc hexanucleotide repeat in noncoding region of c9orf72 causes chromosome 9p-linked ftd and als. *Neuron*, 72(2):245–256.

[13] El Emam, K., Mosquera, L., and Fang, X. (2022). Validating a membership disclosure metric for synthetic health data. *JAMIA Open*, 5(4):ooac083.

[14] Esteves, B., Golpayegani, D., Krog, G. P., Pandit, H. J., Flake, J., and Ryan, P. (2026). Data privacy vocabulary (dpv). Final community group report, Data Privacy Vocabularies and Controls Community Group. W3C Community Group Report.

[15] European Parliament and Council of the European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88.

[16] Faggioli, G., Menotti, L., Marchesin, S., Chió, A., Dagliati, A., de Carvalho, M., Gromicho, M., Manera, U., Tavazzi, E., Nunzio, G. M. D., Silvello, G., and Ferro, N. (2024). An extensible and unifying approach to retrospective clinical data modeling: the BrainTeaser ontology. *Journal of Biomedical Semantics*, 15(1):16.

[17] Fan, J., Chen, J., Liu, T., Shen, Y., Li, G., and Du, X. (2020). Relational data synthesis using generative adversarial networks: a design space exploration. *Proc. VLDB Endow.*, 13(12):1962–1975.

[18] Ganev, G., Annamalai, M. S. M. S., and Cristofaro, E. D. (2025). The elusive pursuit of reproducing PATE-GAN: benchmarking, auditing, debugging. *Trans. Mach. Learn. Res.*, 2025.

[19] Ganev, G. and Cristofaro, E. D. (2023). The inadequacy of similarity-based privacy metrics: Privacy attacks against "truly anonymous" synthetic datasets. *2025 IEEE Symposium on Security and Privacy (SP)*, pages 4007–4025.

[20] Garst, S. and Reinders, M. (2024). Federated k-means clustering. In *International Conference on Pattern Recognition*, pages 107–122. Springer.

[21] Golob, S., Pentyala, S., Maratkhan, A., and Cock, M. D. (2025). Privacy vulnerabilities in marginals-based synthetic data.

[22] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *NIPS*, pages 2672–2680.

[23] Gu, Z., Calvanese, D., Panfilo, M. D., Lanti, D., Mosca, A., and Xiao, G. (2024). OBDF: OBDA + data federation – extended abstract. In *40th IEEE International Conference on Data Engineering Workshops, ICDE 2024*, pages 381–383, Utrecht, Netherlands. IEEE.

[24] Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C.-A., Goronflot, T., Vince, N., Limou, S., Karakachoff, M., Wargny, M., and Gourraud, P.-A. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digital Medicine*, 6(1):37.

[25] Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., and Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493.

[26] Hittmeir, M., Mayer, R., and Ekelhart, A. (2020). A baseline for attribute disclosure risk in synthetic data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, CODASPY '20, page 133–143, New York, NY, USA. Association for Computing Machinery.

[27] Hittmeir, M., Mayer, R., and Ekelhart, A. (2022). Efficient bayesian network construction for increased privacy on synthetic data. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5721–5730.

[28] Jordon, J., Yoon, J., and van der Schaar, M. (2019). PATE-GAN: generating synthetic data with differential privacy guarantees. In *ICLR (Poster)*. OpenReview.net.

[29] Kamenjasevic, E., Biasin, E., Dell'Aglio, D., and Romanovych, A. (2024). Deliverable 7.1: Legal, ethical, and regulatory inventory.

[30] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

[31] Lautrup, A. D., Hyrup, T., Zimek, A., and Schneider-Kamp, P. (2024). Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1).

[32] Liu, K. S., Li, B., and Gao, J. (2018). Generative model: Membership attack, generalization and diversity. *CoRR*, abs/1805.09898.

[33] Lu, P.-H., Wang, P.-C., and Yu, C.-M. (2019). Empirical evaluation on synthetic data generation with generative adversarial network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, WIMS2019, New York, NY, USA. Association for Computing Machinery.

[34] Niederhametner, N. and Mayer, R. (2023). Protecting multiple sensitive attributes in synthetic micro-data. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5529–5538.

[35] Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74(11):1–26.

[36] Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. J., and Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*. OpenReview.net.

[37] Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

[38] Podareanu, D., Dell'Aglio, D., Romanovych, A., and Silvello, G. (2024). Deliverable 2.14: Computing infrastructures.

[39] Qian, Z., Cebere, B.-C., and van der Schaar, M. (2023). Synthcity: facilitating innovative use cases of synthetic data in different data modalities.

[40] Slokom, M., de Wolf, P.-P., and Larson, M. (2023). Exploring privacy-preserving techniques on¬†synthetic data as¬†a¬†defense against model inversion attacks. In Athanasopoulos, E. and Mennink, B., editors, *Information Security*, pages 3–23, Cham. Springer Nature Switzerland.

[41] van Breugel, B., Sun, H., Qian, Z., and van der Schaar, M. (2023). Membership inference attacks against synthetic data through overfitting detection. *ArXiv*, abs/2302.12580.

[42] van der Wal, D., Podareanu, D., Cardenas, B., Biasin, E., and Romanovych, A. (2025). Deliverable 2.15: Communication protocols.

[43] Van der Wal, D., Podareanu, D., Rodríguez, J. M., Silvello, G., and Romanovych, A. (2025). Deliverable 2.11: Federated infrastructure design.

[44] Venugopal, R., Shafqat, N., Venugopal, I., Tillbury, B. M. J., Stafford, H. D., and Bourazeri, A. (2022). Privacy preserving generative adversarial networks to model electronic health records. *Neural Networks*, 153:339–348.

[45] Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network. *CoRR*, abs/1802.06739.

[46] Yale, A., Dash, S., Bhanot, K., Guyon, I., Erickson, J. S., and Bennett, K. P. (2020a). Synthesizing quality open data assets from private health research studies. In Abramowicz, W. and Klein, G., editors, *Business Information Systems Workshops*, pages 324–335, Cham. Springer International Publishing.

[47] Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., and Bennett, K. P. (2019). Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, AIDR '19, New York, NY, USA. Association for Computing Machinery.

[48] Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., and Bennett, K. P. (2020b). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255.

[49] Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., and Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13(1).

[50] Yoon, J., Drumright, L. N., and van der Schaar, M. (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE journal of biomedical and health informatics*, 24(8):2378–2388.

[51] Yoon, J., Jordon, J., and van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

[52] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2014). PrivBayes: private data release via bayesian networks. In *SIGMOD conference*, pages 1423–1434. ACM.

[53] Zhang, Z., Yan, C., Mesa, D. A., Sun, J., and Malin, B. A. (2019). Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, 27(1):99–108.

[54] Zhao, Z., Kunar, A., der Scheer, H. V., Birke, R., and Chen, L. Y. (2021). CTAB-GAN: effective table data synthesizing. *CoRR*, abs/2102.08369.