Risk Analysis of a personal data anonymisation process using the avatar method - Example_report

Octopize MD

Date

Synthesis Table of Risks

Risks	Attack Attempt Likelihood	Occurence Likelihood	Impact	EDPB Criterion
Re-identification through Origin–Synthetic Linkage	**Negligible**	**Negligible**	**High**	Singling-out
Re-identification via Demographic Attributes	**Very Low**	**Negligible**	**High**	Linkability
Synthetic Data Too Similar to Original Data	**Negligible**	**Negligible**	**High**	Singling-out
Synthetic Record Closer to Source	**Negligible**	**Negligible**	**High**	Singling-out
Re-identification via Single Attribute	**Low**	**Negligible**	**High**	Singling-out
Synthetic Data Identical to Original Records	**Low**	**Negligible**	**High**	Singling-out, Linkability

Risks	Attack Attempt Likelihood	Occurence Likelihood	Impact	EDPB Criterion
Inference of Sensitive Attribute	**Very Low**	**Negligible**	**High**	Inference
Synthetic Data for Linkability of Other Datasets	**Very Low**	**Negligible**	**High**	Linkability

Process description

Purpose of the Processing

The objective of the processing is to transform a dataset containing directly identifiable or pseudonymised sensitive personal data into a synthetic dataset referred to as avatars to eliminate the risk of future re-identification. Upon completion of the transformation, the resulting data will be considered **anonymous** within the meaning of the **General Data Protection Regulation** (GDPR), and will therefore no longer qualify as personal data. The transformation relies on the Avatar method developed by Octopize, which has been audited by the French Data Protection Authority (CNIL) against the three criteria defined by the **European Data Protection Board** (EDPB, formerly WP 29), namely:

- Singling-out
- Linkability
- Inference

According to the EDPB, anonymisation processing under the GDPR constitutes a **compatible further processing** of personal data. Consequently, it does **not require** the identification of a **new legal basis** pursuant to *Article 6* of the Regulation. Nevertheless, as a matter of good practice and in line with the principle of transparency (*Article 5(1)(a) GDPR*), it is recommended to inform data subjects of the potential for such further processing. This fosters trust and strengthens compliance with the accountability principle.

According to GDPR, French data protection authority and the European Committee on data protection anonymisation shall be accepted if the following criteria are met :

- Singling-out : meaning that it is not possible to single out one person in the dataset.
- Linkability : meaning that it is not possible to link two datasets regarding one person.
- Inference : meaning that it is not possible to induce any new information about one person.

The latest EU justice court position on anonymisation (*September 4th 2025 (C413/23 P)*) considered pseudonymised data as anonymised if the recipient of those data is unable to re-identify the person. The data controller shall then conduct a risk analysis to assess the feasibility of the re-identification.

Data Controller

The Data Controller is the organisation that initially collects and hosts the personal data prior to the anonymisation process. This entity determines the purposes and means of the processing operations leading to the generation of synthetic data (avatars).

As such, the Data Controller bears primary responsibility for ensuring that the processing of personal data up to the point of anonymisation complies with the applicable provisions of the GDPR, including those relating to lawfulness, fairness, transparency, data minimisation, and security of processing (*Articles 5 and 32 GDPR*).

Brief Description of the Method

The Avatar method is a **transparent** and **explainable** anonymisation technique for generating synthetic data, **compliant** with the GDPR, and has been both audited by the CNIL and peer-reviewed. It is an individual-centric method that generates synthetic data by introducing irreversible randomness within a local modelling zone.

- Further technical documentation is available here
- Peer-reviewed publications available here

Processor

When the Avatar method is used as a **SaaS** Platform: The anonymisation algorithm is maintained and updated by Octopize, which acts as the data processor within the scope of this processing activity.

When the Avatar method is used as an **on-premise** Platform: Octopize does not process any personal data. Its involvement is strictly limited to maintenance and update operations. These activities do not entail any access to, or processing of, the data by Octopize. Updates exclusively consist in the deployment of new code versions and, as such, do not give rise to any processing of personal data by Octopize.

Type of data processed

The data processed originate from health-related datasets containing information on patients or study participants. These datasets typically include demographic, clinical, and behavioural variables, such as age, gender, diagnosis codes, treatment details, medical outcomes, and follow-up data. The data are structured in one or several tabular databases

of mixed type variables (numerical, categorical, boolean, dates), possibly linking multiple sources (e.g., hospital records, laboratory results, or registries). Given their nature, such data qualify as special categories of personal data under *Article 9 GDPR*. Importantly, at the end of the processing, the data are rendered anonymous. As a result, they fall outside the scope of the GDPR, in accordance with *Recital 26*.

Data subject concerned

The data subjects are patients whose personal data are processed in the context of medical research, healthcare provision, or clinical trials. Such data may include information directly or indirectly identifying individuals, together with health-related or demographic variables. Given the sensitive nature of health data within the meaning of *Article 9 GDPR*, these individuals benefit from enhanced protection under the Regulation. The Anonymisation process is implemented to ensure that their personal data are protected in accordance with the principles of lawfulness, fairness, transparency, and data minimisation.

Information and Consent Mechanisms

The anonymisation processing is deemed compatible with the initial purposes for which the personal data were collected, in accordance with *Article 6(4) GDPR*. It does not amount to a substantial modification of such purposes and therefore does **not require renewed consent** or specific re-information of data subjects.

No new legal basis is consequently required for this further processing. Nevertheless, and in line with the principle of transparency set out in *Article 5(1)(a) GDPR*, it constitutes a good practice to inform data subjects of the possibility of such further processing. Doing so fosters trust, reinforces the fairness of the processing, and supports compliance with the accountability principle under *Article 5(2) GDPR*.

Exercise of Data Subject Rights

Personal data are used solely for the duration necessary to perform the anonymisation process, after which they are **permanently deleted** from the dedicated server. The server is owned and managed by the Data Controller. Once anonymised, the resulting synthetic data (avatars) are **no longer considered personal data** and fall outside the scope of the GDPR, as they cannot be linked to any identifiable individual. These synthetic avatars do **not retain** any connection to the original data subjects. Therefore, it will be possible to determine from which database the avatars were derived, given that the avatars' database preserves the structural format of the initial database. However, it will never be possible to identify which specific data subject corresponds to a given avatar. This ensures that the anonymisation process fully complies with *Recital 26* and the core principles of data minimisation and privacy by design.

Data recipient

The recipients of the anonymised data are exclusively internal stakeholders of the Data Controller, such as authorised employees, researchers, or analysts operating within the same organisation. The synthetic datasets are used for internal analytical, research, or operational purposes, in strict compliance with the principles of data protection by design and by default (*Articles 5 and 25 GDPR*). No transfer of personal data occurs beyond the Data Controller's perimeter, as the anonymised data—being fully non-identifiable within the meaning of *Recital 26*—are no longer subject to the GDPR. Appropriate technical and organisational measures are maintained to ensure that only anonymised datasets are accessible within the internal environment.

Processing Characteristics

The original personal data may reside within the Data Controller's information system or temporarily transit through a server certified for the hosting of health data according to the French regulation located in France for the duration of the anonymisation process, after which they are **permanently deleted**. The processing implements an **irreversible mathematical transformation** designed to prevent re-identification, resulting in the generation of synthetic data referred to as avatars. Each avatar is generated based on a local model constructed from a group of individuals exhibiting similar characteristics ("neighbours"). A randomised selection is performed within the range of values observed among these neighbours to create each avatar. The parameter k, defined at the outset, specifies the number of neighbours included in each local model. The Avatar algorithm produces a unique and random transformation for each individual in the dataset. Only the avatarised dataset is subsequently used for analysis, transfer, or other study-related purposes. In accordance with the criteria defined by the EDPB and *Recital 26 GDPR*, the residual risk of re-identification is systematically assessed using state-of-the-art attack simulations and associated metrics. The preservation of statistical properties is also evaluated to ensure the scientific validity of the synthetic dataset. Each processing operation is documented by Octopize, providing details on the methodology applied, the parameterisation used, and the resulting performance metrics, thereby ensuring compliance with the accountability principle (*Article 5(2) GDPR*).

Risk Analysis

This section provides a structured assessment of potential threats that could compromise the anonymisation process and, consequently, the protection of data subjects' privacy. For each identified risk scenario, the analysis examines (i) the nature of the threat and its relevance to the EDPB re-identification criteria, (ii) the likelihood that an attack would be attempted, (iii) the likelihood that such an attack would succeed under standard safeguards and parameterisation, and (iv) the potential impact should the threat materialise. Each

scenario is accompanied by a specific quantitative evaluation metric designed to measure, in technical and empirical terms, the residual exposure of the synthetic dataset to that particular mode of attack. This systematic approach ensures that risks are assessed not only in abstract or theoretical terms, but also through verifiable indicators that support demonstrable accountability and compliance with the requirements of *Recital 26 GDPR* and EDPB guidance on anonymisation.

Parameters Used for Privacy Assessment

Known Variables

No variables could be considered as potentially known by an attacker.

Frozen Variables

No variables were frozen during the avatarization process.

Re-identification through Origin–Synthetic Linkage (Methodological Flaw)

Description of the Threat

An attacker may attempt to exploit a flaw in the anonymisation method to reconstruct or infer a deterministic or probabilistic mapping between original records and their synthetic counterparts (avatars). Such a breach would compromise the irreversibility of the transformation and enable singling-out of individuals.

Relevant EDPB Re-identification Criterion

singling-out

Likelihood of Attack Attempt

Risk: **Negligible**

provided that it would require simultaneous access to: (i) the full pseudonymised source dataset ; (ii) the complete avatar dataset generated from that same cohort to perform an attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

provided that recommended safeguards and parameterisation are applied. Concretely, a successful attack would require simultaneous access to: (i) the full pseudonymised source dataset; (ii) the complete avatar dataset generated from that same cohort; and (iii) detailed

knowledge of the algorithmic parameterisation (e.g., the value of k, random seeds, or stochastic distributions) or the ability to infer these through extensive testing. Such access would almost certainly arise from illegitimate events (data breach, insider misuse, or secure-environment compromise) rather than lawful use. The adversary would also need advanced data-science expertise, significant computational resources, and time to perform iterative model inversion and record-linkage analyses. Feasibility therefore depends on the breadth of data exposure, the structural overlap of the two datasets, and the confidentiality of anonymisation parameters.

$Associated Risk if the Threat Materializes

Risk: **High**

successful linkage would negate anonymisation, returning the dataset to the scope of the GDPR and exposing data subjects to harm (privacy intrusion, profiling, discrimination), with corresponding legal and reputational consequences.

Evaluation Metric

The **Hidden Rate** is a metric specific to the Avatar method, expressly designed to evaluate the robustness of the anonymisation process against distance-based record linkage attacks. It quantifies the extent to which the direct correspondence between an original record and its avatar has been effectively disrupted by the anonymisation process. More precisely, it measures the percentage of individuals for whom the generated avatar is not the most similar entry in the dataset. The **Hidden Rate** serves as a safeguard against algorithmic flaws. It verifies that the synthetic data generation mechanism does not inadvertently preserve a one-to-one mapping between original and synthetic records, which would constitute a methodological weakness. The Hidden Rate also serves as empirical assurance of unlinkability. It provides a measurable indication that the link between original and synthetic data has been statistically broken in practice, in line with *Recital 26 GDPR* and the EDPB singling-out criterion.

Result

if the attack is attempted and the threat occurs (with all implications) -> percentage of protected individuals: 89.06%.

Interpretation: For instance, a **Hidden Rate** of 80% means that in 80% of cases, an avatar is not its original record's closest neighbour. The remaining 20% does not translate into direct risk, since those cases are selected randomly and cannot be predicted by an attacker. Higher values indicate stronger unlinkability, but even lower values do not automatically imply re-identification, as attack success depends on external conditions (auxiliary data, parameter disclosure, etc.).

See the Octopize documentation for more details on this metric.

Re-identification via Demographic Attributes

Description of the Threat

An attacker possessing external demographic information (for example: age, gender, geographic location) may attempt to re-identify individuals in the synthetic dataset by correlating these known traits with attributes reflected in the avatars, thereby enabling linkability between synthetic and original records.

Relevant EDPB Re-identification Criterion

singling-out

Likelihood of Attack Attempt

Risk: **Very Low**

Regarding (i) the number and discriminative power of demographic attributes available in the cohort to perform an attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

Regarding (i) the number and discriminative power of demographic attributes present in the cohort, and (ii) the availability of both the original pseudonymised dataset and the corresponding avatar dataset to the attacker. Successful exploitation of demographic correlation requires several simultaneous preconditions. Access to a complete original pseudonymised dataset containing the demographic attributes from cohort members. Full coverage materially increases re-identification feasibility because it enables exhaustive matching attempts across the same population. Such access is typically obtained through illicit means (data breach/exfiltration), the compromise/misconfiguration of the secure processing environment, or illegitimate insider actions. Access to the corresponding complete avatar dataset. The attacker also needs the full synthetic demographic output generated from that same cohort. The ease of obtaining avatar data depends on the usage context (internal analytics vs public release). Partial availability of avatars (or severe truncation/aggregation) substantially reduces risk. The attacker's method of acquisition matters. Breach/exfiltration or forced access to secure environment (e.g., exploiting misconfigured HDS host) yields high completeness and thus a higher likelihood. Accidental leakage can produce similar outcomes to a breach depending on scope. Scraping of public sources generally yields incomplete auxiliary datasets and will rarely produce a fully aligned dataset.

Associated Risk if the Threat Materializes

Risk: **High**

successful re-identification may enable singling-out and possibly inference of sensitive attributes, with resulting privacy harm and legal exposure under the GDPR.

Evaluation Metric

The **Knowledge Protection Rate** metric evaluates more probable and common attack scenarios, in which an attacker has partial knowledge of the original dataset. It measures the treatment's ability to protect the specific information of an individual against distance-based linkage attacks. The metric quantifies the percentage of individuals for whom the attacker cannot successfully link an original record to a unique approximate avatar, based on the partial information they possess.

The **Knowledge Protection Rate** is expressed as a percentage, where higher values indicate stronger privacy protection.

Result

if the attack is attempted and the threat occurs (with all implications) -> percentage of protected individuals: 86.0%

Interpretation: Consider a dataset where the **Knowledge protection rate** is measured at 85%. This indicates that 85% of avatars could not be linked to an unique similar original record when tested across a selected subset of variables. The remaining 15% of avatars, however, have a unique similar original record, highlighting a potential vulnerability in the synthetic data generation process.

See the Octopize documentation for more details on this metric.

Synthetic Data Too Similar to Original Data

Description of the Threat

There is a risk that synthetic records are so similar to their original counterparts that, despite stochastic transformation, they enable re-identification or inference of sensitive attributes. In such a scenario an attacker may identify the synthetic record that most closely matches a targeted individual and thereby probabilistically infer sensitive or identifying information.

Relevant EDPB Re-identification Criterion

singling-out

Likelihood of Attack Attempt

Risk: **Negligible**

Regarding the ease with which the attacker can obtain full or partial access to the original data relating to one or more individuals, to carry out the attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

in the context of a standard deployment of the Avatar method. Under normal parameterisation, synthetic records are generated within a local modelling zone defined by the original record's neighbours, which ensures that synthetic entries are not systematically closer to their originating records than real records are to one another. For an attacker to succeed in a distance-based re-identification attack, multiple unlikely preconditions must be met simultaneously: (i) one or more original variables would need to have been "frozen" (i.e. left unchanged or insufficiently perturbed) during synthesis; (ii) the attacker must possess the original pseudonymised dataset containing at least those frozen variables for the exact cohort under analysis; (iii) the attacker must also have access to the complete avatar dataset; and (iv) the attacker must have the specialist skills and computational resources required to design and execute large-scale nearest-neighbour / distance-matching attacks on the basis of the available attributes. In practice, partial or incomplete access to the original dataset, absence of frozen attributes, lack of avatar completeness, or insufficient auxiliary information markedly reduce the probability of success. Consequently, absent a concurrent data breach or other exceptional exposure of both source and synthetic datasets (and of any parameterisation that would aid inference), the scenario remains unlikely.

Associated Risk if the Threat Materializes

Risk: **High**

successful identification of synthetic records closely matching individuals could permit singling-out and the inference of sensitive attributes, with resulting legal, ethical and reputational consequences for the Data Controller and Processor.

Evaluation Metric

The **Distance to Closest** metric calculates the median Euclidean distance between an individual and the closest avatar in the dataset, based on all the available s and without considering the link between the original data point and its avatar. This metric provides insight into how closely the synthetic data replicates the distribution of the original data, while avoiding direct connections between individuals and their avatars.

Result

if the attack is attempted and the threat occurs (with all implications): 82.85%.

Interpretation: A value close to **100** or higher indicates that avatar proximity is similar to what is naturally observed in the reference data. This suggests that the avatar generation process does not enable reliable associations between original data points and their avatars. A value of **50**, for example, indicates that avatars are on average only half as distant from their original records compared to other records in the dataset. This reduced distance could potentially facilitate re-identification attempts, as avatars are more closely clustered around their source data points.

Note: The interpretation of this metric should always be considered alongside the **Closest Distance Ratio**, which evaluates the threat associated with the following risk.

See the Octopize documentation for more details on this metric.

Synthetic Record Is Significantly Closer to Its Source Individual than to Others

Description of the Threat

There is a risk that synthetic records remain disproportionately close to their source individuals compared to other data points in the dataset. This proximity may facilitate re-identification by allowing an attacker to establish a unique link between an original record and its synthetic avatar.

Relevant EDPB Re-identification Criterion

singling-out

Likelihood of Attack Attempt

Risk: **Negligible**

Regarding the ease with which the attacker can obtain full or partial access to the original data relating to one or more statistically rare individuals, to carry out the attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

under standard deployment and parameterisation of the Avatar method. By design, the algorithm protects records exhibiting remarkable or outlying characteristics: synthetic generation occurs within a local modelling zone defined by the original record's neighbours, which tends to re-centre synthetic outputs for outliers and prevents them from remaining uniquely extreme. For a successful attack aiming to find synthetic records that are significantly closer to their source individual than to others, several unlikely preconditions must be satisfied simultaneously: (i) one or more discriminative variables would need to have been intentionally or inadvertently "frozen" (insufficiently perturbed) during synthesis; (ii) the attacker must be able to reliably identify the outliers in the original cohort and possess the original pseudonymised records for those outliers including at least the frozen variables; (iii) the attacker must also obtain the complete avatar dataset corresponding to the same cohort; and (iv) the attacker must have the specialist analytical expertise and sufficient computational resources to perform targeted distance-based searches and validate candidate matches. Partial access to originals, incomplete avatar exposure, absence of frozen attributes, or lack of precise auxiliary knowledge substantially

reduce attack feasibility. Absent a concurrent breach exposing both the relevant original records and the full synthetic output (and any parameterisation logs), this threat remains unlikely in practice.

Associated Risk if the Threat Materializes

Risk: **High**

If attackers can reliably identify an avatar as the unique closest neighbor of an original record, they may achieve partial re-identification or infer additional attributes.

Evaluation Metric

The **Closest Distances Ratio** is a local density metric. It calculates the median ratio of the Euclidean distance between an individual and their closest avatar, relative to the distance between that individual and their second closest avatar. It helps assess the ease of pairing an original record with a single avatar in the synthetic dataset. The **Closest Distances Ratio** ranges from 0 to 1: A value close to 0 indicates that the distance between an individual and their closest neighbor is much smaller than the distance between the individual and their second closest neighbor. This suggests clear separation between avatars in the dataset. A value close to 1 means the distance between an individual and their closest avatar is similar to the distance between the individual and their second closest avatar, which could indicate the presence of identical data or duplication.

Result

if the attack is attempted and the threat occurs (with all implications): 95.36%.

Interpretation: A value close to **100** or higher indicates that the local density of avatars relative to the original data is similar to that of the reference data, suggesting that avatars do not enable reliable linkage between original records and synthetic ones. A value of **50**, for instance, indicates that the local density of avatars is on average twice as close to their original records compared to other records in the dataset. This increased proximity could potentially facilitate re-identification attempts, as avatars are more tightly clustered around their source data points.

Note: The interpretation of this metric should always be considered alongside the **Distance to Closest**, which evaluates the threat associated with the previous risk.

See the Octopize documentation for more details on this metric.

Re-identification via a Single Discriminant Attribute

Description of the Threat

An attacker may exploit a single highly discriminant attribute (e.g., postal code, unique age combination, or a technical identifier) to re-identify individuals in the synthetic dataset, even without combining multiple variables.

Relevant EDPB Re-identification Criterion

singling-out, linkability

Likelihood of Attack Attempt

Risk: **Low**

provided that an attacker only needs to have access to a single information to attempt an attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

contingent primarily on the ease with which an attacker can obtain access to a frozen discriminant variable and to the avatar dataset. Under standard parameterisation of the Avatar method, stochastic generation within a local modelling zone prevents the preservation of strict uniqueness: continuous attributes are perturbed and categorical modalities are redistributed (some modalities may disappear, others duplicate), which substantially reduces the chance that an individual remains directly identifiable by a single attribute. However, the probability of successful re-identification increases materially if the following preconditions are met simultaneously: (i) a column containing a highly discriminant attribute (i.e., a proportion of unique values approaching 1) was deliberately or inadvertently frozen during synthesis; (ii) the attacker has complete access to the avatar dataset corresponding to the cohort; and (iii) the attacker possesses the original value(s) of the frozen attribute for the target individuals (obtained legitimately or, more likely, via data exfiltration, misconfiguration, or targeted scraping). In this single-attribute scenario the computational and analytical burden on the attacker is relatively low (matching on one attribute), which elevates feasibility compared with multi-attribute linkage attacks. Partial access to avatars, absence of freezing, or lack of reliable original attribute values substantially reduce success probability. Accordingly, absent concurrent exposure of the frozen attribute in the originals and the full avatar output, the scenario remains unlikely; if such exposure occurs, the likelihood must be reclassified upward and immediate mitigation steps should be taken.

Associated Risk if the Threat Materializes

Risk: **High**

Successful exploitation of a discriminant attribute may enable direct re-identification of individuals, undermining anonymisation.

Evaluation Metric

The **Column Direct Match Protection** metric evaluates a simpler attack scenario, where the goal is to check if a single column in the dataset allows for direct linking between an individual and their avatar. It measures the ability of each column to protect against re-identification by assessing if the most re-identifying column can isolate individuals and their avatars. The **Column Direct Match Protection** returns the percentage of individuals whose data cannot be directly re-identified from the most re-identifying column in the dataset. A high percentage indicates that the most re-identifying column does not allow for re-identifying individuals, while a value close to 0 suggests that one or more columns contain information specific enough to re-identify individuals, such as a forgotten unique identifier.

Result

if the attack is attempted and the threat occurs (with all implications) -> percentage of protected individuals: 85.88%.

List of the most discriminant columns (<50%):

No columns below 50% threshold.

Interpretation: Imagine a dataset where the **Column Direct Match Protection** is 10%. This means that 10% of the individuals in this dataset cannot be directly re-identified from the most re-identifying column. A score this low could indicate that a personal identifier was omitted during the anonymization process, allowing some individuals to be re-identified.

See the Octopize documentation for more details on this metric.

Synthetic Data Identical to Original Records

Description of the Threat

If synthetic records are reproduced identically from the original dataset, trivial re-identification attacks become possible, completely undermining anonymisation.

Relevant EDPB Re-identification Criterion

singling-out

Likelihood of Attack Attempt

Risk: **Low**

provided that an attacker only needs to screen the avatar dataset, knowing information about one or multiple individuals, to perform an attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

in practical deployments of the Avatar method. By design, the algorithm injects stochasticity within each record's local modelling zone, making the generation of a synthetic record that is strictly identical to an original record an essentially random event with vanishing probability. Even in the exceptional case where a synthetic record coincides with one of the original neighbours, this outcome is accidental and — critically — unpredictable and undetectable a priori by an attacker without privileged knowledge. To exploit such an event an adversary would therefore need simultaneous and highly unlikely conditions: (i) complete access to the original pseudonymised dataset for the exact cohort and to the full avatar output; (ii) knowledge of or access to generation seeds/parameterisation (which must remain confidential); and (iii) the ability to test and confirm exact matches at scale. A particular borderline scenario arises when multiple original records are strictly identical within the source dataset: the algorithm may then produce identical synthetic records, but this does not increase re-identification risk because the originals themselves are already non-individualisable within the dataset. Taken together, these factors make the practical probability of exploitable identity-level duplication in synthetic data negligible under normal safeguards; the likelihood rises only in the event of simultaneous dataset exposure combined with disclosure of generation parameters.

Associated Risk if the Threat Materializes

Risk: **High**

Exact replication would directly expose personal data, resulting in a GDPR breach.

Evaluation Metric

Row Direct Match Protection evaluates the simplest re-identification scenario: a direct record-for-record comparison between the original and synthetic datasets. The goal is to ensure that original records are never reproduced identically, thereby preventing trivial re-identification. The **Row Direct Match Protection** is expressed as a percentage, representing the proportion of records from the original dataset that are not identically reproduced in the synthetic dataset. A value of 100% means no original records are reproduced as-is, ensuring strong privacy protection.

Result

if the attack is attempted and the threat occurs (with all implications) -> percentage of protected individuals: 100.0%.

Interpretation: If the **Row Direct Match Protection** is 95%, this means that 95% of original records do not appear identically in the synthetic data. In this case, only 5% of

records are reproduced exactly, which constitutes a residual but limited re-identification risk. A much lower score would represent a critical weakness in the anonymisation process.

See the Octopize documentation for more details on this metric.

Inference of Sensitive Attribute with High Confidence

Description of the Threat

An attacker may attempt to infer sensitive attributes (e.g., health condition, income bracket) with high accuracy using correlations between available demographic attributes and synthetic avatars.

Relevant EDPB Re-identification Criterion

Inference

Likelihood of Attack Attempt

Risk: **Very Low**

provided that it would require simultaneous access to: (i) partial original information related to individuals that were part of the cohort ; (ii) the complete avatar dataset generated from that same cohort, and data science expertise to perform the inference attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

under standard deployment and parameterisation of the Avatar method. A successful inference attack requires multiple simultaneous and unlikely preconditions: the adversary must obtain either the full or a substantial portion of the original dataset (excluding the sensitive target variable) for the exact cohort, together with access to the complete avatar dataset; possess advanced expertise in distance-based record linkage or supervised machine-learning inference methods; and execute an effective attack that yields high-confidence predictions of the target sensitive attribute. Practically, such an attack is feasible only where the target attribute is extremely structuring of the dataset (i.e., effectively determinable from other covariates by causal or near-deterministic relationships), because otherwise predictive uncertainty will prevent high-confidence inference. Crucially, the attacker must also be confident that the individuals they attempt to infer belong to the same cohort represented in the datasets—an additional non-trivial requirement that typically depends on auxiliary information or prior knowledge of cohort membership. The need for substantive auxiliary data, specialised skills, non-trivial compute resources, and time for iterative model training and validation materially reduces the probability of success. Consequently, absent both (a) disclosure or exfiltration of

substantial original records and avatars and (b) an inherently deterministic target attribute and proof of cohort membership, the scenario remains unlikely; if these preconditions are present, the likelihood should be re-evaluated to a higher category.

Associated Risk if the Threat Materializes

Risk: **High**

Accurate inference of sensitive information compromises individual privacy and may reveal special-category data.

Evaluation Metric

The **Attribute Inference**, evaluates the risk of an attacker successfully deducing the value of a target variable using a subset of known attributes. In practice, each individual is assigned the target value of the most similar synthetic data point. The resulting prediction performance serves as the measure of Attribute Inference. This metric is particularly relevant in attribute inference attack scenarios, where adversaries leverage publicly available or likely accessible demographic data to estimate categorical information about individuals.

Result

if the attack is attempted and the threat occurs (with all implications): 108.5%. Percentage of good inference: 18.62%.

Interpretation:

A value close to **100** indicates that inference performance using avatars is similar to that obtained using real data. The anonymization process does not increase inference risk. if the value is around **50**, it indicates that inference performance using avatars is two times better than using real data. If the percentage of good inference is also high, this could indicate a significant inference risk.

See the Octopize documentation for more details on this metric.

Synthetic Data are used for linkability of other datasets

Description of the Threat

If synthetic records allow simple linkage between original individuals across different datasets, an attacker could exploit this to aggregate information from multiple sources. This could lead to re-identification of individuals or inference of sensitive attributes, resulting in a breach of data protection regulations.

Relevant EDPB Re-identification Criterion

Inference

Likelihood of Attack Attempt

Risk: **Very Low**

provided that an attacker needs at least two data sources having common variables with the transformed avatar dataset to perform an attack targeting this vulnerability.

Likelihood of Occurrence

Risk: **Negligible**

regarding (i) the number and discriminative power of demographic attributes present in the cohort, (ii) the completeness and overlap of external auxiliary datasets, and (iii) the availability of both the original pseudonymised dataset and the corresponding avatar dataset to the attacker. Successful exploitation of linkability requires several simultaneous preconditions. Access to multiple external datasets containing overlapping variables with the avatar dataset. The attacker must obtain at least two distinct data sources that share common attributes with the synthetic data. The ease of acquiring such datasets depends on their public availability, the attacker's resources, and the specific context. Full coverage materially increases re-identification feasibility because it enables exhaustive matching attempts across the same population. Such access is typically obtained through illicit means (data breach/exfiltration), the compromise/misconfiguration of the secure processing environment, or illegitimate insider actions. Access to the corresponding complete avatar dataset. The ease of obtaining avatar data depends on the usage context (internal analytics vs public release). Partial availability of avatars (or severe truncation/aggregation) substantially reduces risk.

Associated Risk if the Threat Materializes

Risk: **High**

A dataset allowing linkability would enable an attacker to associate original individuals from multiple data sources, potentially leading to re-identification, inference and a breach of GDPR.

Evaluation Metric

Linkability Protection Rate quantifies the risk of linkability in synthetic datasets by simulating scenarios in which different subsets of variables are used to establish matches between avatars and 2 original subset records. Each original individual from both subsets is then matched to their closest avatar within the feature space defined by the selected variables. If an individual is linked to the same avatar across both subsets, this reveals a correlation risk for that individual. The **Linkability Protection Rate** is expressed as a percentage, representing the proportion of records from the original dataset that could be linked using avatars when spread across different datasets.

Result

if the attack is attempted and the threat occurs (with all implications): 99.77%.

Interpretation:

A score of **100** means that no avatar was linked to the same original record across the different variable subsets — demonstrating complete dissociation between the synthetic and original data. Conversely, lower values indicate a greater linkability risk, as some avatars maintain a stable association with their original record even when the linking variables change.

See the Octopize documentation for more details on this metric.

Additional security measures

In addition to anonymisation through synthetic data generation, further safeguards can be applied to reduce residual risks:

- **Controlled Access**: Restrict access to synthetic datasets to a limited number of authorized users. Access is based on the principle of the need to know : only the users involved in the project have access to the datasets.
- **Time-limited Access**: Provide only temporary access windows to minimize data exposure. Combined with a regular review of the access to be sure access opened matches the users.
- **Encrypted Storage & Sharing**: Store and exchange synthetic datasets within secure, encrypted environments (e.g., SFTP, secure cloud vaults).
- **Access Logs & Monitoring**: Maintain detailed audit logs to track who accessed the dataset and when, paired with a regular review of the logs.
- **Data Minimisation**: Share only the strictly necessary subset of data required for a given use case.
- **Differential Privacy Enhancements**: Add controlled noise to protect against inference attacks without compromising analytical utility.
- **Federated Analysis**: Where possible, avoid sharing data altogether by enabling computation in secure environments and only exporting aggregated results.
- **Two-factor Authentication**: Require strong authentication for anyone accessing the data.
- **Regular Risk Assessments**: Periodically re-evaluate re-identification risks as new techniques emerge.

- **Legal & Contractual Safeguards**: Complement technical controls with NDAs, data processing agreements if the anonymisation is carried out on the SaaS method, clauses to the agreements with third parties prohibiting them from trying to re-identify by any means the individuals, and strict usage policies.
- **Separation of Duties**: Ensure no single person has full control over both data access and processing.
- **Revocation Mechanisms**: Allow rapid removal of access rights in case of misuse or suspicion.