APPENDIX 7: De-identification techniques and privacy evaluation in synthetic data

De-identification

De-identification refers to a collection of technical and organisational approaches intended to reduce the likelihood that data can be associated with an identifiable individual, such that the data may no longer constitute personal data under applicable privacy frameworks [1,2]. While preventing direct identity disclosure (i.e., linking a named individual to a specific record) is an important goal, it is crucial to recognize that data may still be considered personal information even if an individual's identity is not explicitly recovered. For instance, if information remains reasonably linkable to an individual or a small group, it may still fall under the purview of privacy regulations [1].
De-identification is therefore best understood as a risk-management exercise rather than a binary state. A trade-off between privacy risk and data utility is almost always unavoidable; privacy protection generally improves as more aggressive techniques are applied, but at the cost of reduced data utility [3]. There is no single "correct" way to de-identify data.

Examples of de-identification techniques include:

- aggregation: suppression (remove identifiers or other overtly identifying data fields)
- Generalisation: (e.g., replacing exact date of birth with age bands)
- pseudonymisation: using cryptographically protected transformations (e.g., keyed hashing such as HMAC, or other non-reversible mappings with appropriate key management. Plain hashing alone is widely used but is susceptible to brute-force and dictionary attacks and should not be considered sufficient protection [4])
- perturbation: (adding noise, micro-aggregation or data-swapping).

Evaluation of Privacy in Synthetic Data

Overview

Privacy protection is a foundational principle guiding the generation of synthetic data, which inherently represents a trade-off between fidelity, utility, and privacy. Consequently, privacy evaluation is a critical, yet often misunderstood, component of synthetic data governance. While synthetic data aims to protect privacy by generating artificial records without direct links to real individuals, residual privacy risks persist. Robust privacy assessment remains essential, as risks such as membership and attribute inference may arise when synthetic data preserves statistical patterns from the original dataset.

This appendix provides an overview of current approaches to evaluating privacy in synthetic data. It is a rapidly evolving field, and there is no single accepted definition or universal measure of privacy risk. Existing metrics capture only partial and often complementary aspects of risk, and favourable results under one metric do not preclude vulnerabilities under

others. The privacy engineering and security communities have raised significant concerns about the limitations of many commonly used metrics, particularly those based on similarity [5, 6]. Therefore, privacy evaluation should be viewed as the collection of a portfolio of evidence rather than a single, definitive score.

Types of Privacy Risk

Privacy evaluation in synthetic data focuses on several primary types of disclosure. While the terminology can vary, the most common risks are:

- **Identity Disclosure**: This occurs when a synthetic record can be confidently linked to a specific, named individual. In its purest form, this may be less of a direct threat for synthetic data because direct identifiers (like names or social security numbers) should have been removed from the training data. However, the risk is not entirely eliminated. As noted earlier, synthetic data may still be considered personal information if it remains reasonably linkable to an identifiable individual, even if their name is not explicitly revealed [1]. In practice, the more substantive privacy risks in synthetic data often manifest as membership and attribute disclosure.
- **Membership Disclosure**: This occurs when an adversary can determine whether a specific individual's data was included in the training dataset used to generate the synthetic data. This risk becomes particularly acute when membership in the dataset itself is sensitive information (e.g., a list of participants in a clinical trial for a specific disease or a dataset of individuals with a particular political affiliation).
- **Attribute Disclosure**: This occurs when an adversary can infer new and sensitive information about an individual using the synthetic data, often by leveraging some auxiliary information they already possess. A critical challenge here is to distinguish between legitimate knowledge generation (i.e., learning valid statistical patterns about a population) and a privacy violation (i.e., learning something new and specific about an individual who was in the training data) [7].

A Landscape of Evaluation Metrics

A wide range of metrics has been proposed to evaluate privacy in synthetic data [8, 9], with various classification schemes presented in the literature. One of the most practical groupings is summarised in Table 1. These metrics span traditional re-identifiability measures adapted from anonymised datasets, such as k-anonymity, to distance-based metrics and adversarial approaches like membership inference attacks.

It is critical to understand that while these methods provide useful lenses for assessing certain types of risk, they do not offer bounded guarantees of privacy loss, nor do they fully resolve the inherent trade-off between privacy and usability. The field is evolving rapidly, with new attacks, auditing techniques, and theoretical results continuously emerging [5, 10]. The following table is therefore not an exhaustive list or an endorsement of any particular approach. Rather, its aim is to clarify what these common measures assess and, more importantly, what they do not.

.

Table 1: Categories of metrics used for evaluating privacy in synthetic data.

Evaluation Category		Evaluation Method / Metric	Description
I. Non-Adversarial Metrics (often adapted from anonymised datasets.)	**A. Re-identifiability metrics**	k-Anonymity	Checks whether each individual is indistinguishable from at least k−1 other individuals with respect to a set of quasi-identifiers, under the assumption that each individual is represented by a single record. (Note: record-level protection alone may be insufficient for longitudinal or event-level data.)
		l-Diversity	An extension of k-Anonymity ensuring that sensitive attributes within each anonymised group have at least l distinct values. (Note: l-diversity and t-closeness apply to sensitive attributes and do not provide protection for quasi-identifiers; attributes that are both sensitive and quasi-identifying require additional treatment.)
		t-Closeness	Further refines l-Diversity by ensuring the distribution of a sensitive attribute in any group is close to its distribution in the overall dataset. (Note: l-diversity and t-closeness apply to sensitive attributes and do not provide protection for quasi-identifiers; attributes that are both sensitive and quasi-identifying require additional treatment.)
	B. Memorisation, Similarity and distance-based metrics	Hitting Rate (Common Row Proportion)	Measures the direct percentage of exact matching records (overlapping rows) between the synthetic and source data.
		Close Value Ratio	Assesses the probability of having "blurry matches" or similar values between synthetic and source data, defined by a distance threshold.
		Similarity Ratio (ε-identifiability)	Tests whether less than an ε ratio of synthetic observations are "similar enough" to those in the original dataset, often measured using weighted Euclidean distance.
		Nearest Neighbour Accuracy (Adversarial Accuracy)	Evaluates the proximity of a point in the original distribution (P_R) to its nearest counterpart in the synthetic distribution (P_S); Values close to 0.5 are sometimes interpreted as indicating difficulty distinguishing real from synthetic data. However, similarity-based metrics have been shown to miss certain classes of privacy leakage and do not provide bounded privacy guarantees; favourable values should therefore be interpreted cautiously and not taken as evidence of privacy safety [Stadler et al., 2023].
	C. Distinguishability metrics	Data Likelihood	Measures the likelihood of synthetic data belonging to the source data distribution, often using Bayesian Networks or Gaussian Mixture Models.
		Detection Rate	Assesses the difficulty of distinguishing source data from synthetic data using machine learning models like logistic regression.

II. Adversarial Metrics (Attack-Based) (Metrics that involve applying actual privacy attacks and measuring the success ratio, offering a definition of "practical privacy".)	**A. Singling out attacks**	Singling Out Attack (*Univariate Attack)*	Observes the uniqueness of a single attribute (predicate)in the synthetic data to assess whether rare values may be derived from the original data (univariate singling-out attack).
		Singling Out Attack (*Multivariate Attack)*	Involves combinations of multiple attributes (predicates).
	B. Record linkage attacks	Public-Public Linkage	Uses the synthetic dataset (S) to establish connections between records found in two separate external datasets (X_1 and X_2).
		Public-Synthetic Linkage	Links rows in the synthetic dataset (S) to an external dataset (X'), which may be public or privately held by an attacker, using matching criteria, serving as a basis for inference attacks.
	C. Attribute Inference Attacks (AIA)	Exact Match AIA	Determines the value of a missing target attribute by precisely matching overlapping quasi-identifiers (Q) between the synthetic data and the target records.
		Closest Distance AIA	Infers the missing sensitive value by identifying the single most similar data point (k=1) in S to the target record (equivalent to a K-Nearest Neighbour model).
		Nearest Neighbours AIA	Deduces the sensitive value by examining the k nearest neighbours (k > 1) in the synthetic dataset.
		ML Inference AIA	Attackers train a predictive machine learning model on S and use it to predict the target attributes of the target records.
	D. Membership Inference Attacks (MIA)	Closest Distance MIA	Infers membership if the target record is significantly more similar to the synthetic data than to unrelated data.
		Nearest Neighbours MIA	Relaxes the criteria of Closest Distance MIA to include proximity to k neighbours (k > 1).(Note: This attack remains similarity-based and, as with other distance-based approaches, may fail to detect certain classes of privacy leakage and does not provide bounded privacy guarantees; results should therefore be interpreted cautiously [Desfontaines, 2024; Stadler et al., 2023].
		Probability Estimation MIA	A hypothesis testing method assessing the likelihood that a target record belongs to the synthetic data distribution (and thus the original data distribution).
		MIA Shadow Model	Adversaries create "generative shadow models" using reference datasets (one including the target record, labelled 1; one excluding it, labelled 0) to train a classifier that distinguishes membership. (Note: This represents one class of attack within a rapidly evolving landscape of membership inference, reconstruction, and privacy auditing techniques. New attacks and auditing methods continue to be proposed)

The Limitations of Common Metrics

While the metrics listed above are frequently cited in the literature and used in practice, it is imperative to understand their significant limitations. The privacy engineering and security research communities have raised serious questions about the reliability of many of these evaluation methods, particularly those that rely on similarity or provide only average-case results.

- **The Inadequacy of Similarity-Based Metrics**

  Similarity and distance-based metrics (such as Nearest Neighbour Accuracy) are widely used because they are intuitive and simple to compute. However, a growing body of research demonstrates that these metrics can fail to detect serious privacy leakage and do not provide bounded guarantees of privacy loss [5, 6]. An attacker can craft attacks that completely bypass similarity-based detection. As Stadler et al. conclude in their work, "The Inadequacy of Similarity-based Privacy Metrics":

  > *Our work shows that similarity-based metrics are not a reliable measure of privacy. We have demonstrated that it is possible to craft a generator that produces synthetic data that is both useful and private according to similarity-based metrics, but which is vulnerable to a simple and effective attack.*

  These metrics are fundamentally concerned with the *average* case, whereas privacy risk is a worst-case problem, concerned with the potential for harm to any single individual. Therefore, while a poor result on a similarity metric may indicate a likely problem (a weak negative signal), a strong positive result provides no meaningful assurance of safety.

- **The Problem with Average-Case Metrics like F1**

  Membership inference attacks are often framed as a binary classification task, and summary statistics like the F1 score (the harmonic mean of precision and recall) are used to characterize the *average* performance of a given attack. However, this approach has several critical flaws when used as a privacy metric:

  - F1 is an aggregate, utility-oriented metric: It was designed to measure the effectiveness of a classifier on average, not to quantify individual privacy risk.
  - It reflects average, not individual, risk: A low F1 score across an entire dataset can easily mask the fact that a small number of vulnerable individuals are perfectly re-identifiable. Privacy is concerned with the outlier, not the average.
  - A low F1 score does not guarantee safety: As with other similarity metrics, a result that appears "good" (i.e., close to random guessing) does not prove the absence of privacy leakage; it only proves that the specific attack method used did not succeed on average.
  - A high F1 score clearly indicates vulnerability: The only reliable signal from an F1 score is a bad one. If an attack achieves a high F1 score, it is a clear indicator of a significant privacy failure.

For these reasons, the F1 score and other aggregate metrics should be treated as just one signal among many and never as a sufficient condition for declaring a dataset private.

Differential Privacy: A Formal Approach

In contrast to the empirical, attack-based metrics described above, Differential Privacy (DP) offers a formal, mathematical framework for reasoning about and bounding privacy loss. DP is a property of the *process* used to generate data, not a property of the output dataset itself. It provides a probabilistic guarantee that the output of a computation is nearly indistinguishable whether or not any single individual's data is included in the input [14].

Pure vs. Approximate Differential Privacy

A mechanism can satisfy one of two main definitions of differential privacy:

- **Pure ε-DP (Epsilon-DP):** This is the strictest form. A randomized algorithm M satisfies ε-DP if for any two adjacent datasets D_1 and D_2 (differing by one individual), and for any possible output S, the following holds:

  $$P[M(D_1) \in S] \leq e^{\epsilon} P[M(D_2) \in S]$$

  The privacy budget, ε (epsilon), is a non-negative parameter that controls the level of privacy. A smaller ε provides a stronger privacy guarantee. It is critical to understand that the privacy loss scales *exponentially* with epsilon. A seemingly small linear increase in ε can lead to a massive increase in privacy risk. For example, an ε of 3 corresponds to a 20x increase in the probability of distinguishing outputs, while an ε of 5 corresponds to a ~150x increase. For this reason, many practitioners consider ε values much greater than 1 to offer little meaningful privacy protection.

- **Approximate (ε, δ)-DP (Epsilon-Delta-DP):** Many real-world systems, including most synthetic data generators, satisfy this slightly weaker definition. The guarantee is modified as follows:

  $$P[M(D_1) \in S] \leq e^{\epsilon} P[M(D_2) \in S] + \delta$$

  The δ (delta) parameter represents the probability that the ε-DP guarantee fails to hold. It allows for a small chance of a more significant privacy breach. For the guarantee to be meaningful, δ must be a very small number, typically smaller than the inverse of the dataset size (1/n) [15]. The inclusion of δ is often necessary to handle scenarios where the set of possible outputs is not fixed in advance (e.g., new categories appearing in a free-text field), which can create "distinguishing events" that would otherwise violate pure ε-DP [15].

Differential Privacy in Practice

While DP provides a powerful theoretical guarantee, it should not be treated as a "tick-box" solution. Real-world implementations may fail to satisfy their theoretical claims due to design flaws, incorrect assumptions about the data, or simple software bugs. As one study attempting to reproduce a DP-based synthetic data generator found, neither the utility nor the privacy claims were reproducible in practice, with the implementation containing multiple privacy violations [16].

Privacy Auditing: Verifying Claims Empirically

The potential gap between theoretical promises and practical reality highlights the need for privacy auditing. Auditing refers to the empirical evaluation of a system to verify that it behaves consistently with its stated privacy claims, whether it is based on DP or other methods [17, 18].

Auditing complements theoretical analysis by actively searching for concrete failures and unexpected vulnerabilities. It does not *prove* the absence of privacy leakage, but it can provide strong evidence of its presence. Auditing techniques can include:

- Membership inference attacks
- Attribute inference attacks
- Reconstruction attacks
- Testing the mechanism on neighboring datasets that differ by one individual

Canary-Based Auditing

Traditional auditing methods that require training many models on neighboring datasets can be computationally prohibitive. A more recent and efficient approach involves using canaries [17]. This technique, often called "Privacy Auditing with One (1) Training Run," works as follows:

1. Inject Canaries: A set of carefully constructed, artificial records (canaries) are injected into the training data.
2. Train Model: The synthetic data generation model is trained on this augmented dataset.
3. Test for Leakage: Auditors then test whether the canaries can be detected or their information can be reconstructed from the synthetic output.

If the canaries, which were designed to be unique or rare, are detectable in the output, it provides concrete evidence of information leakage or memorisation by the model. This approach offers a practical, efficient way to move beyond theoretical claims and empirically test the privacy of a synthetic data generator.

Practical Considerations for Privacy Evaluation

While no single, universally accepted evaluation workflow exists, the following considerations, drawn from emerging best practices and expert recommendations [19], can help guide a robust and context-aware privacy assessment. The key is to build a portfolio of evidence rather than relying on a single metric or score.

- ## Defining the Scope of Evaluation

  Guideline: Base evaluations on realistic Quasi-Identifiers (QIs). Disclosure vulnerability is often evaluated based on a set of QIs that represent an adversary's

likely background knowledge. This set should be determined by the data controller based on the specific data and its context.

Consideration: It is useful to distinguish between the set of QIs used for privacy protection (e.g., for k-anonymity) and those used for privacy evaluation. For protection, a conservative approach that treats most attributes as QIs is often warranted [1]. For evaluation, however, the goal is to model a realistic attack. Requiring an evaluation metric to match on an overly broad set of QIs may not reflect how a real-world adversary operates (who often seeks the *minimum* information needed to attack) and could lead to an underestimation of the true privacy risk.

- ## Evaluating the Entire Dataset

Guideline: Calculate metrics across all records. To get a comprehensive view and avoid the potential bias of pre-selecting only a subset of "vulnerable" records, it is generally advisable to evaluate the entire dataset.

Consideration: This approach must be paired with a careful choice of metric. When using metrics that produce an average score (such as the F1 score for membership inference), be aware that this average can obscure high-risk vulnerabilities affecting a small number of individuals. Privacy is fundamentally a worst-case concern—the key question is whether *any* individual is at significant risk, not what the average risk is. Therefore, it is crucial to supplement average scores with methods that can identify outliers or to use metrics that are inherently focused on individual, rather than average, risk.

- ## Evaluating Membership and Attribute Disclosure

Guideline: Assess both membership and attribute disclosure risks. A comprehensive evaluation should consider the risk that an adversary can determine if someone is in the dataset (membership) and the risk they can infer a sensitive attribute about them (attribute disclosure).

Consideration: Many common metrics for these tasks are based on similarity or produce aggregate scores like F1. As discussed previously, such metrics provide asymmetric information: a poor score is a clear signal of a privacy failure, but a good score is not a guarantee of safety [5, 6]. They are effective for finding problems but not for proving their absence. Therefore, they should be treated as one part of a larger testing strategy, not as a definitive measure of privacy.

- ## Validating Formal Guarantees

Guideline: Empirically validate Differential Privacy (DP) guarantees. Even when a synthetic data generator claims to be differentially private, it is wise to perform empirical evaluations, especially if the privacy budget (ε) is not very close to zero.

Consideration: This practice is essential because a gap can exist between theoretical promises and real-world implementations [16]. An empirical audit does not replace the formal guarantee but acts as a crucial verification step to ensure the

claimed privacy properties hold in practice. This reinforces the need for privacy auditing as a standard part of the governance process.

- ### Accounting for Randomness

  Guideline: Report metrics from multiple generation runs. Synthetic data generation is a stochastic process. To account for this variation, it is good practice to generate multiple independent synthetic datasets and report on the variation in privacy metrics across them.

  Consideration: When reporting results from multiple runs, it is important not to obscure the worst-case outcome. For instance, reporting the *average* F1 score over ten runs provides less information about risk than reporting the maximum F1 score observed in any single run. The focus should remain on understanding the highest level of vulnerability observed, as this better reflects the potential risk to individuals.

Future Directions and Open Challenges

The field of privacy evaluation for synthetic data is far from settled. The ongoing debate between different research communities highlights the lack of consensus on fundamental questions of risk and measurement. As new generative models, attacks, and auditing techniques emerge, the landscape will continue to shift. Key areas requiring further research and development include:

- **Beyond Average-Case Metrics**: There is a critical need to develop and validate empirical privacy metrics that can effectively capture individual and worst-case risk, moving beyond the limitations of averaging metrics like F1.
- **Practical Privacy Auditing**: While techniques like canary-based auditing are promising, more work is needed to develop automated, accessible, and reproducible tools to support robust privacy auditing in practice.
- **Interpreting Differential Privacy Parameters**: The practical implications of ε and δ values in differential privacy remain poorly understood by many practitioners. Clearer guidance is needed on how privacy budgets are selected, composed over successive queries, and communicated to non-experts.
- **Successive Data Releases**: Methods are needed to better understand and manage the cumulative privacy loss that occurs over successive or repeated releases of synthetic data, including the risks from differencing attacks.
- **Complex Data Types**: Most current evaluation methods focus on static, tabular data. Extending robust evaluation techniques to other data modalities, such as time-series, longitudinal data, and free text, is a significant open challenge. The use of metadata and labels in longitudinal data, for instance, can inadvertently create distinguishing events that undermine privacy protections.

Conclusion: Towards a Responsible Practice

Privacy evaluation in synthetic data demands a systematic, evidence-based approach that moves beyond simplistic metrics and unsubstantiated claims. This appendix has provided a critical overview of the current landscape, highlighting not only the available tools but also their significant and often-underappreciated limitations. Perfect privacy is an unattainable goal; the true objective is to understand, quantify, and manage residual vulnerabilities in a manner appropriate to the specific context.

Organisations implementing synthetic data must recognise that privacy evaluation is not an optional step but an essential component of responsible data governance. There is no single workflow or universal set of metrics that can guarantee safety. Instead, practitioners should adopt a mindset of healthy scepticism and build a portfolio of evidence, drawing from theoretical analysis, adversarial testing, and empirical auditing.

The fundamental principles of rigorous evaluation (realistic threat modelling, transparency of assumptions, and empirical validation of claims) provide a robust foundation. As the field evolves, these principles will be the most reliable guide for navigating the complex trade-offs between data utility and the fundamental right to privacy.

References

[1] Culnane, C., & Leins, K. (2020). Misconceptions in Privacy Protection and Regulation. *IEEE Security & Privacy, 18*(3), 63-68.

[2] 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), Recital 26.

[3] Garfinkel, S. (2015). *NISTIR 8053: De-Identification of Personal Information*. National Institute of Standards and Technology.

[4] Based on feedback from Chris Culnane, noting that plain hashing is insufficient and recommending keyed hashing algorithms like HMACs with appropriate key management.

[5] Stadler, T., Oprisanu, B., & Troncoso, C. (2023). *The Inadequacy of Similarity-based Privacy Metrics: Privacy Attacks Against "Truly Anonymous" Synthetic Datasets*. arXiv preprint arXiv:2312.05114.

[6] Desfontaines, D. (2024). *The Bad, the Ugly, and the Good (Maybe)*. USENIX Conference on Privacy Engineering Practice and Respect (PEPR'24). Available at: https://desfontain.es/blog/bad-ugly-good-maybe.html

[7] Stadler, T., et al. (2025). *Synth-MIA: A Testbed for Auditing Privacy Leakage in Tabular Data Synthesis*. arXiv preprint.

[8] Trudslev, F. M., Lissandrini, M., Rodriguez, J. M., Bøgsted, M., & Dell'Aglio, D. (2025). A Review of Privacy Metrics for Privacy-Preserving Synthetic Data Generation. *arXiv preprint arXiv:2507.11324*.

[9] Osorio-Marulanda, P. A., Epelde, G., Hernandez, M., Isasa, I., Reyes, N. M., & Iraola, A. B. (2024). Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access, 12*, 88048-88074.

[10] Annammalai, M., et al. (2024). *What do you want from theory alone? Experimenting with Tight Auditing of Differentially Private Synthetic Data Generation*. USENIX Security Symposium.
[11] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 557-570.
[12] Based on feedback from Chris Culnane, noting that l-diversity and t-closeness apply to sensitive attributes, not quasi-identifiers.
[13] Desfontaines, D., Haney, S., & Pujol, D. (2026). *The search for better empirical privacy metrics*. Available at: https://desfontain.es/blog/better-empirical-privacy-metrics.html
[14] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265-284.
[15] Desfontaines, D. (2026). *Almost differential privacy*. Available at: https://desfontain.es/blog/almost-differential-privacy.html
[16] Ganev, G., Angelov, A., & Culnane, C. (2025). The Elusive Pursuit of Reproducing PATE-GAN: Benchmarking, Auditing, Debugging. *Transactions on Machine Learning Research*.
[17] Steinke, T., Nasr, M., & Jagielski, M. (2023). *Privacy Auditing with One (1) Training Run*. Advances in Neural Information Processing Systems (NeurIPS).
[18] Ping, H., Stoyanovich, J., & Howe, B. (2020). *Synthetic Data – Anonymisation Groundhog Day*. arXiv preprint arXiv:2011.07018.
[19] Pilgram, L., Dankar, F. K., Drechsler, J., Elliot, M., Domingo-Ferrer, J., Francis, P., ... & El Emam, K. (2025). A consensus privacy metrics framework for synthetic data. *Patterns*.

Further Resources

HealthStats NSW: Privacy issues and the reporting of small numbers

CSIRO & OAIC, *The De-Identification Decision-Making Framework*. Available at: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-decision-making-framework (the OAIC notes that, as this guide was produced in 2017, certain information it contains may now be out of date)

Office of the Victorian Information Commissioner (OVIC), *The Limitations of De-Identification – Protecting Unit-Record Level Personal Information*, available at: https://ovic.vic.gov.au/privacy/resources-for-organisations/the-limitations-of-de-identification-protecting-unit-record-level-personal-information/

Office of the Information Commissioner Queensland, *Report on Privacy and Public Data: Managing re-identification risk*, available at: https://www.oic.qld.gov.au/__data/assets/pdf_file/0016/43045/Privacy-and-public-data-managing-re-identification-risk.pdf

ISO/IEC 27559:2022