APPENDIX 7: De-identification techniques and Evaluation of Privacy in Synthetic Data

De-identification techniques

De-identification aims to break the link between a dataset and an individual in the real world, so that the disclosure of a fact (such as that a patient is being treated at this hospital for HIV) cannot be linked back to an identified individual (the patient is Sally Citizen).

The harm being prevented here is known as 'identity disclosure'. Identity disclosure - which occurs when data is re-identified - can arise in one of two ways: by either matching a person to data, or matching data to a person. Checking the robustness of de-identification techniques should involve testing your dataset for both these types of re-identification risk.[47]

> "De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting dataset."[48]

There is no single 'correct' way to de-identify data. It is an exercise in risk management. Re-identification risks will differ according to the type of data, its context, and other factors. Trade-offs need to be made between minimising the risk of re-identification, and maximising the value of the data. The wrong de-identification method can fail to reduce privacy risk, and/or decrease data utility. As such, the information outlined in this section should be considered as general guidance about de-identification techniques, and not as a specific plan to achieve a 'de-identified' synthetic health dataset.

Examples of de-identification techniques include:

- aggregation
- suppression (remove identifiers or other overtly identifying data fields)
- generalisation (e.g. replace exact date of birth with a date range like '35-44 year olds')
- pseudonymisation (replace direct identifiers with statistical linkage keys (SLKs), or encrypt or hash identifiers), and
- perturbation (adding noise, micro-aggregation or data-swapping).

[47] When considering re-identification risks, the GDPR makes clear that the identifiability of data should not be considered in a vacuum. Instead, "account should be taken of all the means reasonably likely to be used … to identify the natural person directly or indirectly", including "objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments"; see GDPR Recital 26.

[48] Simson L. Garfinkel, NISTIR 8053: De-Identification of Personal Information, National Institute of Standards and Technology, US Department of Commerce, 2015, p.1; available at http://dx.doi.org/10.6028/NIST.IR.8053.

- Evaluation of Privacy in Synthetic Data

Evaluation of Privacy in Synthetic Data

Introduction

Privacy protection is the first principle guiding synthetic data generation, even though synthetic data ultimately represents a trade-off between fidelity, utility, and privacy. Privacy evaluation is therefore 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. Some generative models incorporate privacy-preserving mechanisms by design (for example, DPGAN and PATEGAN implement differential privacy, while ADSGAN targets re-identification risks [1-3]), yet issues such as model overfitting and the inadvertent retention of sensitive patterns can compromise privacy. Consequently, 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 complements the main text by providing a concise overview of current approaches to evaluating privacy in synthetic data. There is no single accepted definition or universal measure of privacy risk; existing metrics capture only partial aspects of it. The aim here is to clarify what these measures assess and what they do not, emphasising that no single score can demonstrate overall privacy safety. The discussion is conceptual rather than implementation-focused; we do not provide algorithmic details or code, although several of the described metrics may be available through open-source R or Python packages.

A wide range of metrics has been proposed to evaluate privacy in synthetic data [4, 5], with various classification schemes presented in the literature. One of the most practical groupings is summarised in Table 1 [6]. These metrics span traditional re-identifiability measures such as *k-anonymity*, *l-diversity* and *t-closeness* to distance-based metrics like Nearest Neighbour Distance Ratio (NNDR) and adversarial approaches, including membership inference attacks. While some studies recommend using individual metrics or combinations of them, others advocate for broader frameworks that assess privacy alongside utility and fidelity [7-9]. Although these methods aim to capture privacy risks, none fully resolve the inherent trade-off between privacy and usability. Developing rigorously validated, context-aware evaluation criteria remains an open challenge [10].

Despite these efforts, there remains no unified standard defining what constitutes adequate privacy protection or how to interpret the results of these evaluations in practice. This lack of consensus is further complicated by the variation of synthetic data generation processes.

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 if every record is indistinguishable from at least k_1 other records based on quasi-identifiers.
		l-Diversity	An extension of k-Anonymity ensuring that sensitive attributes within each anonymised group have at least l distinct values.
		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.
	B. Memorisation 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); an optimal value of 0.5 suggests indistinguishability.
	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	**A. Singling Out Attacks**	Singling Out Attack (*Univariate Attack)*	Observes the uniqueness of a record or attribute combination in the synthetic data to assess the likelihood that this uniqueness is derived from the original data. (Focuses on rare values of a single attribute (predicate).)
		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 a public dataset (X') using matching criteria, serving as a basis for inference attacks.

success ratio, offering a definition of "practical privacy".)	**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).
		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.

Acknowledging the absence of universal guidelines with predefined thresholds for privacy assessment, this appendix outlines a practical approach to evaluating privacy in synthetic data. The proposed guidance is based on a published study [11] developed in collaboration with international experts in privacy and synthetic health data, aiming to provide actionable methods while addressing common misconceptions.

The suggested framework stems from a rigorous Delphi consensus process involving 13 global privacy specialists. Through three structured rounds, the panel reached agreement on ten core recommendations. This consensus-driven approach helps bridge the critical gap in standardised methods for assessing privacy in synthetic data.

Fundamental Concepts

Privacy evaluation in synthetic data focuses on three primary disclosure types:

1. **Identity Disclosure**: Occurs when a synthetic record can be linked to a specific individual. Whilst synthetic data generation should inherently protect against this by design, overfitting can still create vulnerabilities. It has been argued that identity disclosure is less relevant for synthetic data, since synthetic records are artificial and do not correspond to real individuals; the more substantive privacy risks arise from membership and attribute disclosure, which can reveal information about real data subjects. Hence, we focus on these two measures in what follows. Attempts to measure identity disclosure often rely on record-level similarity metrics, but these lack clear interpretation and tend to conflate with membership or attribute disclosure depending on whether correctness of sensitive information is considered.

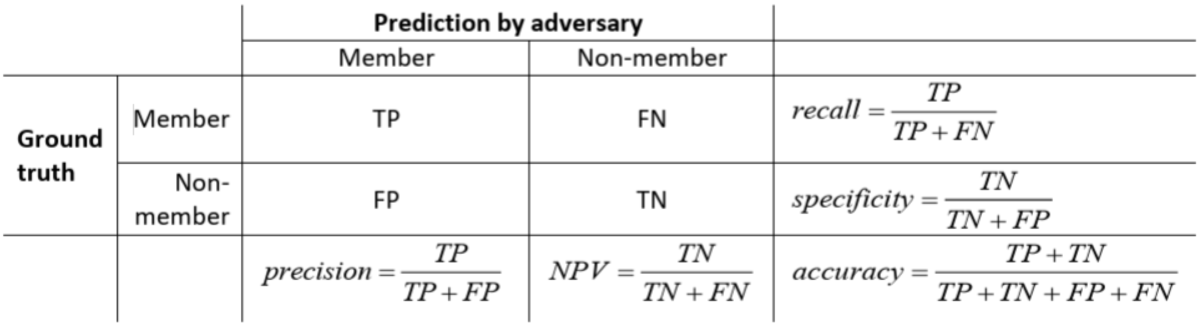
2. **Membership Disclosure**: This occurs when it is possible to determine whether a specific individual's data was included in the training dataset used to generate the synthetic data. This becomes particularly concerning when the training data contains sensitive information (e.g., participants in a clinical trial for a specific condition). Membership disclosure could be treated as a binary classification task where correctly identifying a data member is a "true positive." To quantify this vulnerability, metrics such as the F1 score can be used, combining precision and recall to evaluate the performance of membership inference attacks. A baseline metric, F_{naive}, reflects the expected success of an adversary guessing membership status without using the synthetic data:
   a. **F1 score**: The F1 score is the harmonic mean of precision and recall and therefore measures the balance between correctly identifying positives and avoiding false positives.

   $$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

   b. **$F_{\text{naïve}}$**: The $F_{\text{naïve}}$ score represents the expected F_1 value under random guessing and serves as a baseline to assess whether a model’s performance reflects genuine predictive ability rather than chance.

$F_{na\ddot{\imath}ve} = \frac{2 \times p}{1+p}$ *(p = sampling fraction of the original dataset from the population)*

[11]



		Prediction by adversary		
		Member	Non-member	
Ground truth	Member	TP	FN	$recall = \frac{TP}{TP+FN}$
	Non-member	FP	TN	$specificity = \frac{TN}{TN+FP}$
		$precision = \frac{TP}{TP+FP}$	$NPV = \frac{TN}{TN+FN}$	$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Confusion Matrix in Membership Disclosure Vulnerability. FN: false negative; FP: false positive; NPV: negative predictive value; TN: true negative; TP: true positive.

3. **Attribute Disclosure**: This occurs when an adversary can infer sensitive information about an individual based on the synthetic data. This requires careful distinction between legitimate knowledge generation (learning population patterns) and privacy violations (learning specific information about individuals in the training data).

Other definitions and formulas:

Differential Privacy: Differential privacy (DP) is a mathematical framework that guarantees the output of an analysis is nearly indistinguishable regardless of whether any single individual's data is included in the input. This provides a quantifiable and robust privacy guarantee against a wide range of attacks. Unlike the disclosure risk discussed previously (which are properties of the dataset), DP is an a priori feature of the data generation or analysis process.

Privacy budget (ϵ): The privacy budget is a non-negative parameter that is **set** to control the level of privacy. It's a "budget" in the sense that it's a resource we spend each time we query the data. A randomised algorithm M is **ε-differentially private** if for all pairs of neighbouring datasets D and D' (differing by just one individual), and for all possible outputs $S \subseteq Range(M)$: $\frac{\Pr\,[M(D) \in S]}{\Pr\,[M(D') \in S]} \leq e^{\epsilon}$

Common Misconceptions to Avoid

- **Misconception 1: Synthetic Data Is Inherently Private**

  Synthetic data generation does not automatically guarantee privacy. Models can overfit, memorising specific details rather than learning general patterns, potentially encoding information about individuals in the training data.

- **Misconception 2: Record-Level Similarity Indicates Privacy Risk**

Unlike anonymised data, the similarity between synthetic and source records lacks precise interpretation for privacy risk. A synthetic record might resemble multiple training records, or similarity in quasi-identifiers might not translate to similarity in sensitive attributes. Stand-alone similarity metrics is inadequate to report privacy in synthetic data.

- **Misconception 3: All Attributes Should Be Considered in Privacy Evaluation**

  Assuming an adversary knows all attributes is not necessarily a worst-case scenario. Simulations demonstrate that considering all variables may underestimate disclosure vulnerability. Privacy evaluation should focus on quasi-identifiers relevant to the specific context.

- **Misconception 4: Large Differential Privacy Budgets Ensure Privacy**

  The privacy budget ε lacks uniform translation to empirical privacy unless very close to 0. Values ranging from 0.1 to over 18 in published work demonstrate this parameter's inconsistent interpretation. Empirical evaluation remains necessary even for differentially private synthetic data.

Consensus-Based Recommendations

- **R1: Base Evaluations on Quasi-Identifiers**

  Disclosure vulnerability metrics should be based on quasi-identifiers (QIs) rather than all attributes. QIs represent the adversary's realistic background knowledge and vary by context. The data controller should determine appropriate QIs, though treating all attributes as QIs remains an option if computationally feasible.

- **R2: Evaluate All Records**

  Metrics should be calculated for all records, not pre-selected "vulnerable" subsets. Pre-selection introduces bias and may miss actual vulnerabilities. Maximum vulnerability can only be determined through a comprehensive evaluation.

- **R3: Avoid Stand-Alone Similarity Metrics**

  Record-level similarity metrics without connection to membership or attribute disclosure lack meaningful interpretation and should not be used for privacy reporting.

- **R4: Align Membership Disclosure with Threat Models**

  Evaluate membership disclosure only when the assumptions hold, specifically, when adversaries would learn something new about targets from the same population as the training dataset. Misalignment between threat models and metrics can produce misleading vulnerability estimates.

- **R5: Report Prevalence-Adjusted Scores**

  F1 scores for membership disclosure must be reported relative to an adversary guessing membership (F_{rel}) to account for prevalence dependence. This provides a consistent interpretation across different member prevalence levels.

  $$F_{rel} = \frac{F1 - F_{naïve}}{1 - F_{naïve}},$$

- **R6: Limit Attribute Disclosure to Members**

  Meaningful attribute disclosure applies only to dataset members. Accurate predictions about non-members represent legitimate knowledge generation rather than privacy violations.

- **R7: Use Non-Member Baseline**

  Employ relative vulnerability metrics with non-member baselines to distinguish between information learned from dataset membership versus population membership.

- **R8: Apply Dual Thresholds**

  Consider relative vulnerability unacceptably high only when absolute vulnerability also exceeds its threshold. High relative vulnerability with low absolute accuracy doesn't constitute meaningful disclosure.

- **R9: Validate Differential Privacy Empirically**

  Unless ε is close to 0, empirical evaluation using standard metrics remains necessary even for differentially private synthetic data.

- **R10: Report Stochastic Variation**

  Report metrics for both individual synthetic datasets and multiple generations (averaged with variation measures) to account for the stochastic nature of synthetic data generation.

Practical Evaluation Guides

Guide for Membership Disclosure Evaluation

1. **Define the Threat Model**
   - Specify whether targets come from the training population or the broader population
   - Ensure attack dataset prevalence matches threat model assumptions

2. **Calculate Baseline Metrics**
   - Determine naive membership guess ($F_{naïve}$) based on member prevalence
   - Account for prevalence in interpretation
     $F_{naïve} = \frac{2 \times p}{1+p}$ *(p = sampling fraction of the original dataset from the population)*

3. **Compute Relative Metrics**
   - Calculate $F_{rel} = \frac{F1 - F_{naïve}}{1 - F_{naïve}}$
   - Suggested threshold: $F_{rel} \leq 0.2$ (requires contextual adjustment)

4. **Document Assumptions**
   - Explicitly state the target population
   - Clarify what membership reveals in the specific context

Guide for Attribute Disclosure Evaluation

1. **Define Quasi-Identifiers and Sensitive Attributes**
   - Based on realistic adversary knowledge
   - Consider context-specific factors

2. **Establish Non-Member Baseline**
   - Use holdout or external data
   - Calculate prediction accuracy for non-members ($A_{non-members}$)
     $A_{non-members}$: accuracy of predicting sensitive information for individuals who were not part of the Synthetic Data Generation (SDG) training dataset. It is the probability that the presumed sensitive target value is correct given that the attack record is not a member.

3. **Evaluate Member Vulnerability**
   - Calculate prediction accuracy for members ($A_{members}$)
     $A_{members}$ is represents the absolute accuracy of predicting a sensitive attribute for individuals who were part of the Synthetic Data Generation (SDG) training dataset. It is analogous to the probability that the presumed sensitive target value is correct given that the attack record is a member.

    - Compute relative accuracy: $(A_{rel} = A_{members} - A_{non-members})$

4. **Apply Dual Thresholds**
    - Absolute threshold: $A_{members}$≤ 0.6 (exceeds poor prediction)
    - Relative threshold: A_{rel} ≤ 0.15 (meaningful difference)
    - Both must be exceeded for unacceptable vulnerability

Evaluating Multiple Synthetic Datasets

Given the stochastic nature of synthetic data generation:

- **For Model Evaluation**
    - Generate multiple synthetic datasets (minimum 10 recommended)
    - Report average vulnerabilities and standard deviations
    - Document worst-case scenarios

- **For Data Release Decisions**
    - Evaluate the specific dataset(s) to be released
    - Consider maximum vulnerability across evaluations

Implementation Considerations

- **Computational Efficiency**
    - Start with domain-informed QI subsets
    - Progressively expand evaluation if resources permit
    - Balance thoroughness with practical constraints

- **Threshold Determination**
    - Thresholds must be:
    - Based on empirical precedents where available
    - Adjusted for context (data sensitivity, potential harm, consent)
    - Documented with clear justification
    - Subject to refinement as evidence accumulates

- **Reporting Requirements**

4. Comprehensive reporting should include:
    - Absolute metrics: Raw vulnerability measurements
    - Relative metrics: Baseline comparisons
    - Variation measures: Standard deviations across generations
    - Worst-case estimates: Maximum vulnerabilities observed
    - Context documentation: Threat models, assumptions, and adjustments

Future Directions

- **Immediate Priorities**
    - Empirical validation of suggested thresholds across domains
    - Standardisation of meaningful ε values for differential privacy
    - Development of automated evaluation tools

- **Long-term Goals**
    - Context-aware threshold adjustment systems
    - Joint optimisation of utility and privacy
    - Extension to non-tabular synthetic data types

Conclusion

Privacy evaluation in synthetic data requires systematic, evidence-based approaches that move beyond simplistic metrics and unfounded assumptions. This framework provides practical guidance whilst acknowledging that perfect privacy remains unattainable. The goal is not to eliminate all risk but to quantify and manage residual vulnerabilities appropriately.

Organisations implementing synthetic data must recognise that privacy evaluation is not optional but essential. By following these consensus-based recommendations and practical guides, data controllers can make informed decisions about synthetic data generation and sharing, balancing the tremendous potential of synthetic data with appropriate privacy protections.

The framework acknowledges that synthetic data evaluation is an evolving field. As empirical evidence accumulates and new threats emerge, these methods will require refinement. However, the fundamental principles, rigorous evaluation, realistic threat modelling, and empirical validation, provide a robust foundation for responsible synthetic data governance.

References

1. Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739.* https://arxiv.org/abs/1802.06739
2. Jordon, J., Yoon, J., & Van Der Schaar, M. (2018, September). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*. https://openreview.net/pdf?id=S1zk9iRqF7
3. Yoon, J., Drumright, L. N., & 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. https://ieeexplore.ieee.org/abstract/document/9034117
4. 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*. https://arxiv.org/pdf/2507.11324
5. 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. https://ieeexplore.ieee.org/document/10568134
6. Liao, Q., Van Landuyt, D., & Joosen, W. (2025). Pick Your Enemy: A Survey on Privacy Threat Models of Synthetic Tabular Data. https://www.authorea.com/doi/full/10.22541/au.174893956.65391176
7. Folz, J., Vidanalage, M. D., Aufschläger, R., Almaini, A., Heigl, M., Fiala, D., & Schramm, M. (2025). Scoring System for Quantifying the Privacy in Re-Identification of Tabular Datasets. *IEEE Access*. https://ieeexplore.ieee.org/document/10973096
8. Hernandez, M., Osorio-Marulanda, P. A., Catalina, M., Loinaz, L., Epelde, G., & Aginako, N. (2025). Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Frontiers in Digital Health*, *7*, 1576290. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1576290/full
9. *Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., ... & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. Nature communications, 13(1), 7609.* *https://www.nature.com/articles/s41467-022-35295-1?fromPaywallRec=false*
10. Pierce, D. V., Li, Y., Greenshaw, A. J., Bailey, T., & Cao, B. (2025). Practical Steps in Implementing Privacy Measures With Synthetic Health Data. *World Medical & Health Policy*. https://onlinelibrary.wiley.com/doi/10.1002/wmh3.70023
11. 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*. https://www.sciencedirect.com/science/article/pii/S2666389925001680

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
Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework
https://www.iso.org/standard/71677.html

ISO/IEC 27554:2024
Information security, cybersecurity and privacy protection — Application of ISO 31000 for assessment of identity-related risk
https://www.iso.org/standard/71672.html

ISO/TS 14265:2024
Health informatics — Classification of purposes for processing personal health information
https://www.iso.org/standard/83447.html

ISO 25237:2017
Health informatics — Pseudonymization
https://www.iso.org/standard/63553.html