The Illusion of Anonymity: Quantifying and Managing Re-Identification Risk

Mia Willis

[1]

***Abstract*—Organisations collect, store, and release vast amounts of information about individuals in datasets. Malicious actors can attempt to re-identify individuals from this data, undermining our right to privacy. This project investigates whether quantitative re-identification risk measures can be used to develop practical guidance for organisations making data release decisions. A Python-based pipeline generates synthetic datasets and models data against different levels of attacker knowledge. Re-identification risk is assessed using privacy metrics including linkage, uniqueness, singling-out, and partial re-identification. K-means clustering is used to derive project-specific risk thresholds and classify configurations into LOW, MEDIUM, and HIGH risk bands. Analysis of the experimental results identified key factors influencing re-identification risk and the efficacy of different privacy interventions. These synthesised findings informed an organisation-facing decision framework that enables organisations to assess release risk without requiring them to perform technical re-identification analysis. This framework is intended to complement, rather than replace, existing qualitative privacy governance and guidance by providing a quantitative basis for assessing re-identification risk. Evaluation of the final framework with experts in privacy, cybersecurity, and data ethics found that the framework was clear, practical, and provided a strong proof-of-concept for integrating quantitative risk metrics into organisational release decisions.**

I. Introduction

A. Problem Context and Motivation

Organisations across all sectors collect and release structured datasets for research, service delivery, decision-making, and innovation. While these datasets can provide value, they also contain information that can be used to identify individuals, creating potential privacy harms. To mitigate this risk, organisations apply privacy-preserving transformations to reduce the risk of re-identification [1]. A common method for privacy protection is removal of direct identifiers, but this does not necessarily eliminate re-identification risk. Privacy-preserving techniques can reduce risk, but their efficacy depends on the configuration and the auxiliary knowledge at the attacker's disposal. Additionally, excessive transformation can reduce the usefulness of the data, meaning organisations face a privacy-utility trade-off decision, rather than simply anonymising as much data as possible. Organisations must decide what level of residual re-identification risk is acceptable for the intended use and release context.

Existing legislation and privacy guidance provide principles and obligations for protecting personal information [2], but there remains a gap between quantitatively measuring re-identification risk and translating those measurements into practical organisational guidance for dataset-release decisions. Legal compliance and quantitative risk assessment are related, but not interchangeable. Organisations must consider measured risk alongside other factors such as intended use and access context, data utility, organisational risk appetite, and Māori data governance when determining whether a dataset should be released to support an appropriate, defensible release decision.

B. Project Aim and Objectives

The aim of this project is to investigate whether quantitative re-identification risk measures could be translated into practical, context-aware guidance that complements existing qualitative privacy and governance guidance for organisational dataset-release decisions.

The following objectives guided the project development:

O1: Develop a Python-based pipeline capable of measuring re-identification risk indicators across structured datasets under defined privacy transformations and attacker models.

O2: Experimentally evaluate how anonymisation and differential privacy (DP) configurations affect observed re-identification risk.

O3: Derive project-specific risk bands from experimental results to support interpretation of observed re-identification risk.

O4: Identify and synthesise risk patterns and intervention effects by analysing how changes to age, postcode, gender, DP, and attacker capability affect observed risk.

O5: Translate experimental findings into an organisation-facing framework for supporting context-aware dataset-release decisions.

O6: Evaluate the clarity, utility, and practicality of the framework with privacy, cybersecurity, and data ethics experts.

C. Project Scope and Assumptions

This project is scoped as a research prototype and proof-of-concept, rather than a production-ready privacy assessment framework. The experimental analysis is limited to structured, synthetic datasets and a defined set of attributes, privacy transformations, and attacker knowledge assumptions.

The experiments use age, gender, and postcode as the

[1] This project was supervised by Dr. Qurrat Ul Ain and Dr. Andrew Lensen.

primary attributes for anonymisation interventions. Age is banded into increasingly broad ranges, postcode is truncated by reducing the number of retained digits, and gender is either retained or suppressed. Four pre-defined attacker models represent increasing levels of auxiliary information available to an attacker.

Privacy analysis is limited to the selected re-identification metrics, attacker models, and the anonymisation and DP configurations implemented in the project. DP is evaluated experimentally using a Laplace-noise mechanism and predefined privacy budgets (ε); this implementation is intended to investigate its effect on observed re-identification risk rather than provide a comprehensive production-ready DP implementation.

This project uses synthetic datasets with a defined structure and attribute set. The resulting risk classifications and intervention effects apply to the tested configurations and should not be assumed to generalise directly to other datasets, populations, or organisational contexts, as generalisability is out of scope.

The combination of risk aggregation and K-means clustering is used to derive project-specific risk bands from the experimental results. These bands are intended to provide a consistent basis for interpreting the experimental evidence; they are not universal or externally validated.

The resulting framework does not determine whether a dataset should be released solely from its quantitative risk classification. Release decisions remain dependent on factors including intended use, access context, data utility, organisational risk appetite, legal and sector requirements, and Māori data governance. The framework therefore explicitly incorporates these considerations into the release decision process rather than treating the quantitative risk band as the stand-alone release decision.

D. Contribution and Key Findings

The project has two complementary outputs: a Python-based experimental pipeline for measuring and analysing re-identification risk, and an organisation-facing framework that translates the resulting evidence into practical guidance for data-release decisions. The pipeline evaluates selected re-identification metrics across defined anonymisation, DP, and attacker configurations. These results are used to identify risk patterns, derive project-specific LOW, MEDIUM, and HIGH risk bands, and evaluate the effects of different privacy interventions. The resulting framework presents this evidence in a form that can be used without requiring organisations to perform the underlying technical analysis themselves, while incorporating intended use, access context, data utility, organisational risk appetite, legal and sector requirements, and Māori data governance into the final decision process.

The empirical evaluation found that direct identifier removal alone did not reduce observed re-identification risk in the tested baseline: the raw and de-identified configurations both scored 0.971 under the strong attacker and 0.993 under the very strong attacker, classifying as HIGH risk. Among the tested anonymisation interventions, postcode truncation produced the largest and most consistent reduction in observed risk, occurring in 87.5% of strong-attacker and 100% of very-strong-attacker intervention tests. However, this intervention improved the risk band in 41.7% and 58.3% of tests, respectively, demonstrating that numerical risk reduction does not necessarily translate into a lower risk band. Age banding and gender suppression had more context-dependent effects. The DP experiments showed that observed risk depended on the starting anonymisation configuration, privacy budget, and attacker capability, and that increasing the privacy budget did not produce a universally monotonic change in observed risk.

These findings informed the design of the organisation-facing framework, which treats quantitative risk classification as evidence to support, rather than replace, broader privacy and governance decision-making.

Expert evaluation of the framework provided initial support for its clarity and practical usefulness as a decision-support tool. Participants identified the decision flow, integration of quantitative risk with broader governance considerations, and consideration of attacker knowledge as particularly useful, while also acknowledging the need for further evaluation and scope-widening before broader deployment. Evaluation indicated that the framework provides a strong proof-of-concept for translating quantitative re-identification evidence into organisational guidance.

E. Environmental and Sustainability Considerations

This project aligns with the UN's Sustainable Development Goal 16 [3], "Peace, Justice and Strong Institutions", which emphasises developing effective, accountable, and transparent institutions, and supporting responsive and inclusive decision-making. The project contributes to these principles by supporting evidence-informed privacy governance when organisations assess whether datasets are appropriately configured for release. The framework supports responsible data sharing and accountable decision-making by helping organisations consider quantitative re-identification risk alongside legal, organisational, utility, and Māori data governance considerations, balancing proportionate privacy protections against the value of data sharing and the costs of additional privacy measures.

The project also aligns with Sustainable Development Goal 10 [3], "Reduced Inequalities", particularly through its consideration of Māori data governance and collective interests in data-release decisions. Incorporating Māori data governance recognises that the potential impacts of data use and release may extend beyond individual privacy to include communities and groups. This supports a more inclusive approach to data governance by encouraging organisations to consider cultural context, collective interests, and potential harms alongside technical re-identification risk.

The project has limited direct environmental impact because it is primarily software-based and uses synthetic datasets, requiring no physical infrastructure beyond conventional computing resources. The use of synthetic data also avoids requiring real personal datasets during experimental development and evaluation of the pipeline.

F. Tools and Methodology

The project was implemented in Python, selected for its extensive support for data analysis, statistical processing, and privacy-risk experimentation. Pandas was used for dataset manipulation and analysis, while NumPy supported numerical operations and the implementation of the DP mechanism. scikit-learn was used for K-means clustering to derive the project-specific risk classification band thresholds. These libraries enabled the experimental pipeline to process multiple dataset configurations consistently, keeping the implementation reproducible and extensible. These open-source libraries avoided licensing costs and ran on a standard computer; the trade-off is that in-memory processing with pandas suits the 50,000-record datasets used here but would need revisiting for much larger datasets.

Development and experimentation were conducted using a standard machine running Windows and Python 3.13. The project was developed as a modular Python codebase, separating dataset generation, privacy transformations, risk metrics, experiment execution, and result analysis. This structure supported independent testing and modification of individual components and allowed experimental results to be reproduced across different configurations.

The technical implementation and experimental methodology are described in greater detail in Sections III and IV. The final organisation-facing framework was developed separately from the Python pipeline to ensure that the resulting release guidance could be used without requiring users to interact directly with the underlying technical analysis.

II. Literature Review

This section reviews the existing research surrounding re-identification risk and privacy-preserving techniques. It then examines quantitative risk assessment and privacy governance across jurisdictions to identify the gap between measuring residual risk and supporting practical dataset-release decisions.

A. Re-identification Risk and Attacker Knowledge

The removal of direct identifiers, such as names or identification numbers, does not necessarily prevent re-identification. Sweeney [4] demonstrated that combinations of quasi-identifiers (QIs), such as birthdate, gender, and postcode, can distinguish individuals within a dataset. The well-known Netflix Prize attack by Narayanan and Shmatikov [5] demonstrated that publicly available or externally obtained information can be combined with released datasets to enable re-identification. The systematic review in [6] identified a range of re-identification attacks against supposedly de-identified health datasets. These studies establish that re-identification risk depends not only on the attributes and attribute transformations in the released dataset, but also on the auxiliary information available to the attacker. Consequently, assessing re-identification risk requires consideration of both the characteristics of the dataset and assumptions about the attacker's knowledge. Xia et al. [7] explicitly address this through adversarial modelling, demonstrating how re-identification risk varies substantially depending on the external information and resources available to the attacker. The work also discusses the limitation of assuming attackers possess extensive information; this may overestimate practical risk and lead to unnecessary data modification. This is significant because additional transformation can reduce data utility and may also affect the ability to retain context that is important for appropriate data governance.

Existing research establishes attacker capability as an important consideration in quantitative risk assessment. Sophisticated adversarial models require detailed assumptions about external datasets, attacker resources, and behaviour. Therefore, this project adopts a deliberately simplified attacker model; increasing attacker capability is represented by progressively increasing access to selected QIs. This approach allows the effects of attacker knowledge to be compared consistently across the experimental configurations, without attempting to simulate real-world attacks.

B. Privacy-preserving Techniques and Quantitative Risk Assessment

A range of privacy-preserving models and techniques, summarised in Table I, have been developed and documented to reduce disclosure and re-identification risk.

TABLE I
SUMMARY OF INVESTIGATED PRIVACY-PRESERVING MODELS AND TECHNIQUES

Model	Core idea
k-anonymity	Each record resembles at least k-1 others.
l-diversity	Sensitive values must be diverse within groups.
t-closeness	Sensitive attribute distributions within equivalence classes remain similar to overall distribution.
Differential privacy	Add noise to limit influence of individual records through defined privacy parameter.

These approaches [4,8-10] provide different mechanisms for protecting data, but no single model captures every aspect of contextual re-identification risk. Privacy models and transformations specify how data should be protected but do not necessarily indicate how much residual risk remains under a particular attacker scenario or whether that risk is acceptable under a particular release context.

Quantitative risk assessment provides a complementary perspective to existing guidance by measuring characteristics associated with re-identification. Uniqueness and equivalence-class measures provide indicators of record distinguishability, while linkage and singling-out approaches assess re-identification risk under assumptions about available attacker information [4, 6-7]. Existing research and guidance demonstrate the value of measuring residual risk rather than treating de-identification as a binary property. For example, NIST guidance [1] outlines best practices, recommending evaluating the potential risks created by de-identification and identifying re-identification studies and measurable

TABLE II
COMPARISON OF DE-IDENTIFICATION, ANONYMISATION, AND PRIVACY APPROACHES ACROSS SELECTED JURISDICTIONS

Jurisdiction	Key approach	Strength	Limitation relevant to this project
EU (GDPR) [13]	Contextual assessment of whether identification is reasonably likely, considering factors such as available means, cost, time and technology	Explicitly recognises that identifiability depends on circumstances and available means, incorporating an adversarial perspective	Provides no universal quantitative re-identification threshold or standardised risk metric for determining when risk is sufficiently low
UK (DPA 2018) [14]	Legal controls over re-identification, including an offence for knowingly or recklessly re-identifying de-identified personal data without appropriate authorisation	Provides explicit legal deterrence and enforcement against unauthorised re-identification	Primarily addresses the legal consequences of re-identification rather than providing a quantitative method for measuring residual release risk
Canada (PIPEDA / Privacy Act) [15-17]	Federal privacy frameworks assess whether information relates to an identifiable individual and recognise that information may become identifying when combined with other information; provincial regimes provide additional requirements	Recognises contextual identifiability and the role of linkage and auxiliary information; provincial developments provide more specific anonymisation requirements	Privacy requirements are distributed across federal and provincial frameworks, resulting in variation between jurisdictions and limited standardised quantitative guidance for dataset-release risk
US (HIPAA / NIST) [1,18]	Sector-specific de-identification approaches, including HIPAA Safe Harbor and Expert Determination, alongside broader NIST de-identification guidance	Provides practical and technically informed approaches, including prescriptive identifier removal and contextual assessment of identification risk	Requirements and approaches vary across sectors and jurisdictions, and HIPAA Expert Determination does not prescribe a universal numerical risk threshold and applies only to health data.
Australia (Privacy Act / APPs) [19-20]	Privacy Act 1988 and Australian Privacy Principles, supported by contextual de-identification guidance, assess whether an individual remains reasonably identifiable	Explicitly recognises re-identification risk and provides practical guidance for assessing and managing that risk in context	Guidance remains primarily contextual and does not establish a universal quantitative threshold for dataset-release risk
Aotearoa (Privacy Act / IPPs) [2, 21]	Privacy Act 2020 and Information Privacy Principles, supplemented by sector-specific guidance and Māori data governance principles	Provides broad principles-based privacy protections while Māori data sovereignty frameworks recognise collective interests, governance, provenance and control over Māori data	The Privacy Act does not establish a statutory definition for anonymisation, de-identification, re-identification, or a quantitative re-identification risk threshold

performance levels as possible components of de-identification practice.

Existing tools demonstrate how quantitative risk assessment can be integrated with anonymisation. ARX [11] provides privacy models, data transformations, utility analysis, and re-identification-risk analysis within an integrated anonymisation environment. It can explore transformations such as generalisation and suppression while assessing privacy and utility of potential outputs. Similarly, sdcMicro [12] combines disclosure-risk estimation with anonymisation techniques and information-loss measures, allowing risk and utility to be recalculated as transformations are applied. These tools demonstrate the technical feasibility of combining privacy-preserving transformations with quantitative risk assessment. However, their focus on technical anonymisation and disclosure-risk analysis leaves broader governance considerations, including Māori data sovereignty and appropriate data use in Aotearoa, largely unaddressed. Organisations must therefore interpret the results alongside these considerations when making context-specific dataset-release decisions.

This project takes a different approach. Rather than providing a general-purpose anonymisation tool like ARX [11] or sdcMicro [12], it uses controlled experiments to examine how selected anonymisation and DP configurations affect re-identification risk indicators under different attacker capabilities. The resulting evidence is then synthesised into a stand-alone decision-support framework that translates quantitative findings into practical guidance for organisations making dataset-release decisions. This project therefore complements existing technical tools by focusing on the translation of quantitative risk evidence into higher-level organisational release-decision guidance.

C. Privacy Governance and Data Release in Different Jurisdictions

Technical risk assessments serve a purpose within broader legal and governance requirements. Privacy legislation generally establishes obligations concerning the collection, use, storage, disclosure, and protection of personal information, but jurisdictions differ in how de-identification, identifiability, and re-identification are treated.

Across the jurisdictions examined in Table II, privacy and de-identification requirements generally focus on whether individuals remain identifiable, but the mechanisms used to assess this differ substantially. Some regimes, including the EU and Australia, explicitly require contextual consideration of the circumstances and means by which identification could occur [13, 20]. Several approaches also recognise that information may become identifying when combined with other available information [15-18], highlighting the importance of auxiliary information and potential attacker capability. However, the legislative and governance approaches examined do not establish a single, generally applicable quantitative method for assessing residual re-identification risk as part of a broader dataset-release decision.

Notably, several jurisdictions frame identifiability through qualitative standards, such as whether identification is "reasonably likely" (EU [13]) or whether an individual is "reasonably identifiable" (Australia [20]). These standards are

deliberately broad, leaving the means, resources and thresholds to the assessor's judgement, so two organisations may reach different conclusions about the same dataset. A quantitative approach is useful here: measuring risk under explicitly defined attacker capabilities makes the assumptions behind a "reasonably likely" judgement visible and comparable, without replacing the legal test.

In Aotearoa, the Privacy Act 2020 [2] establishes Information Privacy Principles governing the handling of personal information, while sector-specific requirements and data-governance approaches provide additional controls. Māori data governance further demonstrates why technical re-identification metrics cannot independently determine whether data should be released. Māori data sovereignty [21] emphasises concepts including Māori authority, relationships, collective benefit, guardianship, and protection from inappropriate use. Ngā Tikanga Paihere [22] similarly considers community interests, potential harms, data transformation and culturally appropriate use. These considerations extend beyond whether an individual can be re-identified to include collective interests, cultural context, and potential harms. Therefore, quantitative re-identification risks should be considered as one component of a broader governance decision rather than as a replacement for existing privacy and data-governance processes.

Aotearoa government guidance [23, 24] provides practical direction for the release and reuse of data. The Five Safes framework [25], used by Stats NZ, considers safe people, projects, settings, data, and outputs when assessing appropriate data access and disclosure risk. Existing guidance recognises that different de-identification approaches result in different levels of protection, acknowledges that information may become identifying when combined with auxiliary data, and incorporates privacy into the process of preparing datasets for open release. However, the Aotearoa guidance reviewed provides process and governance direction without quantitative re-identification risk measures. This leaves a gap for an empirically grounded approach that translates measured risk into release decision support, to complement existing governance materials and provide enhanced guidance.

D. Synthesis and Research Gap

The literature provides substantial capability across several components of privacy-preserving data release. Re-identification research demonstrates the importance of QIs and auxiliary knowledge, privacy models provide mechanisms for reducing disclosure risk, quantitative metrics provide ways to measure characteristics associated with re-identification, adversarial modelling introduces context about attacker capability, and technical tools automate privacy analysis and transformation. Alongside these technical approaches, governance frameworks provide legal, organisational, and cultural considerations for responsible data use.

However, these approaches do not necessarily provide an integrated, organisation-facing process for translating quantitative re-identification evidence into dataset-release decisions. The legislation and governance guidance examined does not provide a single, generally applicable quantitative method for assessing residual re-identification risk as part of a broader dataset-release decision. Instead, these approaches rely on contextual judgement, leaving a practical gap between measuring residual risk and communicating what that evidence means for intervention, utility, access restrictions, and organisational release decisions. This distinction informed the direction of this project. The resulting framework uses quantitative re-identification risk as one input to a structured release-decision process, while directing users to consider relevant legislation, sector requirements, governance considerations, and utility before reaching a release decision.

This project addresses this gap by developing a quantitative experimental pipeline and translating its results into an organisation-facing decision framework. The contribution is not a new privacy model or re-identification metric, but an investigation of how selected quantitative risk measures, attacker knowledge, and privacy interventions can be combined to produce project-specific risk classifications and synthesise intervention effects. These results are then translated into practical guidance that supports decision-making in different contexts while incorporating legal, organisational, and Māori data-governance considerations.

III. Design

This section outlines the project requirements and design decisions, experimental architecture, risk measurement approach, and final decision-support framework.

A. Requirements and Constraints

Table III provides requirements and constraints, identified to achieve project objectives and remain within scope.

TABLE III
DESIGN REQUIREMENTS AND CONSTRAINTS

ID	Requirement or constraint
R1	Classify a dataset's observed re-identification risk as LOW, MEDIUM, or HIGH through a lookup that a non-specialist can apply without running code. (O3, O5)
R2	Measure re-identification risk empirically under explicitly defined attacker capabilities rather than using a single theoretical privacy criterion. (O1)
R3	Evaluate privacy interventions applied to quasi-identifiers, including de-identification, generalisation, suppression, and noise addition. (O2)
R4	Provide evidence-based guidance on interventions that may reduce observed risk from different starting configurations. (O4, O5)
R5	Translate technical experimental results into practical organisational release guidance. (O5)
R6	Avoid requiring organisations to perform the underlying technical analysis when applying the framework. (O5, O6)
R7	Support reproducible and auditable experimentation. (O1)
R8	Account for the broader context in which dataset-release decisions are made. (O5, O6)
C1	No real personal data may be used for experimental development and evaluation. (O1)
C2	The project must remain feasible within the scope and resources of an Engineering Honours project. (O1-O6)
C3	The project must use reproducible, accessible software and data resources without dependence on proprietary infrastructure. (O1)
C4	Risk boundaries must be derived from experimental evidence rather than selected solely through subjective judgement. (O3)

B. Design Alternatives and Decisions

The requirements and constraints in Table III informed a series of design decisions (D1-D10) concerning the experimental data, risk measurement, classification, privacy interventions, evidence traceability, and final framework. Alternative approaches were considered, but rejected where they would materially affect the validity, interpretability, reproducibility, or scope of the project.

1) Experimental Data and Attacker Modelling

D1. Experimental data source (R7, C1, C3). Real personal data, public benchmark datasets, and synthetic data were considered. Real personal data was excluded due to the privacy and ethical constraint in C1. Public datasets such as those available through the UCI repository [26] would provide realistic data, but would have reduced control over the attributes, distributions, and transformation levels required for controlled experimentation. A synthetic population of 50,000 records was therefore selected. The dataset contains a broader set of attributes than those used as QIs so that the experimental population represents a structured dataset containing both identifying and non-identifying/contextual information. A fixed seed provides reproducibility, while changing the seed allows structurally comparable datasets to be generated for future experiments.

D2. Attacker Modelling (R2). A single attacker model and a continuous representation of attacker knowledge were considered alongside nested QI sets. Four nested attacker models were selected: age; age and gender; age, gender and postcode; and age, gender, postcode and occupation. This provides an ordered and interpretable representation of increasing auxiliary information while allowing the same risk measures to be applied consistently across attacker capabilities.

2) Risk Measurement and Classification

D3. Risk Measurement (R2, C2). Single-metric approaches, established privacy models such as k-anonymity, and multiple complementary re-identification indicators were considered. The project selected uniqueness, linkage, singling-out, and partial re-identification measures because they capture different characteristics associated with QI-based re-identification risk, including record distinguishability, matching against auxiliary information, isolation through queries, and membership in small equivalence groups. Using multiple indicators reduces reliance on a single representation of risk and provides a broader basis for comparing experimental configurations. l-diversity and t-closeness were also considered as additional privacy models but were excluded because they primarily address sensitive-attribute disclosure within equivalence classes and would broaden the project beyond its primary focus on QI-based re-identification risk.

D4. Composite risk score (R2, C2). Equal weighting, ML-based feature-importance weighting, maximum scoring, and dimensionality-reduction approaches were considered. ML-based weighting would have introduced a substantial modelling and validation problem of its own because no ground-truth risk labels were available. A maximum score would discard information from three metrics. The four metrics were therefore combined using an equal-weight mean to provide a transparent and reproducible composite score. Metric influence was subsequently examined separately through the risk-driver analysis rather than incorporated into the weighting model.

D5. Risk-band derivation (R1, C4). Manually selected thresholds and alternative clustering approaches such as Gaussian Mixture Model were considered. K-means was selected because it provides an empirical method for grouping the observed continuous risk scores without imposing arbitrary round-number boundaries. K = 2-5 were evaluated using silhouette scores. K = 3 was selected because the candidate clusterings had similar separation, and three groups provided a direct and interpretable LOW, MEDIUM, and HIGH classification system. The midpoints between adjacent cluster centres were used as the boundaries between the resulting risk bands.

3) Privacy intervention design

D6. Anonymisation experiment (R3, R4, C2). One-factor-at-a-time testing, a random subset of configurations, and a full factorial design were considered. A full factorial design was selected, covering four age levels, four postcode levels, and two gender states, producing 32 configurations. This provides complete coverage of the selected intervention space and allows combinations of transformations to be compared rather than examining each intervention independently.

D7. Differential privacy (R3, C2). Laplace and Gaussian mechanisms, alongside broader DP approaches, were considered. A Laplace mechanism was selected because it provided a feasible experimental implementation within the project scope and allowed the effect of different ε values to be examined. A comprehensive production-grade DP implementation was outside the scope of the project.

4) Framework and evidence architecture

D8. Evidence traceability (R5, R7). Framework claims could have been supported through conventional citations or an appendix of evidence tables. A Decision Audit Trail was instead designed to link each framework finding to its source evidence file, specific result, interpretation, and resulting organisational guidance. This provides traceability from the final framework back to the empirical results.

D9. Framework form (R1, R5, R6, R8, C2). At project proposal, a PDF lookup framework was selected as the final output because the organisation applying the framework should not need to run the experimental code or provide data to the technical pipeline. This reasoning eliminated producing a tool as the solitary output. The framework presents pre-derived quantitative evidence about re-identification while retaining organisational judgement around intended use, utility, access, legal requirements, organisational risk appetite, and Māori data governance.

Existing tools such as ARX [11] and sdcMicro [12] demonstrate the feasibility of integrating anonymisation, risk assessment, and utility analysis within a technical workflow.

This project instead separates technical evidence generation from the organisation-facing decision framework, reflecting the requirement that the final artefact provide evidence-informed guidance without requiring the organisation to perform the underlying technical analysis.

D10. Pipeline structure (R5, R7, C2, C3). A single script approach was considered, but staged scripts with CSV interfaces were selected to separate dataset generation, risk measurement, evidence derivation, and synthesis. This approach allows individual stages to be rerun and creates explicit intermediate evidence that can be inspected and traced through the repository.

C. System Architecture

The system has two main layers shown in Fig. 1: the Evidence Pipeline (R1-R4, R7) and the Final Framework (R5-R6, R8). The evidence pipeline contains all the quantitative re-identification analysis performed. This is traced to risk framework decisions in the Decision Audit Trail (R5, R7). The Final Framework layer integrates both quantitative and qualitative guidance to support informed release decision-making.

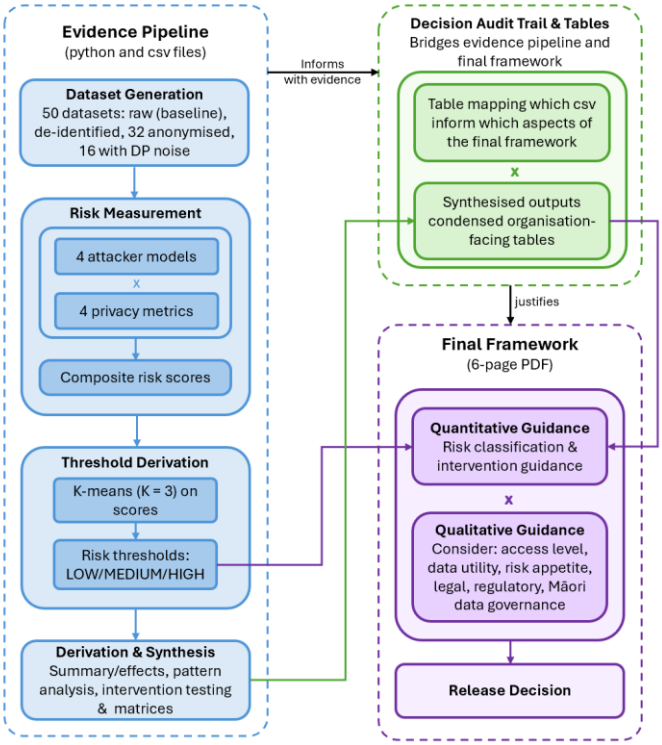

Fig. 1. System architecture diagram.

D. Risk Measurement Design

1) Attacker Models: This project defines four attacker models representing progressively increasing levels of auxiliary information available to an attacker. The models use nested sets of QIs, detailed in Table IV, allowing the effect of additional auxiliary knowledge on observed re-identification risk to be evaluated consistently across the experimental configurations.

TABLE IV
ATTACKER KNOWLEDGE REPRESENTED BY EACH EXPERIMENTAL ATTACKER MODEL

Attacker	Available QIs
Weak	Age
Moderate	Age + gender
Strong	Age + gender + postcode
Very strong	Age + gender + postcode + occupation

Attacker capability is represented by progressively increasing auxiliary information rather than by simulating increasingly sophisticated attack algorithms. This approach was selected because the project focuses on measuring how the information available to an attacker affects the distinguishability and linkability of records, rather than modelling the full behaviour or resources of a real-world attacker.

Using nested QI sets also provides a controlled basis for comparing the same dataset under different assumptions about attacker knowledge. The selected attributes are informed by established re-identification research, including Sweeney's work on the identifying potential of combinations of demographic and geographic attributes [4]. Age was used instead of date of birth to provide a simpler attribute that could be progressively generalised into broader bands while retaining the underlying principle that combinations of quasi-identifiers can contribute to re-identification risk.

Occupation was not investigated as a transformation because it is only included in the very strong attacker model. Analysing occupation suppression as an intervention would therefore introduce an intervention whose effects could only be evaluated for a single attacker model. Intervention analysis instead focuses on transformations with effects that can be compared consistently across the attacker models.

2) Risk Metrics: Four complementary metrics, detailed in Table V, were selected to capture different characteristics associated with QI-based re-identification risk.

TABLE V
SELECTED RE-IDENTIFICATION RISK METRICS

Metric	Purpose / interpretation
Uniqueness	Measures how distinguishable records are using the selected QIs. Calculated as the proportion of records with unique QI combinations.
Linkage	Measures the potential for matching records using external information. A 50% random sample is compared with the target dataset, with risk measured as the proportion of auxiliary records that can be uniquely matched to the target dataset using the selected QIs.
Singling-out	Measures whether individual records can be isolated using QI-based queries. Calculated as the proportion of tested QI-based queries that uniquely isolate a record.
Partial re-identification	Measures the proportion of records belonging to small QI equivalence groups, where fewer possible matches indicate greater distinguishability. Calculated as the proportion of records with $k \leq 5$.

These metrics were selected as they provide complementary indicators of record distinguishability, matching against auxiliary information, isolation through QI-based queries, and membership in small equivalence groups. Together, they provide a broader representation of observed re-identification

risk than relying on a single metric. The metrics are applied using the QIs specified by each attacker model, so increasing attacker knowledge can be evaluated using the same measurement approach.

l-diversity and t-closeness [8, 9] were considered as established extensions of k-anonymity but were not implemented for this project because they primarily address sensitive-attribute disclosure within equivalence classes, whereas this project focuses on re-identification risk associated with QIs, attacker knowledge, and record distinguishability.

3) Composite Risk Score: A consistent scalar measure was required to compare experimental configurations and subsequently derive the project's risk bands. The four metric values are combined using an equal-weight arithmetic mean, as defined in (1); rationale for this approach is discussed in D4.

$$R = \frac{U + L + S + P}{4} \quad (1)$$

where R is the composite risk score, and U, L, S, and P represent uniqueness, linkage, singling-out, and partial re-identification scores, respectively.

All four metrics are normalised to the same [0,1] scale, so the equal-weight mean produces a composite risk score on the same scale.

E. Privacy Intervention, Experimental Design, and Risk Classification

The final experimental design evaluates attribute-based anonymisation and DP as interventions affecting observed re-identification risk (D6, D7). Age is generalised through four levels of banding, postcode through four levels of geographic truncation, and gender is either retained or suppressed. Occupation is retained as auxiliary information for the very strong attacker rather than treated as an intervention, allowing attacker capability and privacy interventions to remain distinct (D2, D6).

The anonymisation experiment uses a full factorial design of four age levels × four postcode levels × two gender states, producing 32 configurations (D6). Each configuration is evaluated under the strong and very strong attacker models. DP is evaluated separately across four starting anonymisation configurations and four predefined ε values using the implemented Laplace mechanism (D7).

Risk classification is applied to the resulting composite risk scores. K-means clustering with K = 3 produces three empirical groups, which are ordered by their cluster centres and assigned the LOW, MEDIUM, and HIGH labels (D5). The midpoint between each pair of adjacent cluster centres defines the boundary between the corresponding risk bands. These thresholds are project-specific and are used consistently throughout the subsequent analysis and framework.

F. Framework Design

The organisation-facing framework (as shown in Appendix A) translates the quantitative experimental evidence into a structured dataset-release decision process. The first page outlines the overall decision flow (Fig. 2). The framework begins with the initial risk classification, then considers intended use, access context, organisational risk appetite, and potential interventions. Where intervention is required, the user is directed to reassess the resulting risk before proceeding to the final release assessment.

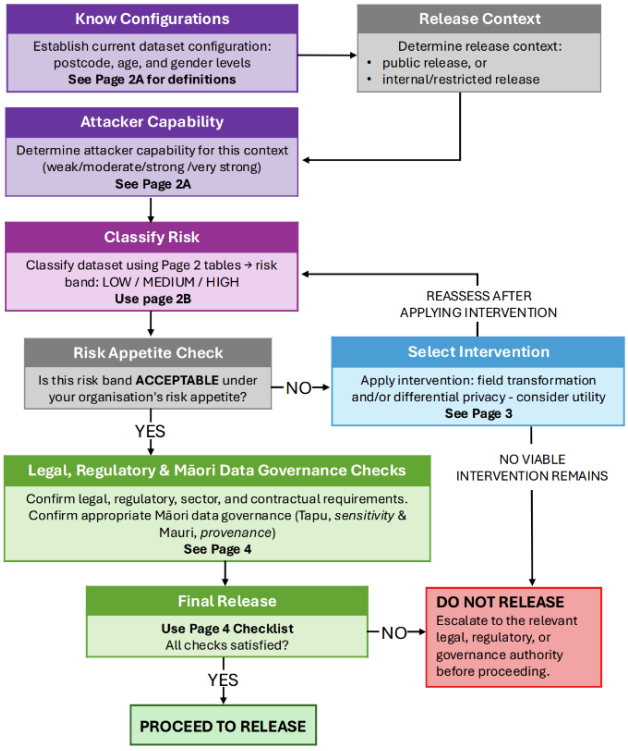

Fig. 2. Decision flow from page 1 of the final framework.

The final decision process incorporates data utility, access controls, applicable legal and sector requirements, and Māori data governance alongside the quantitative risk classification. The framework therefore presents the experimental results as decision-support evidence rather than as an automatic release/rejection mechanism. This approach was selected to avoid treating a numerical risk score as sufficient to determine whether a dataset should be released. It also allows the technical pipeline to remain focused on controlled quantitative analysis while the framework provides a practical interface for applying the resulting evidence in a broader organisational context.

G. Sustainability Considerations

Sustainability considerations were incorporated into the design across environmental, social, economic, and technical dimensions. The use of synthetic data and conventional computing resources limits the project's environmental impact and avoids the resource and privacy costs associated with using real personal data (D1). The use of Python and open-source libraries supports economic accessibility, while the staged and reproducible pipeline supports technical maintainability and future extension (D10).

Social sustainability is particularly relevant to this project because privacy and responsible data release can affect

individuals and communities. The framework therefore incorporates intended use, access context, organisational risk appetite, legal and sector requirements, and Māori data governance rather than treating quantitative risk as the sole basis for release decisions (D9).

Further discussion of these sustainability considerations, including alignment with the UN Sustainable Development Goals, is provided in Section I.E.

IV. Implementation

The artefact comprises a staged Python pipeline that generates experimental evidence which informs a standalone organisational framework, implementing the design in Section III. This section outlines the components, their rationale, and their interactions, showing how the design was realised and results made reproducible.

A. Implementation Environment and Architecture

Each of the five stages in the experimental pipeline is a folder of Python scripts in the GitLab repository, and each writes CSV files that the subsequent stage reads. Fig. 3 shows this implemented pipeline and its relationship to the environment architecture. The first four stages realise the Evidence Pipeline layer of the design; the fifth realises the Final Framework layer. Because stages communicate only through files, any stage can be re-run from its inputs, and every intermediate result can be inspected (D10, R7).

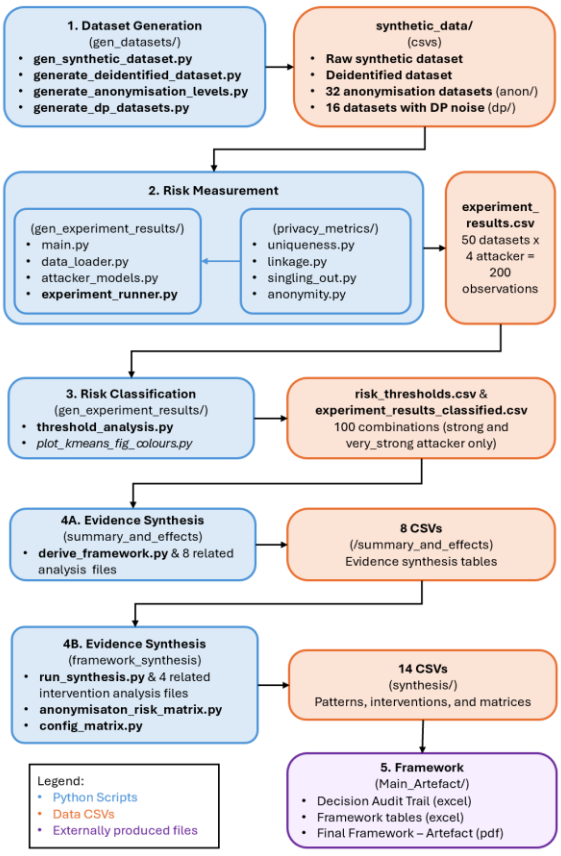

Fig. 3. Implementation of design: scripts and files.

The risk measurement stage is the only one in which several modules interact within a single run. Fig. 4 shows its structure. `main.py` calls `run_experiments`, which takes the attacker QI sets from `attacker_models.py` and passes each, with the dataset, to the four metric modules. The remaining stages of evidence synthesis are sequential. A README documents the run order, inputs and outputs of every script.

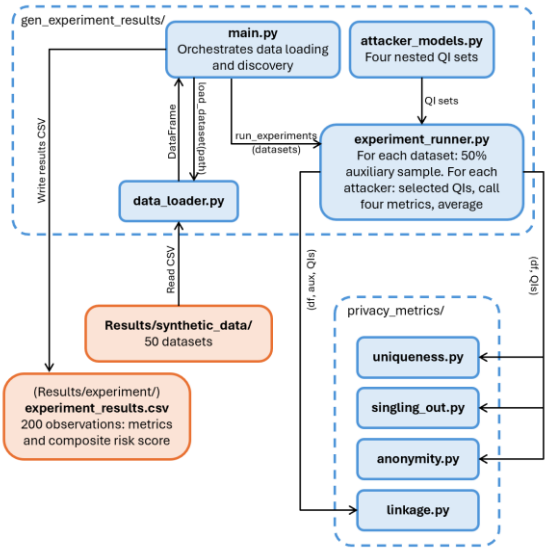

Fig. 4. Module interaction in risk management stage

B. Technology Choices

All selected components are free and open source (C3), run on a single standard computer (C2), and use open file formats, so the artefact can be reproduced and maintained without licenses or special infrastructure. Table VI details the components used and maps them to respective IDs.

TABLE VI
COMPONENTS MAPPED TO REQUIREMENTS, CONSTRAINTS AND DESIGN DECISIONS

Component	Realised as	Rationale
Language	Python 3.13	Mature scientific ecosystem with free libraries; readable code supports auditability (C3, R7).
Data processing	pandas	Grouping records by QI combination supports all four metrics; merging implements linkage (C2).
Random generation, noise	NumPy, seed 42	Seeded generation of the population and of the Laplace noise (D1, D7), giving repeatable results (R7).
Clustering	scikit-learn: K-Means, silhouette score	Established, tested implementation rather than bespoke code; 20 restarts with a fixed seed give repeatable cluster centres (D5, C4).
Figures	Matplotlib	Cluster and score-distribution plots supporting the risk bands.
Stage interfaces	CSV files	Human-readable, tool-neutral and diffable; each stage can be inspected and re-run independently (D10).
Framework	PDF, six pages	Applied without software or transferring data to the pipeline (D9, R6).
Version control	Git, GitLab	Complete history of the work; hosts source, results and framework (R7).

C. Risk Measurement Implementation

Attackers as QI sets (D2). Attacker capability is defined by the information available rather than by attack algorithm, so each attacker is implemented as a list of QI columns. Every dataset is evaluated against all four attackers, giving 200 observations. The weak and moderate attackers scored at most 0.00014 on the composite risk, so they do not inform the risk-band boundaries; the strong and very strong attacker (giving 100 observations) are used for classification and all proceeding analysis.

Metrics and composite score (D3, D4). Each metric is a separate module with the same interface (dataset and QI columns, plus an auxiliary sample for linkage), so attacker strength varies only through the QI set. The auxiliary sample is a seeded 50% random sample of the dataset, an experimental device for controlled background knowledge rather than an assumption about real attackers. The four values are averaged with equal weight and written, with the dataset configuration and attacker, to `experiment_results.csv`.

One code path for every privacy configuration (D6, D7). Transformations were implemented so that every dataset keeps the same columns and the same measurement code runs unchanged. Age and postcode are generalised by banding and digit truncation, and gender suppression replaces values with "Unknown" rather than removing the column. `itertools.product` generates the 32 combinations, with the configuration encoded in each file name so later stages can decode it.

For DP, Laplace noise needs a numeric value, so banded ages are first converted to a random whole age within the band. Noise is then added to each record's age, with a spread equal to a sensitivity (Δ) value divided by ε, and ages are clipped to 0–100. The sensitivity, meaning the most one person's age is assumed able to change the released value, was set to 10 years as a modelling parameter. The sensitivity value of 10 was selected to provide a consistent and interpretable basis for comparing the effect of different ε values, while remaining large enough to produce observable differences in the experimental data. Because smaller ε produces more noise, and the sensitivity was chosen for the experiment rather than derived from the data, ε indexes how much noise was added and does not give a formal privacy guarantee for a released dataset.

Risk classification (D5). `threshold_analysis.py` fits K-means to the 100 strong and very strong risk scores for K = 2 to 5 and compares the fits using the silhouette score, a measure of how well separated the clusters are. The scores were similar (ranging from 0.766 to 0.793), with K = 5 marginally highest (0.793) and K = 3 at 0.771. Because the candidate solutions had comparable separation and the project's required output was interpretable organisational risk bands, K = 3 was selected to prioritise interpretability and alignment with the framework requirements over the marginally higher silhouette score of K = 5. Each cluster's centre is its mean score (0.025, 0.491 and 0.899), and the centres are labelled LOW, MEDIUM and HIGH from lowest to highest. The boundary between adjacent bands is set halfway between their centres, giving 0.2577 (LOW/MEDIUM) and 0.6950 (MEDIUM/HIGH), which are written to `risk_thresholds.csv`. Applied to the 100 observations (50 datasets × 2 attackers), these boundaries classify 44 as LOW, 18 as MEDIUM and 38 as HIGH.

Evidence synthesis (D4, D8). The synthesis stages convert observations into the evidence the framework cites: attacker summaries, transformation effects, risk drivers and framework candidates. `intervention_analysis.py` tests the 128 one-step transformations available from each configuration (age banding, postcode truncation, gender suppression) and records separately whether the score fell and whether the risk band changed, because a score can fall without the band changing.

D. Framework Implementation

The framework is a six-page PDF in `Main_Artefact/` that organisations can apply without running any code (D9, R6). Page 1 (Fig. 2) gives the decision flow: establish the dataset's configuration, release context and attacker capability, classify its risk band, check it against organisational risk appetite, intervene and reassess if needed, then complete the legal, regulatory and Māori data governance checks before release. Pages 2A–2B provide the dataset configuration and classification lookup tables for anonymisation and differential privacy, page 3 requires utility consideration and provides intervention guidance with the quantitative evidence behind each option, and page 4 provides legal and regulatory requirements, and Māori data governance guidance, as well as the final release checklist (R8).

The classification values and intervention guidance were derived from the evidence in `Results/`. The condensed anonymisation and DP classification tables were checked cell by cell against these evidence tables. A 24-row Decision Audit Trail links each framework finding to its main supporting evidence files (22 rows) or design rationale (2 rows), providing traceability for D8.

E. Completeness, Reproducibility and Traceability

Design decisions can be traced to components in the GitLab repository, as shown in Table VII. To verify that the repository reproduces the artefact, all generated results were deleted, and the eleven scripts were run in the documented order. All completed without error, regenerated all 78 result files, and every regenerated CSV matched the committed version to numerical precision. Constraints C1 to C4 hold by construction: only synthetic data is used (C1), the pipeline runs in minutes on a standard computer (C2), all components are open source (C3), and the risk boundaries are computed by `threshold_analysis.py` (C4). Two limitations of the build bound how the results should be read: each DP setting uses a single seeded noise draw, and all evidence comes from one synthetic population. Neither changes the structure of the design.

TABLE VII
MAPS DESIGN DECISIONS TO EVIDENCE IN THE REPOSITORY

Decision	Realised by	Evidence in repository
D1 (R7, C1, C3)	Seeded synthetic population of 50,000 records	`generate_synthetic_dataset.py; raw.csv`
D2 (R2)	Four nested attacker QI sets	`attacker_models.py`
D3 (R2, C2)	Four metric modules	`privacy_metrics/`
D4 (R2, C2)	Equal-weight mean; risk-driver analysis	`experiment_runner.py; risk_drivers.csv`
D5 (R1, C4)	K-means (K = 3) with midpoint boundaries	`threshold_analysis.py; risk_thresholds.csv`
D6 (R3, R4, C2)	Full factorial: 32 anonymised datasets	`generate_anonymisation_levels.py; anon/`
D7 (R3, C2)	Laplace noise: four ε values on four starting configurations (16 datasets)	`generate_dp_datasets.py; dp/`
D8 (R5, R7)	Decision Audit Trail: 24 rows, 12 evidence files	Audit trail workbook and CSV
D9 (R1, R5, R6, R8, C2)	Six-page PDF lookup framework	`Final Framework - Artefact.pdf`
D10 (R5, R7, C2, C3)	Five stages exchanging CSV files	Fig. 3; `README` run order

V. EVALUATION

The evaluation examines two main aspects of the project:

Quantitative risk assessment: Whether the developed pipeline produced useful quantitative evidence for assessing re-identification risk. This was evaluated through empirical analysis of the experimental results, including baseline risk, risk classification, and the effects of anonymisation and DP. Key evaluation measures include the composite risk score, used to measure observed re-identification risk; risk-band classification, used to assess whether configurations move between LOW, MEDIUM, and HIGH risk; numerical risk-reduction rates, used to compare intervention effects; and risk-band reduction rates, used to assess practical changes in classification. These measures assess observed risk and the consistency of the resulting classifications within the experimental dataset; they do not establish real-world re-identification probability or universal privacy thresholds. The silhouette score is used to assess the separation of candidate K-means clusterings, while reproducibility is assessed by comparing regenerated outputs against committed results.

Practical decision guidance: Whether the quantitative evidence could be translated into practical guidance for dataset-release decisions. This was evaluated through expert assessment of the final framework, focusing on its clarity, technical sensibility, assumptions, and practical usefulness.

A. Empirical Evaluation – Quantitative Risk Assessment

1) Baseline and Thresholds: The baseline experiments demonstrated that removal of direct identifiers alone did not reduce observed re-identification risk in the tested datasets. The raw dataset produced risk scores of 0.971 for the strong attacker and 0.993 for the very strong attacker, resulting in both being classified as HIGH risk, as seen in Table VIII. The de-identified configuration produced the same risk classification. As all four metrics use quasi-identifiers only, this follows by construction, and confirms that risk is driven by QI combinations rather than direct identifiers alone.

TABLE VIII
BASELINE OBSERVED RE-IDENTIFICATION RISK BEFORE AND AFTER REMOVAL OF DIRECT IDENTIFIERS

Dataset configuration	Strong attacker	Very strong attacker
Raw	0.971 (HIGH)	0.993 (HIGH)
De-identified	0.971 (HIGH)	0.993 (HIGH)

The risk scores across the 100 strong and very strong attacker observations were used to derive project-specific LOW, MEDIUM, and HIGH risk bands using K-means clustering. Fig. 5 illustrates the resulting clustering and the three project-specific risk groups. Multiple cluster counts were evaluated, with three clusters selected to provide the intended and interpretable three risk categories. The resulting boundaries were 0.2577 between LOW and MEDIUM and 0.6950 between MEDIUM and HIGH. These thresholds provide a consistent empirical basis for classifying the experimental results and were subsequently incorporated into the decision framework. They are project-specific and have not been externally validated as universal privacy-risk thresholds.

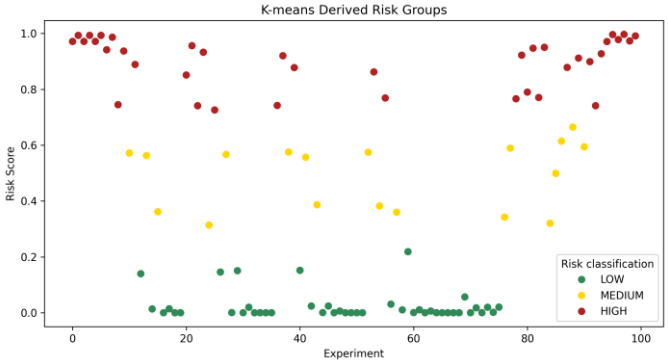

Fig. 5. K-means clustering of experimental risk scores into project-specific LOW, MEDIUM, and HIGH risk bands

2) Anonymisation: The anonymisation experiments demonstrated measurable effects from the tested interventions. Fig. 6 summarises the proportion of intervention tests producing numerical risk reductions and changes in risk classification. Postcode truncation produced the largest and most consistent reduction in observed risk, with numerical reductions (positive values of (2)) in 87.5% of strong-attacker and 100% of very-strong-attacker intervention tests. Risk-band improvements occurred at rates of 41.7% and 58.3%, respectively. Age banding and gender suppression also frequently reduced numerical risk but produced noticeably fewer changes in risk classification.

This project calculates risk reduction as the following:

$$Risk\ reduction = \frac{R_{before} - R_{after}}{R_{before}} \times 100\ \% \qquad (2)$$

where R_{before} is the risk before the intervention is applied, and R_{after} is the residual risk after intervention application.

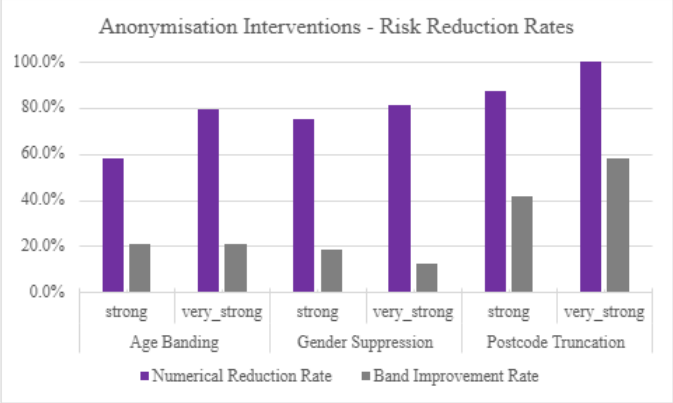

Fig. 6. Percentage of anonymisation interventions producing numerical and band reductions in observed risk.

Empirical evaluation reiterates that numerical reduction in risk did not necessarily produce a lower risk band. This demonstrates that the framework's intervention guidance cannot rely solely on whether an intervention reduces the numerical score; the resulting risk classification, starting configuration, and organisational risk appetite must also be considered

3) Differential Privacy: The DP experiments demonstrated that observed risk is dependent on both the ε value and the starting anonymisation configuration. Fig. 7 illustrates the resulting variation in observed risk across the tested starting configurations and ε values. DP did not consistently produce a LOW risk classification, and the relationship between ε and observed risk was not universally monotonic in the tested configurations. For example, the maximum risk for the LOW/MEDIUM configuration increased from 0.499 at ε = 0.1 to 0.911 at ε = 1, before decreasing to 0.899 at ε = 2. This reinforces the framework's treatment of DP as an intervention requiring reassessment rather than as an automatic means of achieving a particular risk level. This is particularly important to note, as DP is considered a strong mathematical guarantee in literature [10].

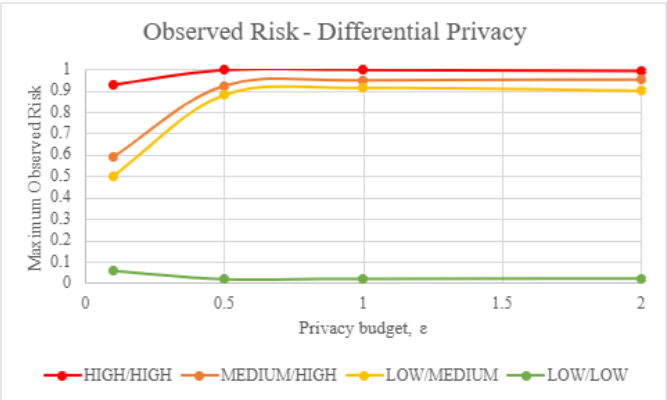

Fig. 7. Observed re-identification risk across different starting risk configurations under DP.

4) Empirical Finding Summary: Overall, the empirical evaluation demonstrated that the pipeline could distinguish changes in observed risk across anonymisation, DP, and attacker configurations, providing the evidence required to construct and populate the decision framework. The results also identified context-dependent intervention effects rather than a single transformation that consistently produced the same risk outcome.

B. Expert Evaluation of the Decision Framework – Practical Decision Guidance

The resulting framework was evaluated through semi-structured interviews with three experts representing privacy practice, public-sector data ethics, and cybersecurity academia. Feedback was analysed thematically in Table IX, rather than by individual interview question.

TABLE IX
EXPERT EVALUATION OF THE ORGANISATION-FACING DECISION FRAMEWORK

Evaluation Theme	Key Findings
Clarity and decision flow	All three experts considered the framework logical and understandable, particularly the sequential decision flow. The decision flow diagram structure was repeatedly identified as a strength, lowering barrier for entry for users with limited privacy expertise.
Practical usefulness	The experts considered the framework potentially useful as a structured starting point for organisations, particularly those with less mature privacy or data-release processes. As intended, framework was viewed as complementary to existing specialist processes rather than a replacement for privacy, legal, or governance expertise. Experts particularly valued the traceability of the framework, noting the reasoning behind a decision is retained, rather than requiring organisations to reconstruct how a decision was reached. One expert noted that completing the framework's initial assessment could be beneficial before engaging specialist teams, as key information and considerations are established before seeking legal or other specialist input.
Context and scope	A consistent theme was that re-identification risk cannot be assessed exhaustively from the three attributes represented in the project. Experts emphasised the importance of considering additional identifiers, data provenance and purpose, and the possibility of group-level harms. This reinforces the framework's explicit positioning as a decision-support tool rather than a standalone release decision mechanism.
Improvement and deployment	Expert evaluators' recommendations fell into three areas. *Governance scope:* Māori data governance principles should be presented as relevant beyond datasets containing Māori data, since small or distinctive communities, such as rainbow communities, can face collective harms that an individual-level risk band does not capture, and this should be made explicit to organisations with examples. *Usability:* provide a worked example or walkthrough, clarify that "attacker" includes non-malicious data users, and make organisational accountability and escalation responsibilities clearer. *Accessibility and deployment:* one evaluator suggested a short how-to video to suit different learning styles and noted compatibility with screen-readers for online versions of the framework as important. Real-world piloting and user testing were consistently identified as the most important next step before broader deployment.

The expert evaluation provides initial support for the framework's clarity and practical relevance while identifying clear boundaries around its current applicability. In particular, the feedback supports the framework as a strong proof-of-concept decision-support artefact but does not constitute validation for unrestricted organisational deployment.

C. Evaluation Against Project Objectives

All project objectives were achieved within the defined scope, as summarised in Table X.

TABLE X
EVALUATION OF PROJECT AGAINST STATED OBJECTIVES

Objective	Evaluation Evidence	Assessment
O1: Develop a pipeline for measuring re-identification risk	Pipeline produced quantitative risk results across the defined datasets and attacker configurations	Achieved within scope
O2: Evaluate anonymisation and DP configurations	Empirical experiments demonstrated changes in observed risk across the tested interventions and DP configurations	Achieved within scope
O3: Derive risk bands from experimental results	K-means clustering produced project-specific thresholds for LOW, MEDIUM, and HIGH risk bands	Achieved within scope
O4: Identify risk patterns and intervention effects	Analysis identified attacker-dependent risk and differing effects of postcode, age, gender, and DP interventions	Achieved within tested configurations
O5: Develop an organisation-facing decision framework	Empirical findings were translated into risk interpretation, intervention, reassessment, and governance guidance	Achieved as proof-of-concept
O6: Evaluate framework clarity and usefulness with experts	Three experts assessed the framework and identified strengths, limitations, and improvements	Achieved

Each requirement and constraint in Table III was met. All 50 datasets were evaluated under four nested attacker models (R2, R3) and 128 one-step interventions were tested (R4). The framework's tables were checked cell by cell against the evidence and every finding is traceable through the Decision Audit Trail (R1, R5, R8). Experts judged the decision flow accessible to users with limited privacy expertise, though non-specialist testing remains future work (R6). All 78 result files were regenerated from the scripts and matched (R7); C1–C4 hold by construction.

D. Limitations and Future Improvement

The evaluation demonstrates performance within the project's defined scope but does not establish general validity across real-world datasets or organisations. The experiments use synthetic datasets and a restricted set of attributes, attacker models, privacy techniques, and risk measures. The risk bands are derived from these experimental results and have not been externally validated. The DP implementation is also an experimental implementation rather than a comprehensive production solution.

Each DP setting used a single seeded noise draw, so the non-monotonic relationship between ε and risk may partly reflect sampling variability. Noise was applied to age only, with a sensitivity (10 years) chosen for the experiment, so ε indexes noise rather than giving a formal guarantee. All evidence comes from one synthetic population of 50,000 records. Because uniqueness depends on population size and attribute cardinality, the lookup tables should not be applied to materially different datasets. The linkage auxiliary sample is a 50% subset of the target, which guarantees overlap and may overstate linkage against real external sources.

Expert evaluation was based on three participants, therefore providing initial qualitative evidence rather than broad organisational validation. The strongest recommendation for further evaluation was real-world piloting and user testing, which could assess whether organisations can apply the framework consistently to actual release decisions. Before potential roll-out of this framework, future work must evaluate additional datasets, attributes, attacker models, privacy measures, organisational contexts, and group-privacy considerations.

VI. CONCLUSION

This project investigated whether quantitative re-identification risk measures could be effectively translated into practical, context-aware guidance for organisations making data release decisions. A Python-based experimental pipeline was developed to measure selected indicators of re-identification risk across defined attacker models, anonymisation interventions, and DP configurations. The experimental results demonstrated that the removal of direct identifiers alone is not sufficient to reduce observed risk. The effects of privacy-preserving interventions varied according to the selected intervention, starting configuration, and attacker capability. Project-specific LOW, MEDIUM, and HIGH risk bands were derived from the experimental results and used to provide a consistent basis for interpreting these findings.

The resulting evidence was translated into an organisation-facing decision framework that steps organisations through the decision-making process, combining quantitative risk classification with considerations including intended use, data utility, access context, organisational risk appetite, legal and sector requirements, and Māori data governance. Expert evaluation provided positive feedback on the framework's clarity, practicality, and potential usefulness as a structured starting point for organisational release decisions. The evaluation also highlighted the importance of further user testing and broader applicability beyond the project's defined scope to progress from a proof-of-concept to an organisation-ready decision-support tool.

Overall, the project demonstrates the potential for quantitative re-identification risk measures to provide useful evidence within a broader privacy, cybersecurity, and data governance decision process. The resulting pipeline and framework provide a proof-of-concept foundation for organisations and future research to further investigate how measurable privacy-risk evidence can support responsible and context-aware decisions about dataset release.

VII. FUTURE WORK

This project provides a foundation for further investigation of quantitative re-identification risk and its application to organisational dataset-release decisions. A natural next step involves expanding the range of quasi-identifiers, auxiliary

information, attacker models, privacy-preserving techniques, and risk metrics considered. It is important to then evaluate the pipeline using real-world or more diverse datasets to determine whether the observed relationships between attacker capability, privacy interventions, and re-identification risk generalise beyond the synthetic datasets used in this project.

Further work could extend the quantitative analysis to examine the privacy-utility trade-off more directly, allowing the effects of different interventions on both re-identification risk and the usefulness of released data to be evaluated together. The DP component could also be extended using more comprehensive implementations and additional privacy mechanisms to investigate how formal privacy guarantees relate to observed re-identification risk.

The organisation-facing framework provides a basis for a subsequent user study or organisational pilot. Testing the framework with a larger and more diverse group of practitioners could assess whether users interpret the risk classifications and decision flow consistently, and whether the framework supports real release decisions in different organisational contexts. Future evaluation could also investigate group and community-level privacy considerations from a quantitative perspective. Current technical implementation focuses on individual privacy. Empirical research focused on group privacy could further support how Māori data governance principles can be incorporated into broader organisational data-release practices.

The next framework iteration should incorporate expert recommendations for a worked example, clarification that attackers include non-malicious data users, clearer accountability and escalation, and more accessible formats.

APPENDIX

The complete framework and supplementary documentation are provided in the submitted appendix.

ACKNOWLEDGEMENT

Thank you to my supervisors Dr. Qurrat Ul Ain and Dr. Andrew Lensen. Thank you to my privacy, data, and cybersecurity experts Katrine Evans, Emma MacDonald, and Dr. Lisa Patterson for helping with evaluation.

REFERENCES

[1] S. Garfinkel, J. Near, A. Dajani, P. Singer, and B. Guttman, "De-Identifying Government Datasets: Techniques and Governance," NIST Special Publication 800-188, Sep. 2023, doi: 10.6028/NIST.SP.800-188.

[2] New Zealand Government, "Privacy Act 2020", New Zealand Legislation, 2020. Available: https://www.legislation.govt.nz/act/public/2020/31/en/latest/#LMS23223

[3] United Nations, "The 17 sustainable development goals," *United Nations*, 2015. Available: https://sdgs.un.org/goals.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy" International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, Oct. 2002, doi: 10.1142/s0218488502001648.

[5] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. 2008 IEEE Symposium on Security and Privacy*, 2008, pp. 111–125, doi: 10.1109/SP.2008.33.

[6] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A Systematic Review of Re-Identification Attacks on Health Data," PLoS ONE, vol. 6, no. 12, p. e28071, Dec. 2011, doi: 10.1371/journal.pone.0028071.

[7] W. Xia, Y. Liu, Z. Wan, Y. Vorobeychik, M. Kantacioglu, S. Nyemba, E. W. Clayton, and B. A. Malin, "Enabling realistic health data re-identification risk assessment through adversarial modeling," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 4, pp. 744–752, 2021, doi: 10.1093/jamia/ocaa327.

[8] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L -diversity: Privacy beyond k-anonymity" *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[9] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106-115, doi: 10.1109/icde.2007.367856.

[10] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends® in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014, doi: 10.1561/0400000042.

[11] "ARX: Data Anonymization Tool" [Online]. Available: https://arx.deidentifier.org/.

[12] T. Benschop and M. Welch, "Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro". [Online]. Available: https://sdcpractice.readthedocs.io/en/latest/

[13] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," *Official Journal of the European Union*, vol. L 119, pp. 1–88, May 4, 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[14] UK Parliament, "Data Protection Act 2018," legislation.gov.uk, 2018. https://www.legislation.gov.uk/ukpga/2018/12/contents

[15] Office of the Privacy Commissioner of Canada, "The Personal Information Protection and Electronic Documents Act (PIPEDA)" Dec. 08, 2021. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

[16] Privacy Act, R.S.C. 1985, c. P-21, "Consolidated federal laws of Canada, Privacy Act," https://laws-lois.justice.gc.ca/eng/acts/P-21/page-1.html#h-397260

[17] "Quebec Adopts the Regulation on Anonymization of Personal Information," Mccarthy.ca, 2024. https://www.mccarthy.ca/en/insights/blogs/techlex/quebec-adopts-regulation-anonymization-personal-information

[18] U.S. Department of Health and Human Services "Methods for De-identification of PHI," *HHS.gov*, https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/

[19] Office of the Australian Information Commissioner, "The Privacy Act," OAIC, 2025. https://www.oaic.gov.au/privacy/privacy-legislation/the-privacy-act

[20] Office of the Australian Information Commissioner, "De-identification and the Privacy Act," [Online]. Available: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-and-the-privacy-act

[21] Te Mana Raraunga, "Principles of Māori Data Sovereignty'" 2018. Available: https://www.temanararaunga.maori.nz/principles-of-maori-data-sovereignty

[22] Stats NZ, "Ngā Tikanga Paihere: A Framework Guiding Ethical and Culturally Appropriate Data Use" 2020 Available: https://data.govt.nz/assets/data-ethics/Nga-Tikanga/Nga-Tikanga-Paihere-Guidelines-December-2020.pdf

[23] New Zealand Digital government, "Making personal information safe for reuse," Mar. 06, 2024. Available: https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

[24] Data.govt.nz, "Releasing open data on data.govt.nz" Available: https://data.govt.nz/catalogue-guide/step-by-step-guides/releasing-open-data-on-data-govt-nz

[25] Stats NZ, "How We Keep Integrated Data Safe | Stats NZ," Govt.nz. Available: https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/

[26] "UCI Machine Learning Repository," *archive.ics.uci.edu*. https://archive.ics.uci.edu/