RAPID: Risk of Attribute Prediction-Induced Disclosure in Synthetic Microdata

Matthias Templ
School of Business
Applied University of Science and Arts
Northwestern Switzerland
matthias.templ@fhnw.ch

Oscar Thees
School of Business
Applied University of Science and Arts
Northwestern Switzerland
oscar.thees@fhnw.ch

Roman Müller
School of Business
Applied University of Science and Arts
Northwestern Switzerland
roman.mueller@fhnw.ch

January 20, 2026

Abstract

Statistical data anonymization increasingly relies on fully synthetic microdata, for which classical identity disclosure measures are less informative than an adversary's ability to infer sensitive attributes from released data. We introduce RAPID (Risk of Attribute Prediction–Induced Disclosure), a disclosure risk measure that directly quantifies inferential vulnerability under a realistic attack model. An adversary trains a predictive model solely on the released synthetic data and applies it to real individuals' quasi-identifiers. For continuous sensitive attributes, RAPID reports the proportion of records whose predicted values fall within a specified relative error tolerance. For categorical attributes, we propose a baseline-normalized confidence score that measures how much more confident the attacker is about the true class than would be expected from class prevalence alone, and we summarize risk as the fraction of records exceeding a policy-defined threshold. This construction yields an interpretable, bounded risk metric that is robust to class imbalance, independent of any specific synthesizer, and applicable with arbitrary learning algorithms. We illustrate threshold calibration, uncertainty quantification, and comparative evaluation of synthetic data generators using simulations and real data. Our results show that RAPID provides a practical, attacker-realistic upper bound on attribute-inference disclosure risk that complements existing utility diagnostics and disclosure control frameworks.

1 Introduction

Open research data (ORD) are increasingly recognized as essential for scientific transparency and reproducibility [Nosek et al., 2015]. Yet the proportion of shared datasets remains low: for instance, only 23% of projects funded by the Swiss National Science Foundation currently provide ORD [Swiss National Science Fund (SNSF), 2024]. Legal constraints and contractual usage restrictions often hinder the release of microdata, while technical barriers to anonymization remain substantial for many research teams.

Fully synthetic microdata – datasets in which all records are simulated rather than perturbed copies of real individuals – offer a promising solution to this problem. However, synthetic data also raise an immediate question for data stewards: how should disclosure risk be quantified when the primary threat is no longer reidentification, but an adversary's ability to infer sensitive attributes from the released data?

1.1 The user perspective: data utility

From the user's perspective, synthetic data must satisfy two core requirements: statistical similarity to the original data and structural plausibility. Beyond reproducing marginal distributions and associations, synthetic data must respect domain-specific constraints– such as realistic household compositions, non-negative expenditures (e.g., on medication), internally consistent demographic characteristics (e.g., no underage individuals with adult children), and valid event sequences (e.g., a PhD obtained after a master's degree). Violations of such constraints can severely limit the credibility and usability of synthetic datasets, even when distributional similarity is high.

Figure 1 illustrates a general workflow for generating synthetic data and assessing its fitness for use. Analysts typically aim to (i) answer predefined research questions by fitting statistical or machine learning models and (ii) explore the data to identify new patterns and relationships. These objectives are attainable only when the synthetic data approximate the original data not merely in distribution, but also in structure.

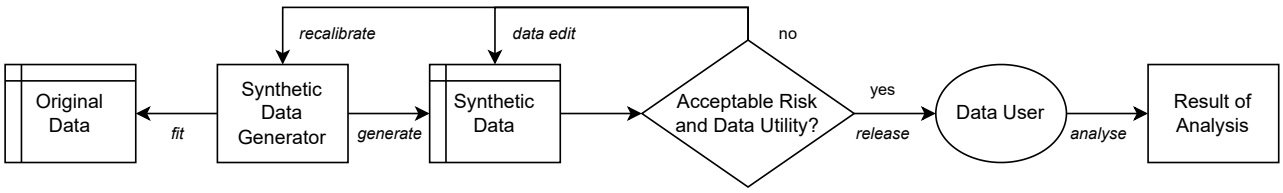

Figure 1: Synthetic data generation and evaluation workflow. A generator is trained on original data to produce synthetic records; analysts then assess whether the synthetic data preserve sufficient statistical properties for their intended analyses.

However, high analytical utility alone does not guarantee safe data release. From the data provider's perspective, increasing utility often entails preserving strong dependencies among variables—precisely the information that may enable an adversary to infer sensitive attributes. Disclosure risk therefore does not necessarily decrease as synthetic data become more useful; in fact, it may increase. In practice, this tension is commonly conceptualized using Risk–Utility (RU) maps [Duncan et al., 2001] and more recent multivariate extensions [Thees et al., 2025], which support informed decisions about acceptable trade-offs between analytical value and disclosure protection. Assessing inference disclosure risk in a way that is commensurate with modern, high-utility synthetic data remains a key methodological challenge.

1.2 Overview of Synthetic Data Generation Methods

Research on synthetic data has progressed considerably since the seminal contribution of Rubin [1993], who introduced the idea of replacing sensitive values with simulated draws from predictive distributions. Synthetic data are now widely used for facilitating open data sharing and long-term archiving.

Today, most synthetic data are generated using machine learning (ML) or artificial intelligence (AI) methods. These fall broadly into two categories:

- Conditional modeling approaches, where each variable is synthesized conditionally on others, often in a sequential fashion. Techniques include decision trees, random forests, gradient boosting (e.g., XGBoost), multiple regression, and more recently, large language models (LLMs). Synthesis proceeds variable-by-variable: each variable is generated conditional on those already synthesized. A few approaches [Templ et al., 2017] allow for accommodating missing data, complex survey designs, clustering, and hierarchical structures. Parameter fitting is performed on the original, non-anonymized data, and synthetic values are drawn from estimated conditional distributions.
- Joint modeling approaches, which attempt to model the entire joint distribution of all variables simultaneously. This class includes generative adversarial networks (GANs) and other deep generative models, including conditional GANs where full records are generated conditional on some known context rather than variable-by-variable. Joint modeling methods generally require large training datasets– especially for synthetic data generation [Mekonnen, 2024]– careful tuning of hyperparameters [Miletic and Sariyar, 2024], and significant computational resources. Moreover, they may struggle with outliers [Stadler et al., 2020], weakly correlated data structures (which are common in practice) [Ward et al., 2025], and learning intricate relationships among variables [Thees et al., 2024].

Comparative evaluations suggest that only a small number of methods currently achieve high utility for synthetic data, particularly in complex tabular settings [Thees et al., 2024].

While this paper focuses on synthetic data as the primary application, the inferential disclosure risk framework we propose applies equally to traditionally anonymized data (e.g., data protected via generalization, suppression, or noise addition) and even to non-anonymized data. In all cases, the core question is the same: given the released (or candidate) dataset, how accurately can an adversary infer sensitive attributes from quasi-identifiers? From a modeling perspective, there is no fundamental difference – the same predictive approach quantifies risk regardless of how the data were produced.

1.3 Disclosure risk measures

Classical SDC distinguishes (i) identity disclosure (i.e., linking a record to a specific individual); (ii) attribute disclosure (i.e., correctly learning sensitive values); and (iii) membership disclosure (i.e., learning whether an individual's data is part of a dataset) [Hundepool et al., 2012a].

For fully synthetic microdata, identity disclosure risk is typically low [Templ, 2014, El Emam et al., 2020], as synthetic records are not literal representations of individuals. As a result, most attention shifts toward analytical validity and attribute disclosure risk.

Existing attribute-disclosure diagnostics fall into two broad categories:

- Match-based measures, which evaluate how often synthetic values exactly match original values [Taub et al., 2018]. Examples include DiSCO [Raab et al., 2025], which flags records as at-risk when a quasi-identifier combination in the original data appears in the synthetic data with the same sensitive value. Such approaches are transparent and integrate naturally with classical SDC concepts, but they may be sensitive to class imbalance and do not directly model the inference process.
- Model-based measures, which train a predictive model on synthetic data and assess its performance on the original data [Hittmeir et al., 2020]. Hittmeir et al. [2020] proposed comparing prediction accuracy against a baseline that represents the attacker's prior knowledge (e.g., marginal class frequencies), thereby quantifying the additional risk introduced by the synthetic release. This baseline-comparison principle – measuring how much better an attacker can do with the released data than without it – is central to the approach we develop in this paper.

For model-based measures, the threat model assumes an attacker who trains a predictive model on the released synthetic data and applies it to individuals whose quasi-identifiers are known. Figure 3 illustrates this scenario. If a model trained solely on synthetic data can reliably predict confidential attributes in the real data, the synthetic release carries privacy risk [Taub et al., 2018, Barrientos et al., 2018, Kwatra and Torra, 2024].

Stadler et al. [2020] operationalized this threat by comparing predictions from models trained on synthetic versus original data, measuring relative risk as the ratio of attack success rates. Bayesian approaches instead formalize the intruder's uncertainty directly: an attacker updates a posterior distribution over the unknown sensitive value given the released data and knowledge of the data generation mechanism [Reiter et al., 2014, Latner et al., 2025].

The risk measure proposed here focuses on inferential disclosure risk– also referred to as predictive disclosure [Willenborg and de Waal, 2001]– which occurs when the publication of microdata enables more accurate or more confident inferences about sensitive attributes than would have been possible without the release [Duncan and Lambert, 1989, Hundepool et al., 2012b]. This concept builds on Dalenius's foundational definition:

> "If the release of the statistic S makes it possible to determine the value [of a sensitive attribute] more accurately than is possible without access to S, a disclosure has taken place." [Dalenius, 1977, p. 432]

Importantly, inferential disclosure can affect individuals who were not part of the original dataset and may concern information of which the affected individual is not even aware. For example, if the released data reveal a strong association between lifestyle indicators and disease risk, an attacker could infer elevated risk for a similar individual outside the dataset.

Although eliminating inferential disclosure entirely would require destroying all meaningful relationships between sensitive and non-sensitive variables [Dwork and Naor, 2010], we argue that inferential risk warrants increased attention in the era of big data and artificial intelligence. Mühlhoff [2021] notes that predictive privacy is violated when sensitive information is statistically estimated against an individual's will, provided that these predictions lead to differential treatment affecting their wellbeing or freedom.

While existing model-based measures typically provide aggregate, dataset-level risk summaries, we argue that record-level assessment is essential for targeted risk mitigation. Identifying which specific individuals face elevated disclosure risk enables data custodians to apply selective protections rather than uniform measures that may unnecessarily degrade utility. This motivates our proposed measure, RAPID, which we introduce after discussing the role of differential privacy.

1.4 Differential privacy and attribute inference

A natural question is whether differential privacy (DP) already addresses the disclosure risks we aim to measure. We argue that DP and RAPID address complementary concerns.

Differential privacy provides a formal guarantee of output stability: the probability of any particular output changes by at most a multiplicative factor $\exp(\varepsilon)$ when a single individual's data is added to or removed from the dataset [Domingo-Ferrer et al., 2021]. This guarantee is independent of the intruder's background knowledge and does not rely on assumptions about data distributions or attacker capabilities.

Crucially, DP does not claim that an attacker cannot learn sensitive information about an individual. Rather, it guarantees that the attacker's information gain is essentially the same whether or not a specific individual's data are included in the dataset [Muralidhar and Ruggles, 2024]. The rationale is that if inclusion does not matter, then no individual-specific privacy breach can occur. This makes DP particularly effective at preventing *membership inference attacks*, where the adversary tries to determine whether a specific individual was present in the training data.

However, *attribute inference attacks* exploit a different mechanism. Rather than asking "Was this person in the dataset?", the attacker asks "What is this person's sensitive attribute, given the released data and what I know about them?" DP does not directly limit the accuracy of such inferences, because the risk stems from population-level patterns that the data preserve, not from any individual's participation. If strong correlations exist in the population – e.g., between age, education, and disease status – then data released under DP (or any other anonymization method) may still encode those relationships, enabling accurate attribute inference [Blanco-Justicia et al., 2022, Muralidhar and Domingo-Ferrer, 2023].

This limitation is not unique to DP. Traditional anonymization methods face the same fundamental tension: preserving analytical utility requires preserving statistical relationships, but those same relationships enable inferential attacks. The question of whether releasing data increases an attacker's ability to infer sensitive attributes – relative to what could be inferred from background knowledge alone – applies regardless of the anonymization technique used. Moreover, DP faces a practical dilemma: stringent privacy guarantees (small ε) require noise levels that render outputs analytically useless, while relaxed guarantees (large ε) offer little meaningful protection [Domingo-Ferrer et al., 2025a].

For this reason, DP (and anonymization more broadly) should be complemented with explicit, scenario-based disclosure risk assessments that directly measure attribute-inference vulnerability. As Domingo-Ferrer et al. [2025b] argue, empirical disclosure risk assessment remains as unavoidable for synthetic or DP-protected data as it was under traditional utility-first approaches. RAPID addresses this need by quantifying how accurately an attacker could infer sensitive attributes from the released data, providing a practical complement to formal privacy guarantees.

Relation to membership disclosure and DP baselines. Recent work on *membership disclosure* situates attribute-inference risk within a differential privacy framework by comparing two anonymized datasets that differ only in the presence of a single individual– a "member" and a "non-member" version– and measuring the marginal improvement in inference accuracy when the individual is included [e.g., Francis and Wagner, 2025]. These approaches aim to quantify per-person *privacy loss due to inclusion*, grounded in DP's stability guarantee.

RAPID addresses a different question. While membership-based evaluations focus on individual-level stability ("Does my inclusion change the risk?"), RAPID focuses on population-level vulnerability ("How often could an attacker be

confidently correct?”). Both approaches contribute to understanding disclosure risk, but they serve distinct purposes: membership-focused methods are suited to certifying DP-style guarantees, while RAPID supports practical risk–utility assessment for public-use data releases.

1.5 Contributions

This paper makes the following contributions.

1. **A new disclosure risk measure under a realistic threat model.** We propose RAPID (Risk of Attribute Prediction–Induced Disclosure), a measure of attribute-inference disclosure risk for anonymized microdata, with a focus on fully synthetic data. RAPID quantifies the proportion of records for which an attacker can confidently infer sensitive attributes. Unlike differential privacy, which measures whether an individual’s inclusion changes the risk, RAPID measures how often inference succeeds across the released dataset – a practical summary of disclosure vulnerability that data custodians can directly interpret and compare.

   Our threat model assumes an intruder who has access only to the released data and to quasi-identifiers of target individuals, but no auxiliary sample of real data for validation or calibration. This reflects the conditions faced by external analysts of public-use files and avoids the practical limitations of holdout-based evaluations, whose conclusions depend heavily on holdout size and representativeness [Hittmeir et al., 2020, Platzer and Reutterer, 2021]. RAPID directly operationalizes the attacker’s objective: confidently inferring sensitive attributes of real individuals using models trained solely on the released data. This prediction-based vulnerability corresponds to *inferential disclosure* in the classical SDC taxonomy [Duncan and Lambert, 1989, Hundepool et al., 2012b].

2. **Baseline-normalized confidence scoring for categorical attributes.** We introduce a normalized gain measure that compares an attacker’s predicted probability for the true class to a baseline determined by class prevalence in the original data. This yields a calibrated notion of *confidence beyond chance* that explicitly accounts for class imbalance and avoids overstating risk in skewed distributions. By focusing on confidence-adjusted correctness rather than raw accuracy, RAPID captures the extent to which released data enable *meaningful* inference about sensitive attributes.

3. **A unified framework for categorical and continuous sensitive attributes.** RAPID provides a consistent formulation for both discrete and continuous confidential variables. For categorical attributes, inference risk is quantified via baseline-normalized prediction confidence; for continuous attributes, it is defined through tolerance-based relative prediction error. This unified framework allows inference risk to be assessed consistently across variable types without discretizing continuous outcomes, introducing arbitrary binning, or requiring counterfactual datasets that compare member versus non-member scenarios.

4. **Threshold-based, policy-interpretable risk summaries.** By summarizing disclosure risk as the proportion of records exceeding a confidence or accuracy threshold, RAPID yields interpretable, bounded metrics that are easy to communicate and tunable to institutional or regulatory risk tolerances. This formulation aligns naturally with risk–utility decision frameworks commonly used in statistical disclosure control and supports consistent comparison across different synthesizers, parameter settings, and data releases.

5. **Record-level risk indicators enabling diagnostic analysis.** Although RAPID reports an aggregate disclosure metric, it is constructed from per-record risk indicators. This design enables granular analyses of inference vulnerability, including the identification of high-risk records, subgroup-specific risk patterns, and combinations of quasi-identifiers that disproportionately contribute to disclosure risk. These record-level signals help data curators understand the drivers of risk and apply targeted mitigation strategies.

6. **Empirical validation and robustness analysis.** Through simulation studies and real-data illustrations, we demonstrate how RAPID scales with dependency strength between quasi-identifiers and sensitive attributes, how threshold choice affects risk classification, and how results remain stable across a range of attacker models. We also show how RAPID can be combined with holdout-based assessments when auxiliary real data are available, supporting internal benchmarking without altering the core threat model.

7. **Open-source implementation.** RAPID is implemented in R and integrates with existing synthetic data generation pipelines. The implementation supports multiple attacker models, bootstrap confidence intervals, and diagnostic outputs. By focusing on population-level vulnerability rather than per-individual privacy loss, RAPID complements formal privacy frameworks such as differential privacy and is particularly suited to evaluating public-use microdata releases.

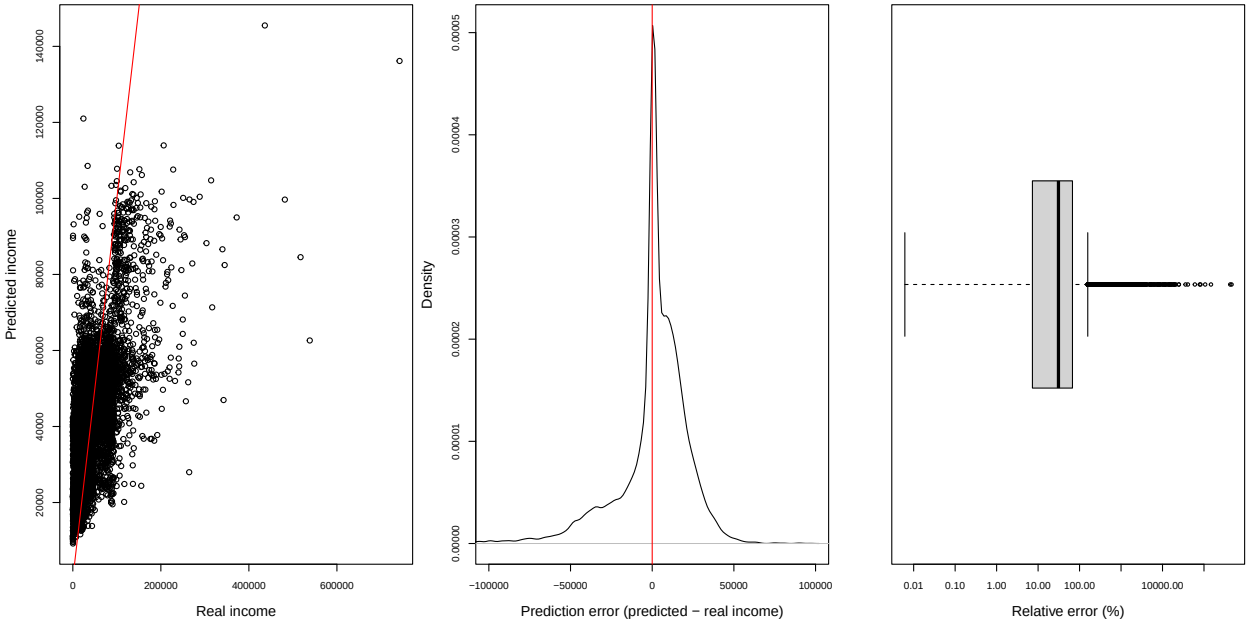

Figure 2: Attribute inference attack for a continuous sensitive variable (`income`) using a model trained on synthetic data and applied to real covariates. Left: predicted vs. true income; middle: prediction error distribution; right: relative error distribution. RAPID provides analogous risk quantification for categorical sensitive attributes.

1.6 Motivating Example

To illustrate the inference threat RAPID addresses, we conducted a simple attack experiment. Using the `eusilc` public-use file – a close-to-reality synthetic population based on the European Union Statistics on Income and Living Conditions (EU-SILC), which contains complex household structures and a rich set of demographic and socioeconomic variables [Alfons et al., 2011] – we generated a fully synthetic dataset with `synthpop` [Nowok et al., 2016]. We then trained random forest models [Wright and Ziegler, 2017] on the synthetic data to predict two sensitive attributes: the categorical variable `marital_status` and the continuous variable `income`.

Applying these synthetic-trained models to real covariate values reveals meaningful disclosure risk. For `marital_status`, the classifier achieved 82% accuracy on real data despite never seeing true labels during training. For `income`, many predictions fell close to true values (Figure 2), indicating that an attacker could approximate real incomes within a practically relevant range.

This experiment demonstrates that standard metrics like accuracy or mean squared error, while useful for evaluating predictive performance, do not adequately characterize disclosure risk. They ignore the attacker's *confidence* in correct predictions, are sensitive to class imbalance, and provide no policy-interpretable threshold. These limitations motivate the formal development of RAPID in the following section, which provides a unified framework for both categorical and continuous sensitive attributes.

2 The RAPID Measure

This section formalizes RAPID (Risk of Attribute Prediction–Induced Disclosure). Recall that our threat model assumes an attacker who trains a predictive model on the released data and applies it to individuals whose quasi-identifiers are known. RAPID quantifies how often such an attacker can make *confidently correct* inferences about a sensitive attribute.

The key idea is to compare the attacker's prediction performance against a baseline that ignores quasi-identifiers entirely. For categorical attributes, this baseline is the marginal frequency of each class in the original data – the probability an attacker would assign to the correct class by simply knowing class prevalences without using any quasi-identifier information. For continuous attributes, the baseline is a reference prediction error.

A record is flagged as at-risk only when the attacker's performance substantially exceeds this baseline, ensuring that RAPID captures genuine information leakage rather than artifacts of class imbalance or distributional properties.

2.1 Setup and Notation

Let the original microdata be denoted by

$$\mathbf{Z} = [\mathbf{X}_Q,\ \mathbf{X}_U,\ \mathbf{y}] \quad ,$$

where $\mathbf{X}_Q \in \mathbb{R}^{n \times p_Q}$ represents quasi-identifiers known to an attacker, $\mathbf{X}_U \in \mathbb{R}^{n \times p_U}$ denotes additional attributes not available to the attacker, and $\mathbf{y} \in \mathbb{R}^n$ is the confidential sensitive attribute.

The released (e.g., synthetic) dataset is

$$\mathbf{Z}^{(s)} = [\mathbf{X}_Q^{(s)},\ \mathbf{X}_U^{(s)},\ \mathbf{y}^{(s)}] \quad .$$

An attacker observes the released data $\mathbf{Z}^{(s)}$ together with quasi-identifiers $\mathbf{X}_Q$ of target individuals (e.g., from external sources), but has no access to $\mathbf{X}_U$ or $\mathbf{y}$.

The attacker trains a predictive model $\mathcal{M}$ on $(\mathbf{X}_Q^{(s)}, \mathbf{y}^{(s)})$ to obtain parameter estimates $\hat{\Theta}^{(s)}$, then applies this model to $\mathbf{X}_Q$ to produce predictions $\hat{\mathbf{y}}$ of the sensitive attribute.

For notational simplicity, we write $\mathbf{Z} = [\mathbf{X},\ \mathbf{y}]$, where $\mathbf{X}$ denotes quasi-identifiers, since $\mathbf{X}_U$ plays no role in risk estimation. By default, we evaluate risk across all n records in the original data. However, RAPID's record-level design allows risk assessment for any target set – a specific subpopulation, a sample, or even individuals not in the original data whose quasi-identifiers are known.

Figure 3 illustrates this threat model from a data custodian's viewpoint.

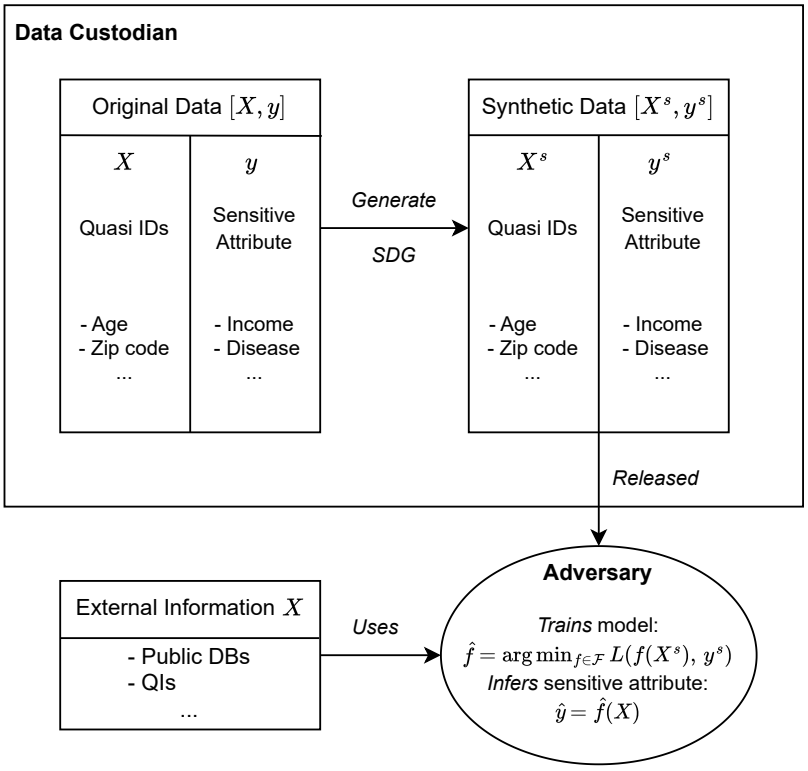

Figure 3: RM's comment on population X not included, Figure needs redrawing, Figure text needs adaptation. Inferential disclosure threat model. A data custodian releases anonymized data derived from original data containing quasi-identifiers ($\mathbf{X}$) and a sensitive attribute ($\mathbf{y}$). An attacker with access to the released data and external knowledge of individuals' quasi-identifiers ($\mathbf{X}^*$, which may include records not in the original data) trains a predictive model to infer sensitive attributes.

2.2 RAPID for a Categorical Sensitive Attribute

When $\mathbf{y}$ is categorical, we define risk in terms of the model-assigned probability to the true class label. For each record i, the attacker's model assigns probability

$$g_i = \Pr(\hat{y}_i = y_i \mid \mathbf{x}_i, \hat{\Theta}^{(s)})$$

to the true class y_i. To evaluate whether this confidence is unusually high, we compare it against a baseline b_i defined as the marginal proportion of class y_i in the original data:

$$b_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(y_j = y_i)$$

This baseline represents the prediction confidence achievable by simply guessing according to the marginal class distribution, without using any quasi-identifier information. We use the original marginals (rather than synthetic marginals) to reflect a conservative scenario in which the attacker has access to the true class distribution through external sources.

We compute a normalized gain score,

$$r_i = \frac{g_i - b_i}{1 - b_i}$$

that measures the improvement in prediction confidence over baseline, normalized by the maximum possible improvement (reaching perfect confidence $g_i = 1$). The score satisfies:

- $r_i < 0$: model performs worse than baseline
- $r_i = 0$: model performs at baseline (no information gain)
- $0 < r_i < 1$: partial information gain from quasi-identifiers
- $r_i = 1$: perfect prediction (complete disclosure)

A record is considered at risk if $r_i > \tau$, where $\tau \in (0, 1)$ is a policy-defined threshold. We recommend $\tau = 0.3$ as a default, meaning a record is flagged when the attacker achieves at least 30% of the maximum possible improvement over baseline. This threshold balances sensitivity (detecting meaningful inference gains) with specificity (avoiding false positives from minor improvements); Section 5 provides empirical guidance on threshold selection. TO BE CHECKED and COMPARED WITH RESULTS SECTION The categorical RAPID metric is then:

$$\text{RAPID}^{\text{cat}}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(r_i > \tau)$$

representing the proportion of records for which the released data enable inference substantially better than the baseline rate.

2.3 RAPID for a Continuous Sensitive Attribute

When the confidential attribute $\mathbf{y}$ is continuous, we assess whether the model prediction $\hat{y}_i$ is sufficiently close to the true value y_i. Several error metrics can be used, depending on the application context.

Relative error. When values are strictly positive and percentage-based interpretation is meaningful (e.g., income, expenditure), we compute the relative prediction error:

$$e_i = \frac{|y_i - \hat{y}_i|}{|y_i|} \times 100$$

A record is at-risk if $e_i < \varepsilon$, where ε is a percentage threshold (e.g., 10%–20%). This metric is undefined or unstable when $y_i = 0$ or $|y_i|$ is very small; in such cases, absolute error should be used instead.

Absolute error. When the scale of $\mathbf{y}$ has intrinsic meaning or values can be zero (e.g., counts, differences), we use:

$$e_i = |y_i - \hat{y}_i|$$

A record is at-risk if $e_i < \varepsilon$, where ε is now an absolute tolerance chosen based on domain knowledge (e.g., "within \$1000" for income).

Normalized error. Alternatively, error can be normalized by the standard deviation or range of $\mathbf{y}$ to yield scale-free comparisons across variables.

In all cases, the continuous RAPID metric is:

$$\text{RAPID}^{\text{cont}}(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(e_i < \varepsilon)$$

representing the proportion of records where predictions fall within the specified error tolerance. Note that while aggregate error metrics (MAE, RMSE) are insufficient for disclosure assessment—as they do not identify which records are at risk—they can inform the choice of ε by characterizing typical prediction accuracy.

2.4 Computation

Algorithm 1 formalizes the RAPID evaluation protocol: train a predictive model on the released data, apply it to target quasi-identifiers, compute per-record risk scores, and aggregate to obtain RAPID.

Algorithm 1 RAPID Evaluation Protocol

Require: Original data $\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$; released data $\mathbf{Z}^{(s)}$
Require: Thresholds $\tau \in (0, 1)$ (categorical; default $\tau = 0.3$) and $\varepsilon > 0$ (continuous; default $\varepsilon = 10\%$)

```
Fit predictive model M on Z^(s) to obtain Θ̂^(s)
Compute predictions ŷ = M(X; Θ̂^(s))
for i = 1, ..., n do
    if y is categorical then
        g_i ← Pr(ŷ_i = y_i | x_i)
        b_i ← n^{-1} Σ_{j=1}^{n} I(y_j = y_i)
        r_i ← (g_i − b_i)/(1 − b_i)
        I_i ← I(r_i > τ)
    else
        e_i ← |y_i − ŷ_i|/(|y_i| + δ)
        I_i ← I(e_i < ε)
```

Output: record-level risks $\{r_i\}$ or $\{e_i\}$, indicators $\{I_i\}$, and

$$\text{RAPID} \leftarrow \frac{1}{n}\sum_{i=1}^{n} I_i$$

RAPID is thus an empirical estimate of the probability that a randomly selected record is subject to successful attribute inference under the specified attacker model.

When evaluating multiple models $\mathcal{M} \in \mathcal{S}$, report both the average and conservative envelope:

$$\overline{\text{RAPID}}(\cdot) = \frac{1}{|\mathcal{S}|}\sum_{m\in\mathcal{S}} \text{RAPID}^{(m)}(\cdot), \qquad \text{RAPID}_{\max}(\cdot) = \max_{m\in\mathcal{S}} \text{RAPID}^{(m)}(\cdot)$$

2.5 Implementation Considerations

RAPID is applicable to any algorithm producing class probabilities or point predictions. To reflect a strong attacker, we recommend evaluating powerful learners (e.g., random forests, gradient boosting) and reporting the maximum observed risk. When multiple released datasets are available, risk can be averaged across them or reported as the worst-case value,

depending on desired conservativeness. Thresholds τ and ε should reflect policy requirements; we recommend defaults of $\tau = 0.3$ and $\varepsilon = 10\%$, which we validate empirically in Section 5. For data-driven selection, one can permute $\mathbf{y}$, recompute RAPID, and choose the threshold placing observed risk above the 95th percentile of the resulting null distribution. Uncertainty can be quantified via bootstrap resampling or, treating RAPID as a binomial proportion, using Wilson score [Wilson, 1927] or Clopper–Pearson confidence intervals [Clopper and Pearson, 1934].

Evaluating Synthetic Data Generators. To assess the general disclosure risk of a synthetic data generator (SDG), we recommend k-fold cross-validation: partition the original data into k folds (stratified by $\mathbf{y}$ if categorical), generate synthetic data from k-1 folds, and evaluate RAPID on the held-out fold. Aggregating across folds yields the expected risk $\mathbb{E}$[RAPID] with confidence intervals, providing a robust assessment of the SDG's average disclosure risk. This approach is implemented via the `rapid_synthesizer_cv()` function and is particularly useful for selecting between alternative synthesis methods (e.g., CART vs. parametric) or optimizing hyperparameters. Once a method is selected, standard RAPID should be applied to the final synthetic data product to obtain record-level risk assessments prior to the data release.

2.6 Properties

RAPID has several structural properties that follow directly from its definition. First, it targets successful and confident attribute inference events rather than aggregate predictive accuracy, aligning the risk measure with inferential disclosure in the classical SDC taxonomy. Second, the normalization by the empirical base rate ensures that categorical risk scores are invariant to class imbalance, so that common outcomes do not spuriously inflate disclosure risk.

Third, RAPID is bounded in $[0, 1]$, as it is defined as the empirical mean of record-level disclosure indicators. This facilitates comparison across datasets, synthesizers, and attacker models. Fourth, the construction is flexible with respect to the attacker model: any predictive model capable of producing point predictions or class probabilities may be used, allowing RAPID to be evaluated under strong and adaptive attacker assumptions.

Finally, RAPID is monotone in the decision thresholds τ and ε, respectively. Increasing the categorical threshold τ or decreasing the continuous tolerance ε can only reduce the number of records classified as at risk. This monotonicity supports transparent sensitivity analysis and makes RAPID well suited for practical disclosure control.

Because RAPID operates at the record level, it also enables subgroup-specific and conditional risk analysis by restricting the indicator average to subsets of the quasi-identifier space (see, e.g., Figure 6).

3 Toy Example

We demonstrate the categorical RAPID metric on a simple example. Consider a dataset with 100 records where class `healthy` has 60% prevalence (marginal proportion). A model trained on released data is applied to three original records, all truly belonging to class `healthy`:

- Record 1: $g_1 = \Pr(\hat{y}_1 = \text{healthy} \mid \mathbf{x}_1) = 0.70$
- Record 2: $g_2 = \Pr(\hat{y}_2 = \text{healthy} \mid \mathbf{x}_2) = 0.85$
- Record 3: $g_3 = \Pr(\hat{y}_3 = \text{healthy} \mid \mathbf{x}_3) = 0.55$

The baseline is the marginal frequency of class `healthy` in the original data:

$$b_i = 0.60 \quad \text{for all three records (same class)}$$

We compute the normalized gain for each record:

$$r_1 = \frac{0.70 - 0.60}{1 - 0.60} = \frac{0.10}{0.40} = 0.25$$
$$r_2 = \frac{0.85 - 0.60}{1 - 0.60} = \frac{0.25}{0.40} = 0.625$$
$$r_3 = \frac{0.55 - 0.60}{1 - 0.60} = \frac{-0.05}{0.40} = -0.125$$

Interpretation:

- Record 1: 25% of maximum improvement over baseline

- Record 2: 62.5% of maximum improvement (high confidence!)
- Record 3: Below baseline (model worse than guessing)

Using the default threshold $\tau = 0.3$, we flag records where $r_i > 0.3$:

$$\mathbb{I}\{r_1 > 0.3\} = 0, \quad \mathbb{I}\{r_2 > 0.3\} = 1, \quad \mathbb{I}\{r_3 > 0.3\} = 0$$

The RAPID metric is:

$$\text{RAPID}^{\text{cat}}(0.3) = \frac{1}{3}(0 + 1 + 0) = 0.33$$

This indicates that 33% of records have predictions substantially better than the baseline rate. Record 2, with 62.5% normalized gain, represents a high disclosure risk: the released data enables confident inference beyond what marginal class frequencies alone would allow. This example illustrates how RAPID identifies records where quasi-identifiers provide meaningful information gain, rather than merely measuring prediction accuracy.

Example (continuous attribute):

Now consider the case where the sensitive variable is continuous, such as income. Suppose an attacker trains a regression model on released data and applies it to the original covariates of three individuals. Let the true incomes and the model's predictions be:

- Record 1: $y_1 = 50{,}000$, $\hat{y}_1 = 47{,}000$
- Record 2: $y_2 = 35{,}000$, $\hat{y}_2 = 39{,}000$
- Record 3: $y_3 = 80{,}000$, $\hat{y}_3 = 90{,}000$

The relative prediction errors (as percentages of true values) are:

$$e_1 = \frac{|50{,}000 - 47{,}000|}{50{,}000} \times 100 = 6\%$$
$$e_2 = \frac{|35{,}000 - 39{,}000|}{35{,}000} \times 100 = 11.4\%$$
$$e_3 = \frac{|80{,}000 - 90{,}000|}{80{,}000} \times 100 = 12.5\%$$

Using a threshold of $\varepsilon = 10\%$, we flag records where predictions fall within 10% of the true value:

$$\mathbb{I}\{e_1 < 10\%\} = 1, \quad \mathbb{I}\{e_2 < 10\%\} = 0, \quad \mathbb{I}\{e_3 < 10\%\} = 0$$

The continuous RAPID metric is:

$$\text{RAPID}^{\text{cont}}(10\%) = \frac{1}{3}(1 + 0 + 0) = 0.33$$

This indicates that for one-third of individuals, the attacker's model predicted income within 10% relative error. Record 1, with only 6% error, represents a disclosure risk: the released data enables an attacker to infer income with high precision. This form of attribute disclosure is directly relevant for risk assessment but would not be captured by aggregate metrics like mean absolute error alone.

3.1 Software and defaults

The proposed risk measure is implemented in R. A dedicated package, `RAPID`, accompanies this paper and is publicly available on GitHub.[1] The reference implementation builds on well-established libraries: `ranger` [Wright and Ziegler, 2017] for random forests, and optionally `xgboost` [Chen et al., 2025] for gradient boosting. The package provides functions for computing RAPID for both categorical and continuous sensitive attributes, including baseline normalization, threshold calibration, and bootstrap-based uncertainty quantification.

Unless otherwise noted, we adopt the following default settings in our experiments: the attacker model $\mathcal{M}$ is a random forest with 500 trees and probabilistic outputs enabled. For categorical attributes, we use the default threshold $\tau = 0.3$; for continuous attributes, we use the default relative error tolerance $\varepsilon = 0.10$. Uncertainty is quantified via a nonparametric bootstrap over the original dataset (500 replicates), with percentile-based confidence intervals.

[1] `https://github.com/XXX/RAPID`

4 Real data illustration: UCI Adult (Census Income)

We illustrate the workflow on the *Adult* dataset (UCI Machine Learning Repository; 48,842 rows, 14 attributes; binary confidential attribute $\mathbf{y} = \mathbb{I}(\text{income} > \$50\text{K})$) [Becker and Kohavi, 1996]. The covariates, $\mathbf{X}$, include age, education, hours-per-week, marital status, etc. We treat $\mathbf{y}$ as the confidential, sensitive variable and $\mathbf{X}$ as potentially known quasi-identifiers.

Pre-processing. We standardize continuous features within the original file (means/SDs computed on $\mathbf{Z}$ and then applied to $\mathbf{Z}^{(s)}$ to avoid leakage), and one-hot encode categorical predictors consistently across original and synthetic files.

Synthesizers. We consider a CART-based tabular synthesizer (as implemented in `synthpop` [Nowok et al., 2016]), which is trained solely on $\mathbf{Z}$ and produces $M = 5$ synthetic replicates.

Attack models. We evaluate an attacker suite $\mathcal{S} = \{\text{RF, GBM, } \ell_1\text{-logistic}\}$, trained on each synthetic replicate and scored on the real covariates $\mathbf{X}$.

Metrics and reporting. We compute $\mathrm{RAPID}^{\mathrm{cat}}(\tau)$ across a range of threshold values and visualize the results as a threshold curve (Figure 4). For each synthetic replicate, we average RAPID across the $M = 5$ replicates and report 95% bootstrap confidence intervals. We additionally stratify RAPID by true class y to reveal whether disclosure risk differs across outcome categories—for example, whether high-income individuals are more identifiable than low-income individuals.

Sensitivity and diagnostics. To understand how RAPID responds to threshold choices, we plot $\mathrm{RAPID}^{\mathrm{cat}}(\tau)$ across a grid of τ values (Figure 4), visualizing how disclosure risk decays as stricter normalized gain thresholds are imposed.

Practitioners may consider additional diagnostics:

- **Class balance:** Reporting the baseline probability b_k for each class k exposes the influence of class prevalence on the normalized gain.
- **Joint utility–risk view:** To contextualize disclosure risks, utility metrics such as predictive accuracy of models trained on $\mathbf{Z}^{(s)}$ but evaluated on $\mathbf{Z}$ can be reported alongside RAPID.

To assess inferential disclosure risk, we applied RAPID to the UCI Adult dataset using `income` as the sensitive attribute. Five synthetic datasets were generated via the CART synthesizer (`synthpop` package). For each replicate, we trained a random forest attacker model on the synthetic data to infer the sensitive attribute from quasi-identifiers, then evaluated the model's predictions on the original dataset.

We computed RAPID across a range of threshold values $\tau \in [0, 1]$ to examine how disclosure risk varies with the stringency of the threshold. To quantify uncertainty, we performed non-parametric bootstrapping with $R = 500$ resamples from the original dataset, providing robust percentile-based confidence intervals.

Figure 4 shows how RAPID varies with the normalized gain threshold τ. As τ increases from 0 to 1, fewer records exceed the threshold, demonstrating the monotonicity property discussed in Section 2.6. This threshold curve enables the data provider to calibrate disclosure risk according to their privacy requirements.

5 Simulation study

5.1 Overview and Design

To validate RAPID and investigate factors influencing attribute-inference risk (per row), we conducted four simulation studies:

1. **Dependency strength:** How does risk scale with QI–sensitive attribute relationships? ($\kappa \in [0, 100]$)
2. **Threshold sensitivity:** How does τ affect risk across dependency regimes? (5 κ levels × 19 τ values)
3. **QI attribution:** Which quasi-identifiers drive risk? (Regression-based analysis)
4. **Attacker robustness:** Is RAPID consistent across models? (RF, CART, GBM comparison)

All simulations use synthetic health microdata with six variables (gender, age, education, income, health score, disease status). We control dependency strength via a global parameter $\kappa \geq 0$, with full details in Appendix A.

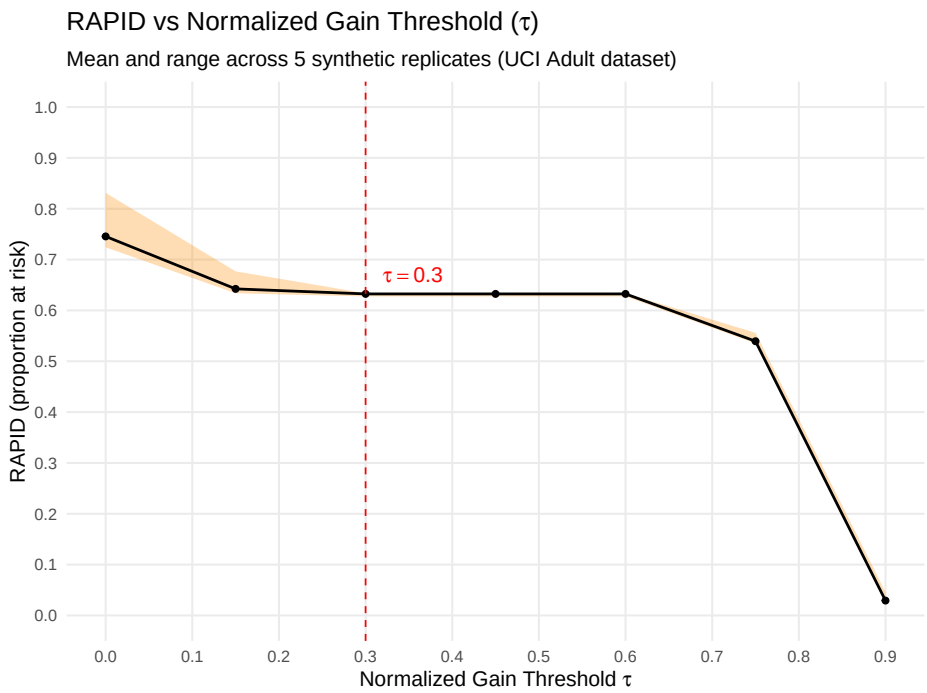

Figure 4: rerun with finer-grained tau values, increaded bootstraps and rf instead of cart - in real-data-application-new-rapid.RRAPID as a function of the normalized gain threshold τ for the UCI Adult dataset. The curve demonstrates how disclosure risk decreases as stricter thresholds are imposed.

5.2 Data Generation Process

We simulate n independent records with six variables: gender (G), age (A), education (E), income (I), health score (H), and disease status (D). The design encodes realistic dependencies through a latent socioeconomic status (SES) variable, with dependency strength controlled by parameter $\kappa \geq 0$.

Signal and noise weights are derived from κ as:

$$w_{\text{signal}} = \sqrt{\frac{\kappa}{1+\kappa}}, \quad w_{\text{noise}} = \sqrt{\frac{1}{1+\kappa}}.$$

At $\kappa = 0$, relationships are weak ($w_{\text{signal}} \approx 0$); at $\kappa \gg 1$, dependencies approach deterministic strength ($w_{\text{signal}} \to 1$). Age follows a truncated normal; education is ordinal derived from a latent variable; income is log-linear; health score uses sigmoid transformation; disease status is generated via multinomial logit with κ-scaled coefficients; gender is binary with mild SES dependency. Complete mathematical specifications appear in Appendix A.

5.3 Dependency Strength

We varied dependency parameter κ from 0 to 100 (101 values, 10 replications each) to investigate how attribute-inference risk scales with quasi-identifier–sensitive attribute relationship strength. Subplot a) in Figure 5 shows a S-shaped trajectory demonstrates that risk escalates rapidly when transitioning from weak to moderate dependencies, then saturates as relationships approach deterministic strength. Saturation near 0.97 rather than 1.0 reflects inherent noise from the CART synthesizer's sampling. Attacker accuracy follows a similar trajectory (0.70 to 0.98), demonstrating that RAPID reliably reflects actual prediction performance: datasets with high RAPID face genuinely elevated disclosure risk.

Having established that risk scales with dependency strength, we next investigate how the threshold parameter τ interacts with this dependency structure. Simulation 2 addresses this challenge by examining threshold sensitivity across the full dependency spectrum.

5.4 Threshold Sensitivity

We examined how confidence threshold τ affects RAPID across five dependency levels ($\kappa \in \{0, 5, 10, 20, 50\}$), varying τ from 0.05 to 0.95 in 0.05 increments (10 replications per combination). Subplot b) in Figure 5 reveals a qualitative shift in curve geometry: at low dependency ($\kappa = 0$, gray), RAPID decreases convexly, dropping rapidly from 0.50 to near-zero by $\tau = 0.60$, indicating diffuse attacker confidence. At high dependency ($\kappa \geq 5$, blue/red), curves become concave, remaining elevated (> 0.85) until stringent thresholds ($\tau > 0.70$), reflecting concentrated confidence distributions near certainty. This convex-to-concave transition marks a shift from weak to strong adversarial inference capability. Practically, data curators should choose higher thresholds ($\tau \geq 0.70$) for strongly dependent data to avoid flagging nearly all records, while lower thresholds ($\tau \approx 0.30$–0.40) suffice for weakly dependent data where they effectively separate high-confidence from baseline predictions.

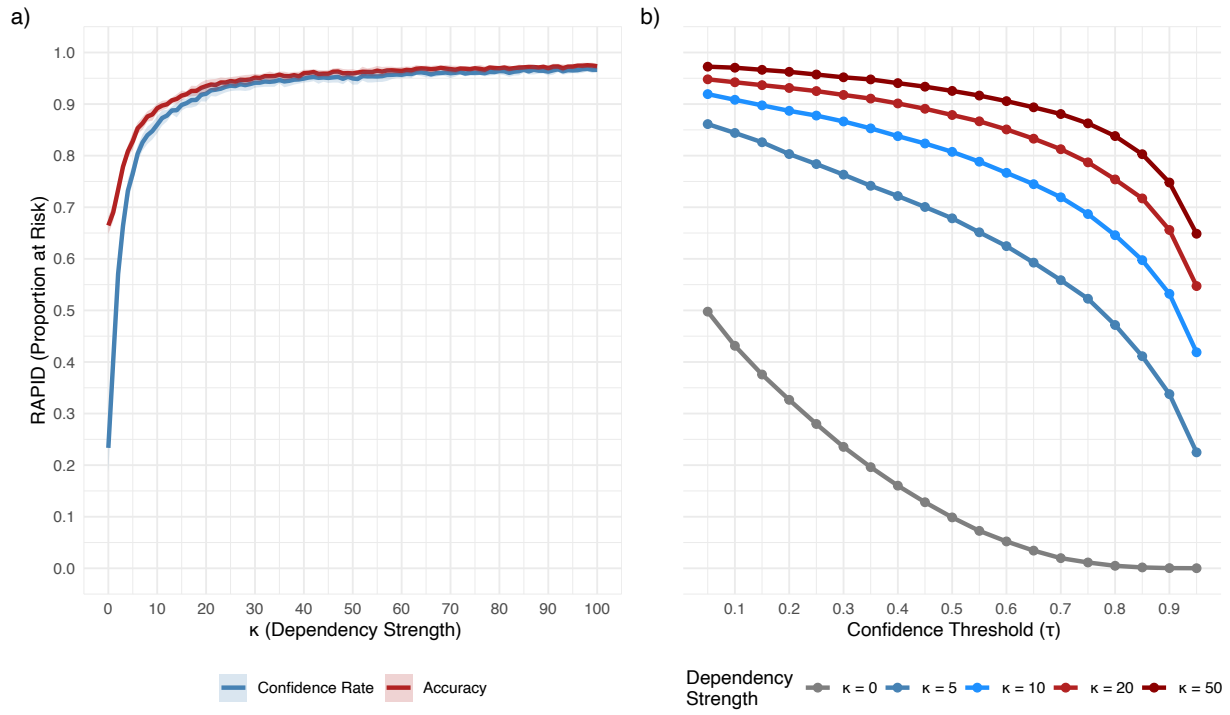

Figure 5: Impact of dependency strength and confidence threshold on RAPID:

(a) RAPID and attacker accuracy increase monotonically with dependency strength κ. RAPID rises from 0.25 at $\kappa = 0$ to 0.97 at $\kappa = 100$, with steepest increases at low κ values. This S-shaped growth demonstrates that attribute-inference risk escalates rapidly when transitioning from weak to moderate quasi-identifier–sensitive attribute relationships, then saturates as dependencies approach deterministic levels.
(b) RAPID vs confidence threshold τ for varying κ. At low dependency ($\kappa = 0$, gray), the curve is convex, reflecting diffuse attacker confidence where most records are filtered out at moderate thresholds. At high dependency ($\kappa \geq 5$, blue/red), curves become concave, remaining elevated until stringent thresholds ($\tau > 0.7$) are applied. This transition reflects a qualitative shift in attacker confidence distributions as dependencies strengthen.
Both panels: Mean $\pm$ 1 SD over 10 simulations; $n = 1000$ records, CART synthesizer, Random Forest attacker. Panel (a): $\tau = 0.3$.

5.5 Quasi-Identifier Attribution

To identify which quasi-identifiers drive attribute-inference risk, we fitted logistic regression models predicting at-risk status from demographic characteristics across 50 simulations ($\kappa = 10$, $\tau = 0.3$). Figure 6 displays coefficient distributions relative to reference categories (marked in red). This regression-based approach demonstrates how RAPID's per-record risk flags enable granular attribution analysis: by modeling what predicts at-risk status, data curators can identify which quasi-identifier combinations elevate disclosure vulnerability. The analysis reveals heterogeneity across

quasi-identifier levels. This differentiation is valuable for understanding risk drivers in synthetic data and assessing which demographic characteristics the attacker exploits most effectively.

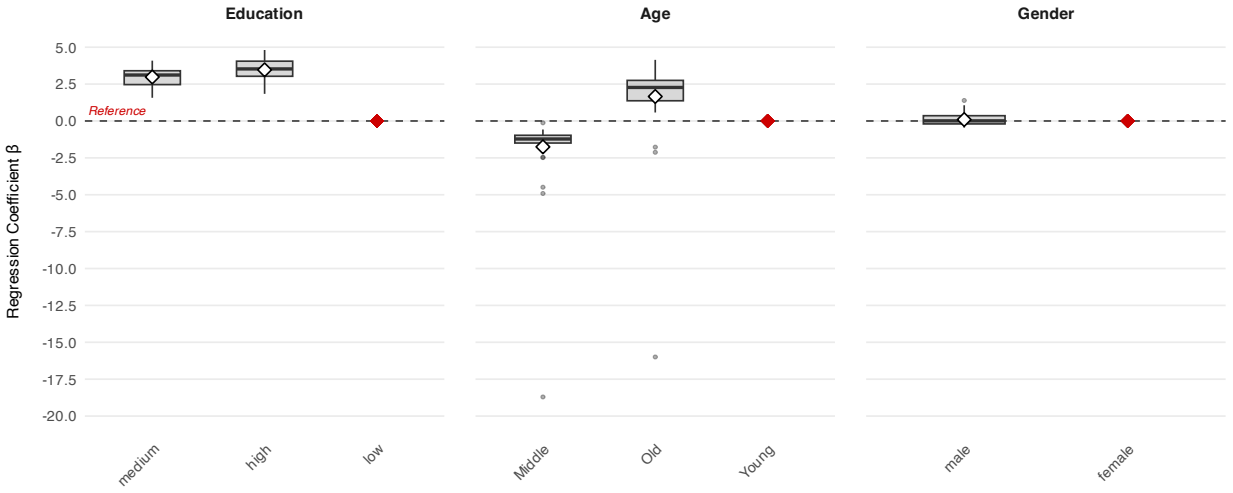

Figure 6: Attribution analysis of re-identification risk by quasi-identifier. Boxplots display regression coefficients (β) from logistic models predicting at-risk status across 50 simulations (κ=10, τ=0.3, n=1000). White diamonds = means; red points = reference categories (β=0). Positive β indicates elevated re-identification risk.

5.6 Attacker Model Robustness

To assess whether RAPID is sensitive to attacker model choice, we compared three tree-based models (Random Forest, CART, GBM) at $\kappa = 10$ and $\tau = 0.3$ across 50 replications. Table 1 shows that all three models achieve similar RAPID values within a narrow range (0.864–0.876). This robustness is a desirable property for practical risk assessment: RAPID provides consistent risk estimates regardless of which specific tree-based method models the attacker, reducing sensitivity to modeling assumptions about adversarial capabilities. The narrow variation demonstrates that RAPID captures a stable notion of attribute-inference vulnerability rather than idiosyncrasies of particular algorithms. Data curators can thus apply RAPID confidently without requiring precise knowledge of adversary methods, as the metric reliably quantifies disclosure risk across reasonable attacker specifications.

Table 1: Attacker Model Comparison

Attacker	RAPID	SD	Accuracy	SD
CART	0.876	0.013	0.879	0.011
GBM	0.875	0.011	0.881	0.011
Random Forest	0.864	0.010	0.891	0.010

5.7 Computational cost

Let T_{train} be the time to fit $\mathcal{M}$ on $\mathbf{Z}^{(s)}$ and T_{score} the time to score $\mathbf{X}$. Per replicate, the dominant cost is T_{train}. Overall complexity is $O\big(M \cdot |\mathcal{S}| \cdot T_{\text{train}}\big)$. Bootstrap intervals add a multiplicative factor of B light-weight recomputations (reusing $\hat{\Theta}^{(s)}$, only resampling rows of $\mathbf{Z}$), so walltime scales roughly as $O\big(M \cdot |\mathcal{S}| \cdot T_{\text{train}} + B \cdot n\big)$.

5.8 Reproducible R sketch

Below we give a minimal R sketch for the categorical case; continuous y differs only in the score definition.

RAPID is implemented as an open-source R package (`rapidsynthpop`) with straightforward workflows integrating with the widely-used `synthpop` package for generating synthetic data. A typical evaluation requires only original data, synthetic data, and specification of quasi-identifiers and the sensitive attribute:

```
library(rapidsynthpop)
rapid_result <- rapid(
  original_data = data_orig,
  synthetic_data = data_syn,
  quasi_identifiers = c("age", "education", "gender"),
  sensitive_attribute = "disease_status",
  model_type = "rf",
  cat_tau = 0.3
)
```

The function returns sample-level risk rates and per-record risk scores. For categorical sensitive attributes, the primary output is the confidence rate: the proportion of records for which the attacker's predicted probability (relative to the model-estimated baseline) exceeds threshold τ. For continuous attributes, RAPID reports the roportion of records where relative prediction error falls below a specified tolerance.

6 Discussion

Table-based measures like DiSCO [Raab et al., 2025] assess attribute disclosure by flagging records as disclosive when synthetic data imply a unique target value for a given quasi-identifier combination q—that is, when the synthetic column proportion $p_{stq} = 1$ within that q-group. DiSCO then counts the number of original records in those q-groups with that target t ("Correct Original"). Formally,

$$\text{DiSCO} = 100 \times \sum_q \sum_t \frac{d_{tq} \mid p_{stq} = 1}{N_d},$$

where d_{tq} denotes original records in quasi-identifier group q with target t, and N_d is the total number of original records. For example, if every synthetic record for "60-year-old male smoker from region A" shows depression ($p_{\text{depression}|q} = 1$), and five original records match these quasi-identifiers with depression, those five records contribute to DiSCO. Related measures include DCAP and TCAP, which use different denominators to express conditional disclosure probabilities.

This approach offers transparency and integrates naturally with classical Statistical Disclosure Control concepts of key variables and within-group homogeneity. However, it has several limitations for modern synthetic data assessment.

First, DiSCO requires discretizing continuous sensitive attributes and quasi-identifiers into categorical bins to construct contingency tables, making results sensitive to these binning choices. Coarse bins (e.g., three broad age groups) yield larger equivalence classes (i.e., groups of records sharing the same quasi-identifier combination) that are less likely to exhibit perfect homogeneity ($p_{stq} = 1$), potentially underestimating risk. Fine bins (e.g., five-year age intervals) create smaller, more homogeneous equivalence classes but also sparser ones with fewer observations. There is no objective criterion for optimal binning, introducing subjectivity into risk assessment.

Second, DiSCO's deterministic threshold ($p_{stq} = 1$) is strict: it only flags risk when synthetic data exhibit perfect within-class homogeneity for the sensitive attribute. This misses inferential vulnerabilities from strong but non-deterministic patterns that adversaries using machine learning could exploit. Consider an equivalence class where 85% of synthetic records show depression and 15% show anxiety. An adversary training a predictive model on the synthetic data would learn to predict depression for this quasi-identifier pattern with high confidence (0.85), as the model captures the strong probabilistic association. This poses genuine disclosure risk for original records with these quasi-identifiers who have depression. Yet because $p_{stq} \neq 1$, DiSCO does not flag these records as at-risk. More generally, flexible models can exploit complex interactions that don't yield perfect homogeneity in individual equivalence classes. RAPID captures these vulnerabilities by training flexible models that mirror realistic adversarial capabilities and flagging records where model confidence substantially exceeds baseline, regardless of whether synthetic data exhibit $p_{stq} = 1$.

Third, while DiSCO identifies which original records are at-risk, it provides only binary classification (disclosive or not) without quantifying the strength of adversarial inference. All records with $p_{stq} = 1$ are treated identically regardless of the broader context: whether the quasi-identifier pattern is common or rare in the data, whether the sensitive attribute has high or low marginal prevalence, or whether alternative quasi-identifier combinations would yield similar predictions. RAPID produces individual-level confidence scores that vary continuously, enabling risk stratification: curators can identify not just which records are at elevated risk, but which face the highest vulnerability, facilitating targeted disclosure control and subgroup-specific diagnostics.

Fourth, DiSCO can flag records as at-risk due to synthetic data artifacts rather than genuine inferential vulnerability. If a synthesizer happens to produce a degenerate cell ($p_{stq} = 1$) for a quasi-identifier combination through sampling variability or model idiosyncrasies—even when the original data for that combination are heterogeneous—DiSCO

counts all matching original records as disclosive. This may overstate risk for combinations where no real adversary could reliably infer the sensitive attribute. RAPID avoids this by requiring that a predictive model trained on the synthetic data actually achieves high-confidence correct predictions on the original quasi-identifiers. Risk is only flagged when the attacker's model, using the released synthetic file, demonstrates genuine predictive capability—more closely aligned to a successful inference attack than to synthetic table quirks.

Fifth, DiSCO does not calibrate for class imbalance or baseline prediction difficulty. In datasets with skewed sensitive attribute distributions (e.g., 90% no condition, 8% depression, 2% anxiety), simply predicting the majority class yields high accuracy without genuine inference. DiSCO treats all $p_{stq} = 1$ cells equally regardless of whether the unanimous value is common or rare. RAPID addresses this through class-normalized confidence scoring: for categorical targets, it compares the attacker's predicted probability for the true class against a model-estimated marginal baseline. This ensures that risk reflects predictive advantage beyond naive baseline strategies, analogous to `synthpop`'s baseline scaling for utility metrics but applied to disclosure risk. Records are only flagged when the attacker is substantially more confident than chance, calibrating risk assessment to class prevalence.

Sixth, as quasi-identifier dimensionality or cardinality increases, DiSCO's contingency table approach becomes computationally and interpretatively unwieldy. With many high-cardinality categorical variables (e.g., occupation codes, geographic units), tables explode combinatorially and most cells become sparse or empty, making alignment across original and synthetic data fragile. RAPID trains a single predictive model once and evaluates predictions—often more scalable when quasi-identifiers would generate thousands of cells. Moreover, RAPID naturally provides actionable diagnostics through standard machine learning interpretability tools: feature importance rankings identify which quasi-identifiers drive risk most strongly, partial dependence plots reveal nonlinear relationships, and subgroup analyses highlight vulnerable demographic combinations. DiSCO, being table-based, offers transparency but less guidance on how to mitigate risk beyond coarsening keys.

Seventh, DiSCO's results depend critically on quasi-identifier selection and specification choices. Different quasi-identifier sets or category definitions (e.g., coarsening age from 10 bins to 3) can yield substantially different risk assessments for the same data, as cell homogeneity ($p_{stq} = 1$) changes with table granularity. Additionally, aligning quasi-identifier levels across original and synthetic datasets (including unioning categories that appear in one but not the other) introduces further complexity and sensitivity. RAPID sidesteps this fragility by treating quasi-identifiers as model features rather than pre-defined keys: the predictive model automatically learns which variables and interactions matter most for inference, without requiring hand-crafted contingency table specifications. Risk estimates remain stable across reasonable quasi-identifier definitions because the model adapts flexibly to the provided feature set.

RAPID addresses these limitations by explicitly modeling the adversary's inference task: training predictive models on synthetic data and evaluating them on original quasi-identifiers. This attack-realistic approach naturally handles continuous targets without discretization, detects risk even when $p_{stq} < 1$ through confidence thresholding, produces individual-level confidence scores enabling nuanced risk stratification, aligns with genuine predictive capability rather than synthetic artifacts, calibrates for class imbalance through baseline-normalized scoring, and scales to high-dimensional settings while providing actionable diagnostics. Where DiSCO asks "does this quasi-identifier group uniquely determine the sensitive value in synthetic data?" RAPID asks "could an adversary trained on synthetic data accurately infer this individual's sensitive attribute with confidence substantially exceeding baseline?"—directly simulating the inferential disclosure pathway under realistic modeling capabilities and class-aware evaluation.

Case: angreifer augment synthetic data wiht public data.

DiSCO flags records as attribute-disclosive when the synthetic data imply a unique target value for a given quasi-identifier combination q (i.e., the synthetic column proportion $p_{stq} = 1$ within that q-group), and it then counts the number of original records in those q-groups with that target t (hence "Correct Original"). Formally (their Eq. 11), DiSCO $= 100 \times \sum_q \sum_t (d_{tq} \mid p_{stq} = 1)/N_d$, with related measures like DCAP and TCAP, alignment of q/t levels across GT (original) and SD (synthetic), and options such as denominator limits and grouping continuous targets before cross-tabulation.

Why RAPID can be preferable over DiSCO:

1) Attack-realistic and model-based. RAPID explicitly trains a predictor on synthetic $(\mathbf{X}^{(s)}, \mathbf{y}^{(s)})$ and evaluates it on original $\mathbf{X}$, mirroring how an intruder would use released synthetic data to infer sensitive attributes. DiSCO is table-based on chosen keys q and does not model $y \mid \mathbf{X}$. This makes RAPID closer to a modern inference attack than a key-uniqueness check.

2) Works natively for continuous y. RAPID handles continuous targets via regression and an accuracy/risk threshold (e.g., relative error), avoiding discretization. DiSCO needs binning/grouping of continuous targets before cross-tabulation, and results can be sensitive to the grouping choice.

3) Individual-level risk, not just group flags. RAPID produces per-record scores (e.g., relative confidence r_i or accuracy indicators g_i) and summary curves over thresholds, enabling granular diagnostics and subgroup auditing. DiSCO is primarily an aggregated share of originals that become disclosive under synthetic q-groups.

4) No fragile dependence on quasi-identifier design. RAPID can use all non-sensitive features (or any subset) without hand-crafted q. DiSCO requires selecting and aligning q across GT/SD (including unioning levels), and results can change markedly with key choice or coarsening.

5) Captures complex structure. With flexible learners (RF/GBM/GLM, etc.), RAPID detects nonlinearities and interactions in $y \mid \mathbf{X}$. DiSCO's logic is based on degenerate synthetic conditional distributions within q (i.e., $p_{stq} = 1$), which can miss inferrability arising from complex, high-dimensional relations that don't show as perfect within-cell certainty.

6) Better aligned to "true" risk (fewer false alarms from synthetic artifacts). DiSCO can count records as risky when the synthetic data happen to be degenerate within a q cell—even when the original isn't uniquely determined by that q (the paper notes a distinct DiSDiO variant that requires $p_{dtq} = 1$). RAPID only flags risk when a model trained on SD actually predicts the original y well from original $\mathbf{X}, whichisclosertoasuccessfulattackratherthanasyntheticquirk.$

7) Calibrated to class imbalance via a predictive baseline. For categorical y, RAPID's relative-confidence score compares the model's probability for the true class to a marginal baseline estimated over the original $\mathbf{X}$. This is analogous in spirit to synthpop's baseline scaling (e.g., $baseCAP_d$), but operates directly on predicted probabilities, yielding interpretable "how much better than chance" risk summaries.

8) Scales in wide/high-cardinality settings. Building and aligning large contingency tables for many keys or high-cardinality factors (as DiSCO requires) can be heavy. RAPID trains a model once and evaluates predictions—often more scalable when q would explode combinatorially.

9) Actionable diagnostics. RAPID naturally provides feature importance, partial-dependence, and subgroup risk profiles—useful for custodians to decide which variables or relations to weaken. DiSCO is transparent and simple, but offers less guidance on how to mitigate beyond coarsening keys.

When DiSCO is still attractive • Transparent, key-based story for stakeholders comfortable with SDC notions of keys/uniques. • Built-in to synthpop with DCAP/TCAP comparators and options like denominator limits. • Quick categorical screening where clear quasi-identifiers are known a priori.

Bottom line:
DiSCO asks: "Does the synthetic table make a target value appear uniquely determined for this key?"
RAPID asks: "Could an attacker trained only on the synthetic data accurately infer my sensitive attribute from non-sensitives?"

For modern inference-attack realism, continuous targets, high-dimensional structure, and record-level diagnostics, RAPID offers clear advantages. For simple, explainable, key-centric screening that integrates with synthpop's tooling, DiSCO remains useful.

Critique of the holdout approach (for privacy/risk)

- Measures memorization, not inference. Its privacy statistic is the fraction of synthetic records that are closer to training than to holdout records (distance-to-closest-record, DCR). A value near 50
- Sensitive to discretization and distance choices. The framework discretizes all variables (with a cap on cardinality) and then uses Hamming (or another) distance. Risk conclusions can vary with binning thresholds, category lumping, and metric choice—especially in mixed-type, high-dimensional data.
- High-dimensional sparsity issues. Nearest-neighbor distances become unstable as dimensionality grows (curse of dimensionality). Even with binning, "closeness" can be hard to interpret and may over- or under-state record proximity.
- Requires internal holdout access. Proper evaluation needs a real holdout split of the original data. That's fine for an internal steward, but it doesn't mirror what an external attacker can do with only the released synthetic file.
- Utility–risk conflation gaps. The holdout framework's fidelity side (TVDs of k-way marginals) is useful for representativeness, but it's orthogonal to attribute-inference risk: a synthesizer can pass fidelity checks yet still leak labels via strong predictive relations; conversely, failing a DCR test doesn't quantify per-variable harm.
- Potential to be gamed by post-processing. Small perturbations or aggressive discretization can inflate DCR to look "safer" without materially reducing the ability to infer sensitive attributes.

Advantages of RAPID over holdout-based evaluation

- Directly measures attribute-inference risk. RAPID asks the attacker's question ("Can I correctly infer a confidential attribute from quasi-identifiers?") by training on the synthetic file and scoring on the real covariates. The holdout method instead assesses (i) fidelity via marginal distribution distances and (ii) privacy via nearest-neighbor closeness to training vs. holdout records; it does not quantify how well an adversary could predict the confidential label.
- Interpretable, policy-ready metric. RAPID yields a single, bounded risk rate—e.g., the share of people for whom an attacker is times more confident than the model's class-baseline (categorical) or within relative error (continuous). This is easier to explain than distributions of total-variation distances or nearest-neighbor gaps.
- Base-rate calibration. RAPID's class-normalized confidence ("relative confidence") avoids spurious risk inflation under class imbalance or miscalibration—issues common when using raw probabilities or accuracy. The holdout framework is model-free and therefore cannot calibrate risk to class prevalences or prediction confidence.
- Aligned attacker model. RAPID constrains the intruder to the released synthetic file (train-on-synthetic, test-on-original), matching real public-release settings. The holdout approach requires access to an internal holdout of the original data during assessment and targets memorization/overfitting rather than inferential disclosure.
- Comparable across releases and synthesizers. Because RAPID normalizes confidence by a model-implied baseline and reports a thresholded rate, numbers remain comparable even when marginal class distributions or calibration differ across releases; holdout TVDs and nearest-neighbor shares can shift with discretization, binning caps, and sample splits.

Bottom line. Holdout-based evaluation is valuable for fidelity and for detecting memorization/overfitting to training records in a model-free way. RAPID complements—and for disclosure risk, improves upon—this by quantifying exactly what many stewards and regulators care about: how often a capable intruder, using only the released synthetic file, could be confidently correct about a person's confidential attribute, with class-imbalance calibration and a single interpretable rate.

Public Use Files (PUFs) are datasets released openly, often on public websites, with minimal barriers to access. These files are heavily anonymized through traditional Statistical Disclosure Control (SDC) techniques such as generalization, suppression, and top/bottom coding. The emphasis is on eliminating both direct identifiers and key indirect identifiers to ensure that the risk of reidentification is extremely low. While some advocate for the use of DP to strengthen protections here, it's important to recognize that DP is not inherently designed for static data release. Applying DP to create "DP-anonymized" PUFs (e.g., DP-synthetic data) is technically possible but offers limited practical utility unless access is tightly controlled and a privacy budget is enforced, which contradicts the very openness of PUFs.

Scientific Use Files (SUFs), by contrast, are made available only under formal agreements, typically to accredited researchers with a legitimate need. These datasets are usually more detailed than PUFs and may retain some potentially identifying indirect variables. Though still anonymized, SUFs acknowledge the residual risk of linkage attacks, especially when external datasets are available. While DP could theoretically be applied to SUFs to enhance protections, its formal guarantees only hold under carefully controlled, bounded interaction models—not static release. Hence, using DP here may give a false sense of security, especially if data are copied or used repeatedly in analysis pipelines without strict budget enforcement.

RM: possible limitations:

- RAPID is not model-agnostic, i.e., the risk measure is strongly dependent on the fitted model. This also includes feature engineering. A sufficiently motivated attacker could derive new features from existing variables, potentially boosting prediction performance substantially. RAPID is therefore not a worst-case, conservative measure like the Bayesian risk, which provides an upper bound on risk. Instead, in comparison with Bayesian risk (Reiter, 2014) it can be interpreted as a lower-bound estimate: the true risk will be at least this large.
- What if there are multiple sensitive variables, which is basically always the case in practice?
- A comparison with bayesian risk measures for synthetic data could also be interesting, see:
  - Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data (Hu, 2021)
  - Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data (Reiter et al., 2014)
  - Implementation as R package: Bayesian Estimation of Attribute Disclosure Risks in Synthetic Data with the AttributeRiskCalculation R Package (Hornby Hu, 2022)

6.1 Ethical and legal considerations

The evaluation assumes the attacker does *not* access original labels beyond those implicit in $\mathbf{Z}$ for scoring. Risk numbers should be interpreted in the context of legal thresholds and harm models for the specific domain (e.g., health, labor). When reporting, avoid releasing per-record scores r_i or h_i; publish only aggregated metrics and confidence intervals.

6.2 Summary

The proposed workflow – train on synthetic, score on real covariates, normalize by model-implied baselines, and summarize by RCIR with uncertainty – yields an interpretable, model-agnostic measure of inference disclosure risk. The real data illustration shows how to operationalize the protocol on a standard benchmark; the same steps apply to domain datasets (e.g., longitudinal panels or mobility traces) with choice of τ/ε tailored to sensitivity.

Pilgram et al. [2025] erwaehne, dass es dazu passt als Mosaikstein.

7 conclusion

We introduced a calibrated, attacker-realistic measure of inference disclosure for synthetic microdata. By normalizing correct-class confidence to a class-specific baseline and summarizing the share above a threshold, the metric is interpretable, robust to class imbalance, and easily integrated with existing synthetic-data workflows. It complements utility diagnostics and aligns with modern intruder models from both official statistics and machine learning. Future work includes formal power analyses for threshold selection, extensions to multi-label outcomes, and integration with end-to-end risk-utility dashboards.

References

Andreas Alfons, Stefan Kraft, Matthias Templ, and Peter Filzmoser. Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20(3):383–407, 2011. doi: 10.1007/s10260-011-0163-2.

Andrés F. Barrientos, Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *The Annals of Applied Statistics*, 12(2):1124 – 1156, 2018. doi: 10.1214/18-AOAS1194.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Computing Surveys*, 55(8):Article 160, 1–16, 2022. doi: 10.1145/3547139. URL `https://doi.org/10.1145/3547139`.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, Jiaming Yuan, and David Cortes. *xgboost: Extreme Gradient Boosting*, 2025. URL `https://CRAN.R-project.org/package=xgboost`. R package version 3.1.2.1.

Charles J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistisk tidskrift*, 15:429–444, 1977.

Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM*, 64(7):33–35, June 2021. ISSN 0001-0782. doi: 10.1145/3433638. URL `https://doi.org/10.1145/3433638`.

Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. Do Privacy Models Deliver? *arXiv preprint*, 2025a. URL `https://arxiv.org/abs/2510.11299`. Argues that DP incurs unacceptable utility loss for small privacy budgets, while its guarantee becomes meaningless for large budgets.

Josep Domingo-Ferrer, David Sánchez, and Krishnamurty Muralidhar. Statistical Disclosure Control: Moving Forward. *Journal of Official Statistics*, 41(3):820–826, September 2025b. ISSN 0282-423X, 2001-7367. doi: 10.1177/0282423X241312023.

George Duncan and Diane Lambert. The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, 7(2):207–217, 1989. doi: 10.2307/1391438.

G.T. Duncan, S.A. Keller-McNulty, and S.L. Stokes. Disclosure risk vs. data utility: the r-u confidentiality map. Technical report la-ur-01-6428, Los Alamos National Laboratory, 2001. URL http://www.heinz.cmu.edu/research/122full.pdf.

Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107, September 2010. ISSN 2575-8527. doi: 10.29012/jpc.v2i1.585.

Khaled El Emam, Laura Mosquera, and Jason Bass. Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research*, 22(11):e23139, 2020. doi: 10.2196/23139. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7704280/.

Paul Francis and David Wagner. Towards better attribute inference vulnerability measures, 2025. URL https://arxiv.org/abs/2507.01710.

Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. A baseline for attribute disclosure risk in synthetic data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, CODASPY '20, page 133–143, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371070. doi: 10.1145/3374664.3375722.

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sabine Giessing, Rainer Lenz, Jim Longhurst, Eric Nordholt, Kevin Spicer, and Peter-Paul de Wolf. *Statistical Disclosure Control*. John Wiley & Sons, Chichester, UK, 2012a. Foundations; archive attacker model.

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Patrick de Wolf. *Statistical Disclosure Control*. Wiley, 2012b.

Saloni Kwatra and Vicenç Torra. Empirical evaluation of synthetic data created by generative models via attribute inference attack. In Felix Bieker, Silvia de Conca, Nils Gruschka, Meiko Jensen, and Ina Schiering, editors, *Privacy and Identity Management. Sharing in a Digital World*, pages 282–291, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-57978-3.

Jonathan Latner, Marcel Neunhoeffer, and Jörg Drechsler. Buyer Beware: Understanding the trade-off between utility and risk in CART based models using simulation data. In *UNECE Expert Meeting on Statistical Data Confidentiality*, pages 1–12, Barcelona, 2025. United Nations. ISBN 978-92-1-001079-5. doi: 10.18356/9789210010795.

Kidist Amde Mekonnen. Conditioning GAN without training dataset, 2024. URL https://arxiv.org/abs/2405.20687.

Marko Miletic and Murat Sariyar. Challenges of using synthetic data generation methods for tabular microdata. *Applied Sciences*, 14(14), 2024. ISSN 2076-3417. doi: 10.3390/app14145975. URL https://www.mdpi.com/2076-3417/14/14/5975.

Rainer Mühlhoff. Predictive privacy: Towards an applied ethics of data analytics. *Ethics and Information Technology*, 23(4):675–690, December 2021. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-021-09606-x.

K. Muralidhar and S. Ruggles. Escalation of commitment: A case study of the united states census bureau efforts to implement differential privacy for the 2020 decennial census, 2024. URL https://doi.org/10.48550/arXiv.2407.15957. arXiv preprint.

Krishnamurty Muralidhar and Josep Domingo-Ferrer. Database reconstruction is not so easy and is different from reidentification. *Journal of Official Statistics*, 39(3):381–398, 2023. doi: 10.2478/jos-2023-0017.

B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. Promoting an open research culture. *Science*, 348(6242): 1422–1425, 2015. ISSN 0036-8075. doi: 10.1126/science.aab2374. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550299/.

Beata Nowok, G. M. Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. 74(11):1–26, 2016. ISSN 1548-7660. doi: 10.18637/jss.v074.i11. URL https://doi.org/10.18637/jss.v074.i11.

Lisa Pilgram, Fida Kamal Dankar, Jörg Drechsler, Mark Elliot, Josep Domingo-Ferrer, Paul Francis, Murat Kantarcioglu, Linglong Kong, Bradley Malin, Krishnamurty Muralidhar, Puja Myles, Fabian Prasser, Jean Louis Raisaro, Chao Yan, and Khaled El Emam. A consensus privacy metrics framework for synthetic data. *Patterns*, page 101320, 2025. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2025.101320. URL https://www.sciencedirect.com/science/article/pii/S2666389925001680.

M. Platzer and T. Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in Big Data*, 4: 679939, 2021. doi: 10.3389/fdata.2021.679939. URL https://doi.org/10.3389/fdata.2021.679939.

G. M. Raab, Beata Nowok, and Chris Dibben. Practical privacy metrics for synthetic data, 2025. URL http://arxiv.org/abs/2406.16826. Pages: 1-23.

Jerome P. Reiter, Quanli Wang, and Biyuan Zhang. Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6(1), June 2014. ISSN 2575-8527. doi: 10.29012/jpc.v6i1.635.

D.B. Rubin. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data - anonymisation groundhog day. In *USENIX Security Symposium*, 2020. URL https://api.semanticscholar.org/CorpusID:235391080.

Swiss National Science Fund (SNSF). Open Research Data: ein erster Blick auf die aktuelle Praxis. https://data.snf.ch/stories/open-research-data-2023-de.html, 2024. accessed October 15, 2025.

Jennifer Taub, Mark Elliot, Maria Pampaka, and Duncan Smith. Differential correct attribution probability for synthetic data: An exploration. In Josep Domingo-Ferrer and Francisco Montes, editors, *Privacy in Statistical Databases*, pages 122–137, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99771-1.

M. Templ. Providing data with high utility and no disclosure risk for the public and researchers: An evaluation by advanced statistical disclosure risk. *Austrian Journal of Statistics*, 43(4):247–254, 2014. doi: 10.17713/ajs.v43i4.43.

M. Templ, B. Meindl, A. Kowarik, and O. Dupriez. Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software*, 79(10):1–38, 2017. doi: 10.18637/jss.v079.i10. URL http://dx.doi.org/10.18637/jss.v079.i10.

O. Thees, R. Müller, and M. Templ. Beyond the trade-off curve: Multivariate and advanced risk-utility maps for evaluating anonymized and synthetic data. *Arxiv*, 2025.

Oscar Thees, Jiří Novák, and Matthias Templ. Evaluation of synthetic data generators on complex tabular data. In Josep Domingo-Ferrer and Melek Önen, editors, *Privacy in Statistical Databases*, pages 194–209, Cham, 2024. Springer Nature Switzerland.

J. Ward, Y. Yang, C. Wang, and G. Cheng. Ensembling membership inference attacks against tabular generative models, 2025. URL https://doi.org/10.1145/3733799.3762977.

Leon Willenborg and Ton de Waal. *Elements of statistical disclosure control*, volume 155 of *Lecture Notes in Statistics*. Springer, 1 edition, 2001. ISBN 978-1-4613-0121-9. doi: 10.1007/978-1-4613-0121-9. URL http://link.springer.com/10.1007/978-1-4613-0121-9.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

A Simulation Data Generation

We simulate n independent microdata records with six variables: gender (G), age (A), education (E), income (I), health score (H), and disease status (D). The design encodes realistic, policy-relevant dependencies through a latent socioeconomic status (SES) variable while remaining transparent and tunable via a global dependency parameter $\kappa \geq 0$. Throughout, $\mathsf{TN}(\mu, \sigma^2; [a, b])$ denotes a normal distribution truncated to $[a, b]$, and we use standardized predictors $\tilde{X} = (X - \mu_X)/\sigma_X$ when indicated.

Dependency mechanism. All variable dependencies flow through a shared latent variable:

$$\text{SES}_i \sim \mathcal{N}(0, 1),$$

representing unobserved socioeconomic status. The strength of dependencies is controlled by signal and noise weights derived from κ:

$$w_{\text{signal}} = \sqrt{\frac{\kappa}{1+\kappa}}, \qquad w_{\text{noise}} = \sqrt{\frac{1}{1+\kappa}}.$$

At $\kappa = 0$, relationships are driven purely by noise ($w_{\text{signal}} = 0$, $w_{\text{noise}} = 1$). As $\kappa \to \infty$, dependencies become deterministic ($w_{\text{signal}} \to 1$, $w_{\text{noise}} \to 0$). At the default $\kappa = 1$, signal and noise contribute equally ($w_{\text{signal}} = w_{\text{noise}} = 1/\sqrt{2}$), yielding approximately 50% explained variance.

Age. We draw age from a truncated normal to reflect adult populations:

$$A_i \sim \mathsf{TN}(\mu_A, \sigma_A^2; [a_{\min}, a_{\max}]), \quad \text{defaults: } \mu_A = 45,\ \sigma_A = 12,\ a_{\min} = 18,\ a_{\max} = 85.$$

Education. Education is an ordinal categorical variable with three levels $\{0, 1, 2\}$ corresponding to *low*, *medium*, and *high* attainment. We generate E_i from a latent continuous variable that depends on both SES and age:

$$L_i = w_{\text{signal}} \cdot (0.8 \cdot \text{SES}_i - 0.4 \cdot \tilde{A}_i) + w_{\text{noise}} \cdot \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1),$$

$$E_i = \begin{cases} 0 & \text{if } L_i < c_1, \\ 1 & \text{if } c_1 \le L_i < c_2, \\ 2 & \text{if } L_i \ge c_2, \end{cases} \quad c_1 = -0.3,\ c_2 = 0.7.$$

The negative age coefficient reflects the empirical pattern of younger cohorts having higher educational attainment.

Income. We generate log-income using a linear model:

$$\log I_i^\star = w_{\text{signal}} \cdot (0.5 \cdot \text{SES}_i + 0.3 \cdot \tilde{A}_i + 0.25 \cdot E_i) + w_{\text{noise}} \cdot \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1),$$

$$I_i = \exp(10 + \log I_i^\star).$$

The constant 10 centers income around realistic values (approximately \$20,000–\$40,000).

Health score. H is a continuous health measure on $[0, 100]$ created via sigmoid transformation of a linear predictor:

$$H_i^\star = w_{\text{signal}} \cdot \big(0.6 \cdot \text{SES}_i - 0.5 \cdot \tilde{A}_i + 0.2 \cdot E_i + 0.2 \cdot \widetilde{\log I_i}\big) + w_{\text{noise}} \cdot \zeta_i, \quad \zeta_i \sim \mathcal{N}(0, 1),$$

$$H_i = \frac{100}{1 + \exp(-H_i^\star)}.$$

The negative age coefficient reflects declining health with age, while positive SES, education, and income coefficients represent protective effects.

Disease status. $D_i \in \{\texttt{healthy}, \texttt{diabetic}, \texttt{hypertensive}\}$ is sampled via a multinomial logit with `healthy` as the baseline. Unlike other variables, disease dependencies scale *linearly* with κ rather than through signal/noise weights:

$$\log \frac{\Pr(D_i = \texttt{diabetic})}{\Pr(D_i = \texttt{healthy})} = -1.5 + \kappa \cdot \big(0.8 \cdot \tilde{A}_i - 0.3 \cdot \widetilde{\log I_i} - 0.2 \cdot E_i\big),$$

$$\log \frac{\Pr(D_i = \texttt{hypertensive})}{\Pr(D_i = \texttt{healthy})} = -1.3 + \kappa \cdot \big(1.0 \cdot \tilde{A}_i - 0.2 \cdot \widetilde{\log I_i} - 0.1 \cdot E_i\big).$$

This linear scaling creates stronger dependency effects at high κ: older age raises risk, while higher income and education are mildly protective. The fixed intercepts (-1.5 and -1.3) induce a baseline class imbalance favoring `healthy` outcomes.

Gender. Gender is binary with mild dependency on SES, age, and education:

$$\eta_i = w_{\text{signal}} \cdot (0.3 \cdot \text{SES}_i - 0.2 \cdot \tilde{A}_i + 0.2 \cdot E_i),$$

$$G_i \sim \text{Bernoulli}\big(\text{logit}^{-1}(\eta_i)\big), \quad G_i \in \{\texttt{female}, \texttt{male}\}.$$

Controlling dependence strength. The global parameter $\kappa \ge 0$ controls the strength of all dependencies:

- $\kappa = 0$: Variables retain weak dependencies due to fixed intercepts in the disease model, but signal contributions vanish ($w_{\text{signal}} = 0$).
- $\kappa = 1$ (default): Balanced signal-to-noise ratio, yielding moderate dependencies.
- $\kappa \gg 1$: Near-deterministic relationships as $w_{\text{signal}} \to 1$ and disease coefficients grow large.

Because disease logits scale linearly with κ while other variables use the signal-to-noise transformation, disease dependencies strengthen more rapidly at high κ.

Defaults and realism. With default $\kappa = 1$, the simulation produces realistic marginal distributions and moderate dependencies: income rises with SES, age, and education; health declines with age but improves with socioeconomic status; and the probabilities of `diabetic` and `hypertensive` increase with age and decrease with income and education. These defaults can be adapted to domain-specific baselines by adjusting variable-specific parameters without changing κ.