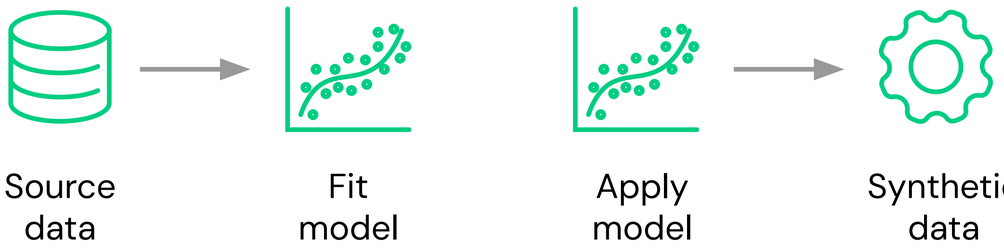
A Record-Level Risk Metric for Partially Synthetic Clinical Data Using Inverse-Square Similarity Decay

Ade Adeoye, Lucy Mosquera
Aetion, a Datavant Company

BACKGROUND

DEFINITIONS

Synthetic data is generated by training models on real data, but is not real data. It has the same patterns and statistical properties as real data. For certain use cases it can act as a proxy for real data.

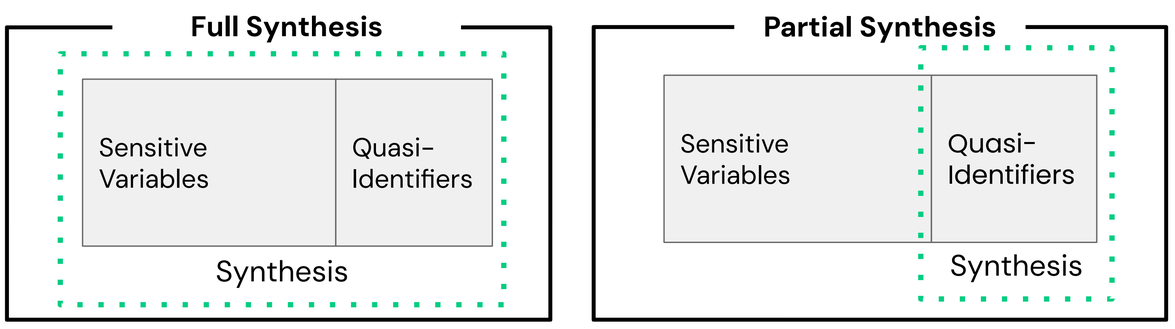


Direct Identifiers are pieces of information that can uniquely pinpoint a single individual (e.g. social security number, healthcare number, etc.)

Quasi-Identifiers (QIs) are pieces of information that do not uniquely identify an individual alone, but can, when combined with other QIs (e.g. age, sex, etc.)

Sensitive variables are pieces of information that may not be identifying but can cause significant harm to an individual if disclosed (e.g. HIV status, etc.)

Partial synthetic clinical trial data synthesizes only quasi-identifiers (QIs), such as age, sex, and country, to mitigate re-identification risk while preserving data utility. Synthesis of complex clinical trial data benefits from partial synthesis, as it allows the maintenance of complex longitudinal relationships with smaller patient counts.



Necessary for **maximum privacy** or when amplifying **smaller datasets**

Sometimes used **when the data is complex** or **for small datasets** with very large number of variables

PROBLEM

Partially synthesized datasets may carry a re-identification risk as they contain some real patient information. A quantitative assessment of re-identification risks should be conducted to prove a dataset is de-identified.

Traditional privacy approaches like *k*-anonymity or l-diversity provide group-level guarantee, but fail to quantify record-level risk, particularly when synthetic records closely mimic real counterparts.

This study proposes an interpretable record-level risk metric which leverages pairwise record comparison, incorporates structured QI similarity, and models decline in attacker confidence and increasing ambiguity using intuitive components such as distance, tie counts, and attack probability.

IMPACT

Accurate assessment of re-identification risk is important for the responsible sharing and secondary use of data. As the adoption of (partially) synthetic data grows for varying uses from software testing to augmentation of underrepresented subgroups, it becomes important to ensure that patient privacy remains protected.

A granular privacy risk metric designed to qualify partially synthetic data enhances trust, supports compliance with data-sharing standards, and enables broader and safer use of clinical trial data for innovation and research.

OBJECTIVES

To develop and demonstrate a record-level risk metric that integrates:

- Structural similarity between real and synthetic records via pairwise distance,
- Ambiguity through tie counts (number of equally close synthetic matches),
- Adversarial intent modeled by probability of attack, while applying an inverse-square similarity decay to capture the nonlinear decline in identifiability as ambiguity increases.

METHODS

- This novel risk estimator evaluates the re-identification risk of partially synthetic data relative to a real dataset using a similarity metric derived from pairwise distances and record ties, equivalence class sizes, and a probability of attack.
- A $N \times N$ distance matrix over a select number of QIs produces a scaled distance, d_i^{scaled}, with record ties, T_i, quantifying the uncertainty an adversary will encounter counting the number of synthetic records that share the same distance to a real record. This results in a similarity metric, sim_i ,given as:

$$\text{sim}_i = (1 - d_i^{\text{scaled}}) \cdot \frac{1}{T_i^2}$$

  where $\frac{1}{T_i^2}$ reflects decreasing attacker confidence as ambiguity increases
- The record-level risk score is calculated by combining similarity with the probability of attack, $P(A)$, and the inverse of the real data equivalence class size, f_i, ensuring that rare, highly similar records receive a high-risk score.

$$\text{Average Risk} = P(A) \cdot \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{f_i} (1 - d_i^{\text{scaled}}) \cdot \frac{1}{T_i^2} \right]$$

- Overall dataset risk is summarized using maximum record-level risk and compared against the European Medicine Agency and Health Canada acceptable re-identification risk threshold of 0.09 to determine if the synthetic dataset can be considered de-identified.
- This risk metric captures how closely a synthetic record matches a real one, how difficult it is for an attacker to tell exactly which synthetic record is this true match, and how unique that real record is in the original data, weighted by the chance that an attacker would attempt re-identification.

RESULTS

The proposed record-level metric successfully quantified privacy risk in partially synthetic data derived from 16 phase III trials in diabetes. Consistent with theory, results showed that risk decreased nonlinearly with increasing ambiguity and increased for highly similar, rare records.

Records with low scaled distances (<~0.1-0.2) and unique synthetic matches (T=1) exhibited the highest individual risk, particularly when they belong to small equivalence classes. Records with multiple equally close synthetic equivalents (high T) or greater QI variability demonstrated substantially reduced risk.

Risk increases as the number of QIs used in an attack increases. When considering a reasonable number of QIs (e.g. 7), maximum record-level risk remained below the acceptable risk threshold of 0.09 for most QI combinations in all trial datasets.

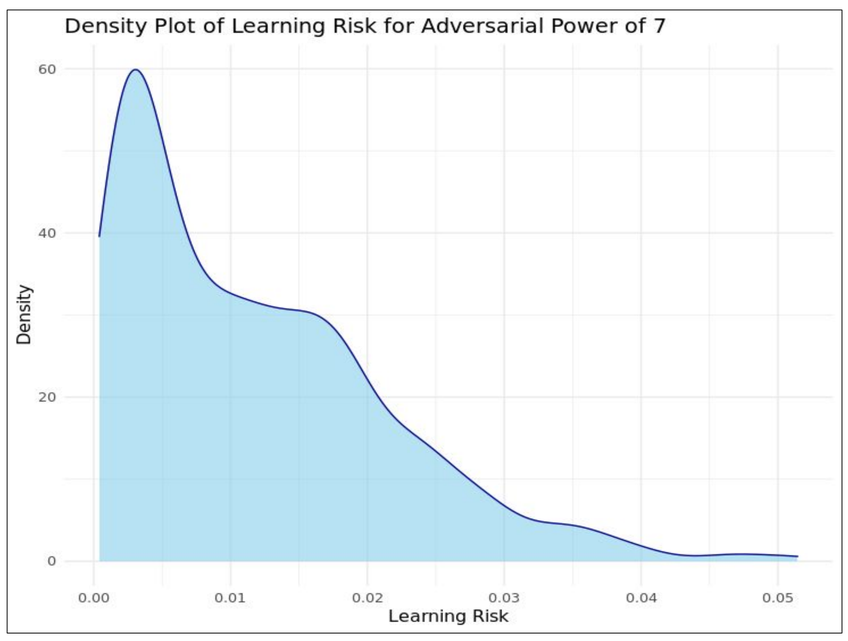


Figure 1. DENSITY plot of learning risk for partially synthetic data using n(QI)=7

CONCLUSIONS

- Demonstrated a record-level risk metric that integrates structural similarity between real and synthetic data, ambiguity through tie counts and adversarial intent.
- Incorporated an inverse-square similarity decay to represent the nonlinear reduction in identifiability as ambiguity increases.
- Generated per-record risk scores, providing a more granular and interpretable assessment of disclosure risk than traditional metrics like k-anonymity.
- The method provides a scalable and empirically grounded tool to support regulatory and ethical confidence in (partially) synthetic data sharing.
- Validation on 16 trial datasets demonstrated that the proposed metric provides a sensitive yet balanced risk indicator at the record level, identifying potentially vulnerable records while maintaining low overall disclosure risk for the dataset.

REFERENCES

Charu C. Aggarwal. 2005. On k-anonymity and the curse of dimensionality. In Proceedings of the 31st international conference on Very large data bases (VLDB '05).

El Emam K et. al. 2020. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation J Med Internet Res

AETION® | a Datavant Company