JAMIA Open, 2026, **9**(2), ooag066
https://doi.org/10.1093/jamiaopen/ooag066
Research and Applications

Aim high, stay private: differentially private synthetic data enables public release of behavioral health information with high utility

Mohsen Ghasemizade, MS[1,*], Juniper Lovato, PhD[1], Chris Danforth, PhD[2], Peter Sheridan Dodds, PhD[1], Laura S. P. Bloomfield, PhD[3], Matthew Price, PhD[4], Joseph Near, PhD[1]

[1]Department of Computer Science, University of Vermont, VT, 05405, United States
[2]Department of Mathematics and Statistics, University of Vermont, VT, 05405, United States
[3]The Gund Institute for the Environment, University of Vermont, VT, 05405, United States
[4]Department of Psychological Science, University of Vermont, VT, 05405, United States
*Corresponding author. Mohsen Ghasemizade, MS, Department of Computer Science, University of Vermont, VT, 05405, United States (mghasemi@uvm.edu)

Abstract

Objective: **Sharing behavioral health and wearable data poses privacy challenges, as traditional de-identification remains vulnerable to re-identification. Differential privacy (DP) provides mathematical guarantees through a tunable privacy budget, ε. This study evaluates the feasibility of generating and releasing DP synthetic behavioral health data with high analytical utility, identifying practical ε values for public data sharing.**

Materials and methods: **We analyzed physiological data from wearable devices and self-reported data from Phase 1 of the Lived Experiences Measured Using Rings Study (LEMURS), which tracked sleep, stress, and well-being among first-year college students. Three DP synthetic data generators: AIM, MST, and PATECTGAN, were evaluated across privacy budgets ranging from $\varepsilon = 1$ to 100. Utility was assessed using L1/L2 errors, correlation, regression, UMAP, and assessed vulnerability via privacy attacks.**

Results: **AIM outperformed MST and PATECTGAN in preserving both statistical and analytical properties of the original data. For the Survey dataset, the lowest marginal errors occurred at $\varepsilon = 5$ and 10. Correlation, regression, and UMAP analyses confirmed that AIM-generated data closely replicated original relationships at moderate ε values.**

Discussion: **Choice of privacy budget is still an open question, and it is task-agnostic and dataset-specific. Moderate privacy budgets ($5 \leq \varepsilon \leq 10$) maintained key associations between physiological and psychological measures while ensuring privacy. AIM's workload-aware design effectively allocated noise toward relevant features, enhancing performance.**

Conclusion: **A privacy budget of $\varepsilon = 5$ offers a practical balance between data utility and participant privacy for LEMURS behavioral health data sharing.**

Key words: health behavioral study, differential privacy, wearable devices, utility-privacy tradeoff

Lay Summary

The central finding of this work is that we can safely share behavioral health data by generating differentially private synthetic datasets that protect participants' privacy while preserving scientific value. Participating in studies that gather biometric and survey measures often exposes people to re-identification risk, even when identifiers are removed, unusual combinations of traits can single someone out. That risk has forced many researchers to keep valuable datasets locked away.

This study tested a mathematical privacy technique called differential privacy to create synthetic versions of the health data. By adding controlled random noise into the data so that any individual's contribution becomes indistinguishable, while the underlying relationships we care about (e.g., how disrupted sleep tracks with elevated stress) remain intact.

Received: November 26, 2025. **Revised:** March 4, 2026. **Accepted:** April 6, 2026

This requires a balance between privacy guarantees and empirical fidelity, and that balance is rarely tested with real behavioral health data.
Using records from 600 students, we stress-tested multiple privacy levels and evaluated both scientific utility and vulnerability to privacy attacks. Stress prediction models worked nearly as well on synthetic data as on original data, and privacy attacks failed to re-identify participants. Our framework provides practical guidance for institutions deciding whether their health datasets can be safely released for scientific research.

Introduction

Behavioral health and wearable datasets offer valuable fine-grained views into how physiology, stress, and daily experience coevolve, yet they are among the most difficult data to share responsibly. Traditional de-identification has repeatedly proven insufficient, leaving study participants vulnerable to re-identification. Our aim in this work is to examine whether differentially private synthetic data can provide a practical solution. We show that, for this use case, moderate privacy budgets can sustain key behavioral and physiological relationships while resisting linkage and membership attacks, pointing toward a feasible and reproducible framework for public release of complex health data.

The Lived Experiences Measured Using Rings Study (LEMURS)[1] recruited roughly 600 college students to explore the connections between well-being, health, heart activity, and sleep, employing Oura rings and surveys. This study generated a dataset with more than 100 attributes related to the participants' daily routines and demographics. While participants are assigned anonymous numerical IDs within the dataset, there is concern that certain column combinations could disclose the identities of some participants, particularly those in minority groups, such as transgender individuals and people of color.

Due to privacy concerns, datasets like LEMURS are typically not available to the public. Traditional de-identification methods have proven vulnerable,[2,3] with well-known re-identification failures in datasets.[4,5,6] These concerns highlight the need for robust privacy-preserving techniques for wearable health data.[7] Differential Privacy (DP)[8,9] can protect against unforeseen vulnerabilities.

In this work, we demonstrate the feasibility of generating and releasing high-quality DP synthetic data for the LEMURS study, based on existing task-specific non-DP research on the dataset. We suggest this generalizable framework for data custodians:

- Defining real-world use cases and baseline hypotheses;
- Selecting task-agnostic and task-specific (epistemic parity) utility metrics;
- Generating synthetic data across a range of ε values and synthetic data generators;
- Evaluating the utility-privacy tradeoff against attacks;
- Selecting a practical ε range based on the task;
- Ensuring reproducibility of methodology and availability of data, available on our GitHub repo.[1]

[1] https://github.com/mghasemizade/lemurs_dp

Background and significance

LEMURS dataset

LEMURS began in the fall of 2022 and is ongoing at the University of Vermont.[10] To date, over 600 first-year college students have participated in this longitudinal experiment, which assesses changes in students' sleep, stress, mental health, and other outcomes through a series of weekly surveys.

Additionally, researchers used Oura rings to collect Total Sleep Time (TST), resting heart rate, heart rate variability, sleep stages, body temperature, and other physiological data from the participants. The data is longitudinal, where each row corresponds to a single weekly entry for a participant. Each participant, identified by a consistent but pseudo-anonymized numerical ID, contributes multiple rows to the dataset.

In the LEMURs project, researchers aimed to predict stress levels or mental health deterioration before symptoms become severe, using sleep patterns and wearable-derived metrics to detect early signs of risk. Prior work on this dataset has successfully predicted stress and anxiety using sleep and physiological metrics.[11–13]

Our work uses two datasets collected in phase 1 of the LEMURS study. The first dataset contains 108 columns with a mix of qualitative and quantitative features. The second dataset published via,[12] is more compact, consisting of 19 purely quantitative columns containing physiological measurements from the Oura ring and self-reported Perceived Stress Scale score (PSS). For simplicity, we refer to the larger dataset as the Survey dataset and the smaller one, composed almost entirely of Oura-based measurements, as the Oura dataset.

Differential privacy

DP[8,9] is a widely adopted framework that safeguards the privacy of individuals within a dataset. DP introduces controlled random noise scaled by privacy budget ε, ensuring outputs cannot fluctuate by more than e^{ε}. Formally, a mechanism $\mathcal{M}$ satisfies DP if for any pair of neighboring datasets D_1 and D_2 and any possible set of outcomes S:

$$\frac{\Pr[\mathcal{M}(D_1) \in S]}{\Pr[\mathcal{M}(D_2) \in S]} \leq e^{\varepsilon} \tag{1}$$

Intuitively, we can think of $\Pr[\mathcal{M}(D_1)]$ as the probability of outcome S when an individual contributes their data, while $\Pr[\mathcal{M}(D_2)]$ denotes the probability of outcome S when the individual does not contribute their data to the dataset. The DP definition bounds the ratio of these probabilities, ensuring that they must be similar—with ε controlling the degree of similarity.

Smaller values of ε provide better privacy but generally result in less utility. Values of ε in the single digits are preferred for strong privacy.

DP effectively prevents the re-identification of individuals within the dataset and has found extensive applications in medical settings to protect individual privacy.[14–17]

Methodology

Linkage attack

A linkage attack is a privacy attack in which an adversary uses external data sources, such as public records or survey data, to re-identify individuals in an anonymized dataset. Even when direct identifiers (names or email addresses) are removed, quasi-identifiers (age, sleep duration, or stress scores) can still be used to match individuals across datasets and uncover sensitive information.

Assuming the Oura dataset from Bloomfield et al.[12] as the target dataset, and the Survey dataset as an auxiliary dataset available to an adversary, a linkage attack becomes feasible. Since both datasets contain overlapping attributes, such as TST and PSS, an attacker could compare those features and successfully re-identify individuals in the Oura dataset.

Membership inference attack via closest distance

To evaluate the privacy of our DP synthetic data, we conducted a membership inference attack via closest distance from the TAPAS framework,[18] and infused the AIM generator in it. Our goal is to test whether an attacker, given only the differentially private synthetic data, can determine whether a particular individual was included in the original training set.

The attacker knows the target's characteristic values and holds a 50% random sample of the real data for calibration. They repeatedly ask the AIM mechanism for synthetic draws of the same size, and in each draw, they find the single synthetic record closest to the target (using either Hamming or Euclidean distance). Intuitively, if the target was part of the training data, its synthetic twin will tend to appear closer in the release. By collecting these "nearest neighbors" distances from draws where the target was included versus excluded, the attacker builds a Receiver Operating Characteristic (ROC) curve and selects the distance threshold that best separates "in" from "out". Finally, with that threshold fixed, a fresh synthetic draw is used to decide membership: if the nearest neighbor is within the threshold, the attacker guesses "in" otherwise "out".

Utility evaluation methods

Recently, Rosenblatt et al.[19] introduced the concept of "epistemic parity" as a novel lens for assessing the efficacy of DP synthetic datasets. Instead of solely focusing on statistical metrics such as marginal L1 and L2 errors, or correlation preservation, they contend that the utility of synthetic data should be gauged by whether the scientific conclusions derived from the original data remain reproducible after applying DP.

We generated 10 versions of the dataset for each algorithm (AIM, MST, PATECTGAN) for each privacy budget ($\varepsilon = [1, 2, 5, 10, 20, 50, 100]$) for the Survey and Oura datasets. For each privacy budget, we generated 10 synthetic datasets to account for algorithmic randomness, resulting in a total of 210 datasets.

The algorithms employed in our generators are limited only to quantitative values because DP introduces numerical noise. So, qualitative columns, such as text-based survey responses, have been removed.

Regression models

We replicated the mixed-regression model by Bloomfield et al.[12] that predicts PSS on the original data and confirmed that we recovered their published coefficients. We then applied the same model to each synthetic dataset, using replication fidelity as a case-study metric.

In addition to this replication, we developed a second model using random forest regression to predict stress levels in the larger Survey dataset. We selected the top 12 stress-related features based on Spearman correlations, including anxiety and sleep measures. We chose a random forest model because of its ability to capture non-linear relationships and its robustness to noise. To assess model performance, we used the R^2 score, which ranges from 0 (no predictive power beyond the mean) to 1 (perfect prediction). Higher R^2 values indicate stronger preservation of predictive utility in the synthetic datasets.

Umap

Uniform Manifold Approximation and Projection (UMAP)[20] is a nonlinear dimensionality reduction technique that enables visualization of high-dimensional data in lower dimensions. By projecting these datasets into a 2D space, we can visually assess the extent of structural distortion introduced by DP. UMAP also uncovers latent cluster structures, enabling us to examine whether individuals with similar attributes, such as comparable PSS, remain grouped together after applying DP. This unsupervised clustering approach is particularly useful for evaluating latent patterns, similar to how Ghasemizade et al.[21] leveraged UMAP to uncover hierarchical groupings in complex, belief-driven datasets.

Results and analysis

Linkage attack

To demonstrate the vulnerability of the original Oura dataset, which was only de-identified, we assessed its similarity to the Survey dataset. With the two datasets in hand and knowing the matching columns "week" and "PSS", we looked for rows that exactly match in those columns. To make the attack robust, we added a third condition: the similarity score between the reported sleep time from the Survey dataset and the measured sleep time from the Oura ring. Since they are not expected to match exactly, we allow some error between the two values in each dataset.

Using Record Linkage Toolkit,[22] we compared "TST" values with ± 0.5 hour tolerance, average sleep time (origin) of 7, with a scale of 1.5 beyond this threshold.

We configure this function with an offset of 0.5, meaning that if the difference in sleep hours between the two datasets is within ± 0.5 hours, it is considered a perfect match (similarity score = 1.0). We assumed the average sleep time for college grad students is 7, and *origin* = 7 reflects that. Beyond this threshold, the similarity score decreases linearly, with a scale parameter of 1.5 that controls the rate of decay. This configuration allows for partial credit when sleep values differ by more than 0.5 hours but still remain close.

Using the configuration shown in the Appendix, available as supplementary data at [*JAMIA Open*] online, we identified four matching rows between the original Oura and Survey datasets based on exact matches for week and perceived stress scores, along with a high similarity threshold for reported sleep duration. These matches demonstrate a successful linkage attack under relatively simple assumptions. Four out of 100 is still a small ratio, but the primary objective is to protect the privacy of every individual, even a single person, in the dataset. If the auxiliary dataset had been released under DP, the added noise would have prevented such confident re-identification, as DP ensures that no individual record can be reliably linked, even when some attributes overlap across datasets.[23–25]

Synthetic data generators and L1, L2 errors

L1 and L2 errors measure how closely the synthetic data matches the real data's distribution over all marginals, computed by comparing each attribute-pair distribution and then averaging across all pairs. L1 captures the average absolute discrepancy and L2 captures the average squared/Euclidean discrepancy, so lower values indicate higher fidelity to the original pairwise structure.

To determine the most effective synthetic data generator for this task, we evaluate the L1 and L2 errors for all ε values and selected algorithms across the generated synthetic datasets. We then select the algorithm with the lowest errors for further analysis.

Based on benchmarks by Rosenblatt et al.,[19] we selected three candidates: AIM,[26] MST,[27] and PATECTGAN.[28] For our experiments, we used the default and most common configurations for Aim, MST, and PATECTGAN. A notable hyperparameter of AIM is its tunable preprocessing privacy budget, which allows analyzing the dataset and its marginals before generation. Based on the argument of AIM's creators,[26] we provide only 10% of the privacy budget for the preprocessing and selecting the marginals step, and the rest is used for the measurement step.

Table 1 reports the mean of L1 and L2 errors along with their standard deviation for all ε values across AIM, MST, and PATECTGAN, on 2-way marginals. AIM generally produces lower reconstruction errors compared to MST and PATECTGAN. AIM achieves the smallest errors at $\varepsilon = 5 - 10$ for Survey data. PATECTGAN performs better on lower-dimensional Oura data. Given the Survey dataset's higher dimensionality and the goal of releasing a utility-preserving synthetic version, AIM offers the most reliable balance between privacy and marginal errors. Therefore, we select AIM for the remaining analyses of this paper.

Note that Utility is not monotone in ε for synthetic data pipelines. While DP noise scale decreases with ε the final L1/L2 reflects a combination of DP noise, model/approximation error, and sampling/training randomness; once DP noise is small, the remaining terms can cause small up or down rather than a clean decrease.

Table 1. L1 and L2 errors for Oura and Survey datasets across different ε values.

Oura L1	AIM	MST	PG
$\varepsilon = 1$	1.81±2e-3	1.83±1e-3	1.90±2e-3
$\varepsilon = 2$	1.81±2e-3	1.81±7e-4	1.87±3e-3
$\varepsilon = 5$	1.80±5e-4	1.80±5e-4	1.83±7e-4
$\varepsilon = 10$	1.80±2e-3	1.80±3e-4	1.85±3e-4
$\varepsilon = 20$	1.80±3e-3	1.79±3e-3	1.86±8e-4
$\varepsilon = 50$	1.80±2e-3	1.79±2e-4	1.85±9e-4
$\varepsilon = 100$	1.80±6e-4	1.80±3e-4	1.86±8e-4
Oura L2	**AIM**	**MST**	**PG**
$\varepsilon = 1$	0.30±8e-4	0.29±1e-2	0.06±5e-5
$\varepsilon = 2$	0.30±9e-3	0.28±1e-2	0.05±1e-3
$\varepsilon = 5$	0.26±7e-4	0.26±6e-4	0.04±3e-4
$\varepsilon = 10$	0.26±2e-3	0.27±3e-4	0.05±2e-4
$\varepsilon = 20$	0.27±3e-3	0.27±3e-4	0.05±3e-4
$\varepsilon = 50$	0.28±5e-4	0.29±2e-4	0.05±3e-4
$\varepsilon = 100$	0.30±6e-4	0.30±2e-4	0.05±3e-4
Survey L1	**AIM**	**MST**	**PG**
$\varepsilon = 1$	0.50±1e-2	0.50±5e-3	1.01±2e-4
$\varepsilon = 2$	0.43±5e-3	0.45±6e-3	0.79±6e-4
$\varepsilon = 5$	0.38±1e-3	0.39±1e-3	0.58±1e-3
$\varepsilon = 10$	0.39±3e-3	0.40±4e-3	0.63±6e-3
$\varepsilon = 20$	0.40±8e-4	0.41±1e-3	0.60±2e-2
$\varepsilon = 50$	0.40±1e-4	0.41±9e-4	0.60±1e-4
$\varepsilon = 100$	0.41±6e-4	0.41±4e-4	0.63±2e-4
Survey L2	**AIM**	**MST**	**PG**
$\varepsilon = 1$	0.15±4e-3	0.15±3e-3	0.32±1e-3
$\varepsilon = 2$	0.13±3e-3	0.13±3e-3	0.22±3e-3
$\varepsilon = 5$	0.10±1e-3	0.11±6e-4	0.13±8e-4
$\varepsilon = 10$	0.10±9e-4	0.10±7e-4	0.16±3e-4
$\varepsilon = 20$	0.10±4e-4	0.10±8e-4	0.14±8e-3
$\varepsilon = 50$	0.11±1e-3	0.11±7e-4	0.14±7e-4
$\varepsilon = 100$	0.12±6e-4	0.12±1e-3	0.16±8e-3

Values are reported as mean ± standard deviation across 10 generated datasets for each ε value.

Regression models

For each dataset, we created an original regression model without noise and 60 distinct regression models for each AIM DP-generated synthetic dataset, designed for each epsilon.

Oura dataset

This model uses a centered version of TST by subtracting each participant's mean sleep duration (tst_dev) to account for within-person variation. We then fit a linear mixed-effects model

predicting PSS as a function of week, gender, and within-participant TST deviation, with participant ID as a random effect and week included in the random slope. This model serves as a reference to evaluate how well the synthetic datasets preserve these key associations. The mean coefficients for "week" and "tst_dev" with their standard deviations for varying epsilons are outlined in Table 2. In the original model, the coefficient for "week" is -0.33, indicating that perceived stress scores tend to decrease slightly as the semester progresses, while the coefficient for "tst_dev" is -0.89, suggesting that students who sleep less than their personal average report significantly higher stress levels.

When comparing the mean and standard deviation of the regression coefficients across privacy budget levels, we observe the trend converging to the original model coefficients from $\varepsilon = 10$ for "tst_dev" and $\varepsilon = 20$ for "week". The "tst_dev" stabilizes by $\varepsilon \geq 10$ but remains attenuated relative to the original; week shows weaker recovery and higher variability at $\varepsilon = 10$. The coefficients reach the closest distance to the original at the $\varepsilon = 100$.

Survey dataset

Table 3 shows that the R^2 score of the random forest model trained on the original Survey dataset is 0.71. The R^2 scores for the synthetic datasets follow a peaking trend to reach the original value starting from the beginning, and there are no surprises in the trend. Starting from $\varepsilon = 5$ we can say it reaches a stable plateau, making it a good candidate to select as our privacy budget. We also analyzed the permutation feature importance of R^2 along each dataset, that can be found in the appendix, available as supplementary data at [*JAMIA Open*] online.

Table 2. Comparison of the mean ± standard deviation of "week" and "tst_dev" coefficients across ε values for AIM.

ε	week	tst_dev
Original	−0.33	−0.89
1	−0.06 ± 0.14	−0.01 ± 0.23
2	−0.04 ± 0.11	−0.11 ± 0.22
5	−0.15 ± 0.06	−0.33 ± 0.25
10	−0.10 ± 0.12	−0.56 ± 0.29
20	−0.30 ± 0.07	−0.59 ± 0.23
50	−0.44 ± 0.25	−0.59 ± 0.14
100	−0.33 ± 0.15	−0.60 ± 0.17

Table 3. Comparison of the mean ± standard deviation for R^2 scores on random forest regression across ε values for AIM.

ε	R^2 Score
Original	0.71
1	0.32 ± 0.15
2	0.58 ± 0.05
5	0.64 ± 0.02
10	0.68 ± 0.02
20	0.68 ± 0.02
50	0.71 ± 0.02
100	0.72 ± 0.02

Spearman correlation

We generated Spearman correlation heatmaps for both the original and synthetic versions of the Oura and Survey datasets, for each privacy budget, we picked the dataset with lowest L1 and L2 errors. These heatmaps allow us to assess whether key correlation structures are preserved after applying DP.

Figures 1 and 2 present Spearman correlation heatmaps for the original datasets and their DP synthetic versions across various ε values. At low privacy levels ($\varepsilon = 1-2$), correlation

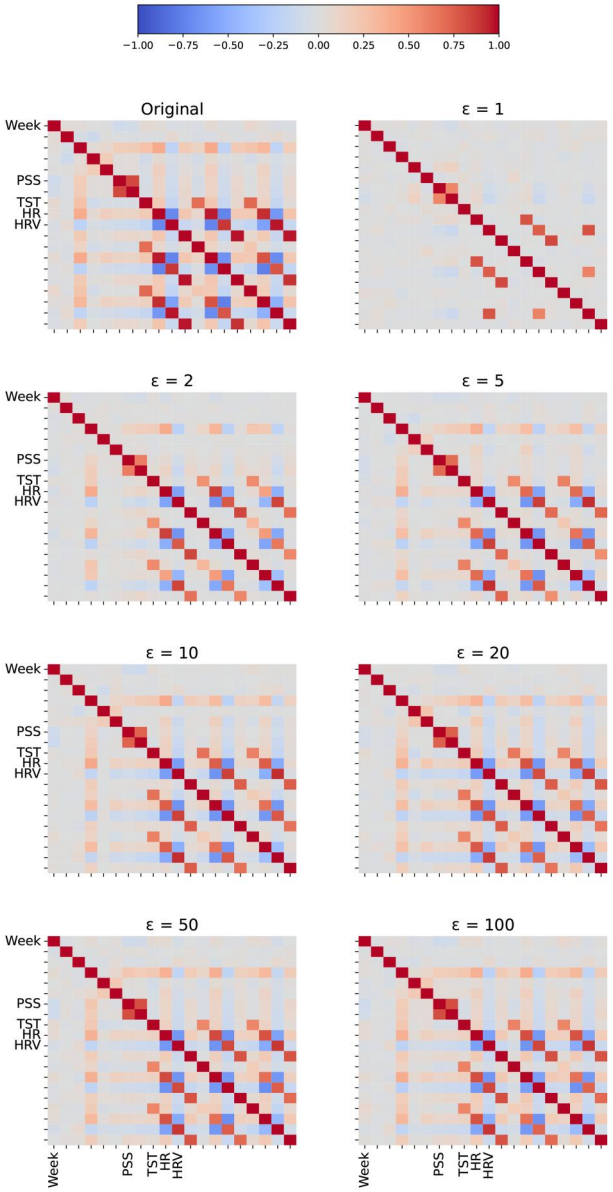


Figure 1. Spearman correlation heatmaps for the original and synthetic Oura datasets across various ε levels for AIM. Only selected variable labels are shown to improve readability. HR: Heart Rate; PSS: Perceived Stress Score; TST: Total Sleep Time; HRV: Heart Rate Variability.

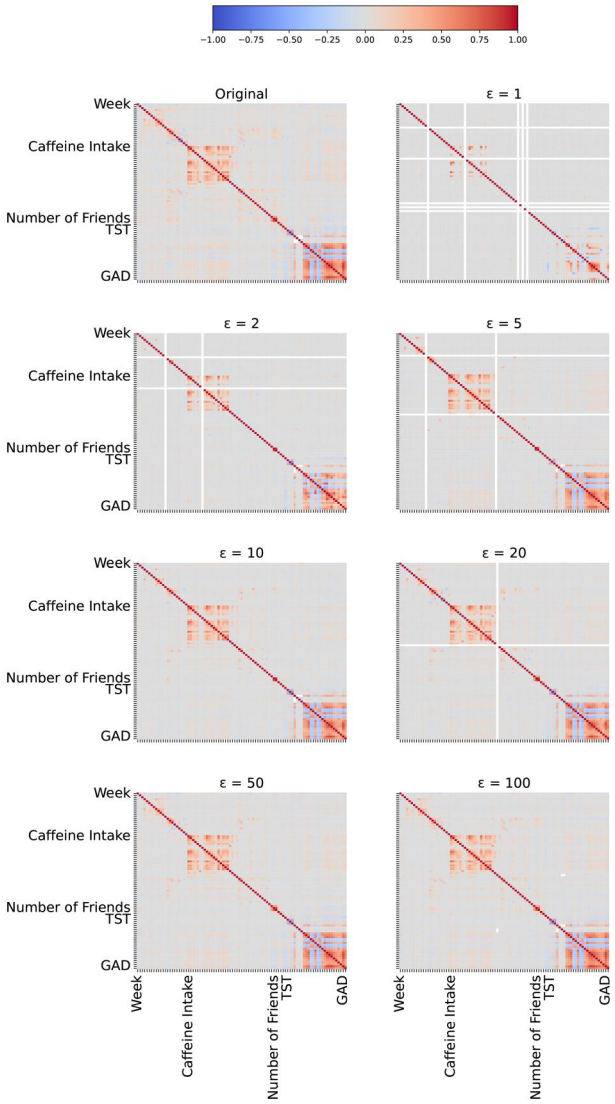


Figure 2. Spearman correlation heatmaps for the original and synthetic Survey datasets at various ε values for AIM. All variables are included, but only selected axes labels are shown for clarity: Week, PSS: Perceived Stress Score; GAD: Generalized Anxiety Disorder; TST: Total Sleep Time, caffeine intake, and number of friends.

patterns degrade, but from $\varepsilon = 5$ onward, the overall structure remains consistent with the original data.

For the Survey dataset, the bottom-right cluster of variables, which includes PSS (Perceived Stress Score), GAD (Generalized Anxiety Disorder), and stress-related features, shows consistent structures starting from $\varepsilon = 5$. Among these, the heatmaps for $\varepsilon = 5$ and $\varepsilon = 10$ most closely resemble the original, suggesting that these privacy levels strike an effective balance between privacy protection and statistical utility for this dataset.

UMAP

UMAP hyperparameters were selected through a systematic grid search to produce interpretable visualizations that reflect both the continuity of PSS and the topological preservation of structure across synthetic datasets. UMAP is particularly effective on high-dimensional data, such as our 108-column Survey dataset. For each privacy budget, we performed UMAP only on the dataset with the lowest L1 and L2 errors.

Oura dataset

To observe the structural fidelity of the synthetic Oura datasets, we projected the original and DP versions into 2D using UMAP with *n_neighbors* = 45 and *min_dist* = 0.025, shown in Figure 3. This configuration prioritizes global coherence while allowing for tight, well-defined local clusters. The choice of a very small *min_dist* = 0.025 is particularly suitable here because the Oura dataset is relatively low-dimensional, and the data shows naturally compact clusters in its latent structure.

In the original dataset, we observe a clear separation into four primary clusters. The top left cluster predominantly consists of people with low PSS, ranging from 0 to 15. The remaining three clusters are populated mostly with higher perceived stress scores, particularly in the 15–20 range, with some individuals reaching scores up to 40. Among the synthetic datasets, those generated with $\varepsilon = 2$, 5, and 10 show the strongest structural resemblance to the original. The cluster geometry and distribution of perceived stress levels are highly consistent with the original projection, suggesting that these privacy budgets strike a compelling balance between utility and privacy in the Oura dataset.

Survey dataset

To evaluate the structure of the synthetic Survey datasets, we applied UMAP, shown in Figure 4, with *n_neighbors* = 6 and *min_dist* = 0.6. The original dataset shows well-defined clusters by stress level (10-15, 15-20, 20-25, 30-35). Among the synthetic datasets, the projections for $\varepsilon = 5$, $\varepsilon = 10$, and $\varepsilon = 100$ most closely resemble the original structure, preserving both the spatial arrangement and perceived stress scores color distribution of the clusters. In contrast, lower ε values (e.g., $\varepsilon = 1$ or $\varepsilon = 2$) lead to noticeable dispersion and less consistent cluster geometry, indicating a loss of utility from strong privacy constraints.

Membership inference attack via closest distance

Based on the previous sections, we chose the DP-generated datasets with $\varepsilon = 5$ as our candidate datasets, as it was the lowest ε that maintained utility and resemblance to the original datasets. We performed the membership inference attack via closest distance from the TAPAS framework[18] to test their resilience. In this attack, the AIM synthesizer is fixed at $\varepsilon = 5$, then used as a black-box generator to repeatedly produce synthetic draws of the same size as the real data.

In contrast to our earlier linkage attack, where we focused only on records with unique attributes, here we randomly select target records, because privacy attacks should ideally be conducted on all records of a dataset rather than focusing only on vulnerable records with unique attributes.[29] Due to computational constraints, we selected only 10 random targets from each dataset to perform the attack. The AIM generator was trained once on the full dataset, and synthetic samples were drawn repeatedly without retraining, ensuring computational

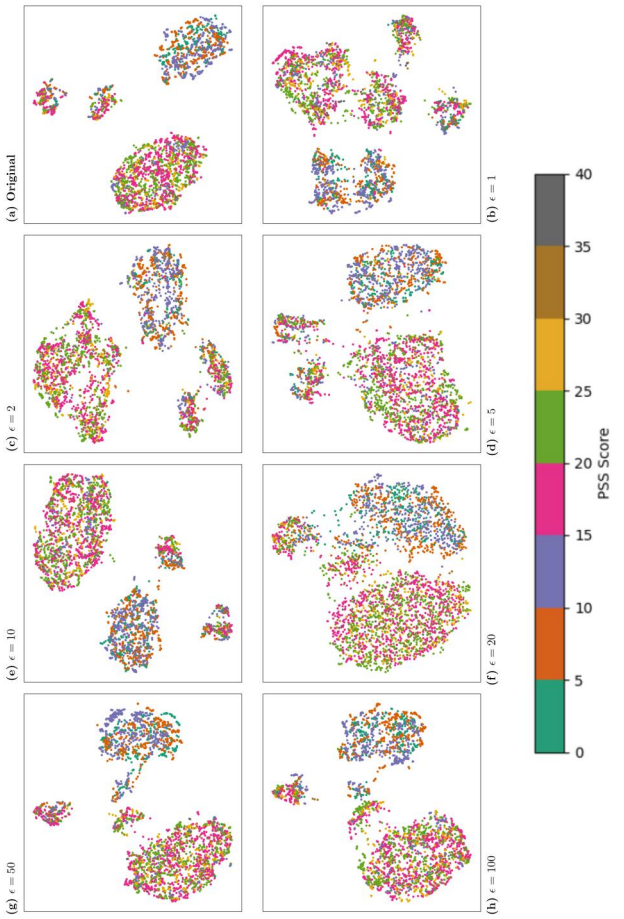


Figure 3. UMAP projections of the original and synthetic Oura datasets for different ε values. Vertical captions are placed beside each image, and the legend on the right shows perceived stress scores (PSS).

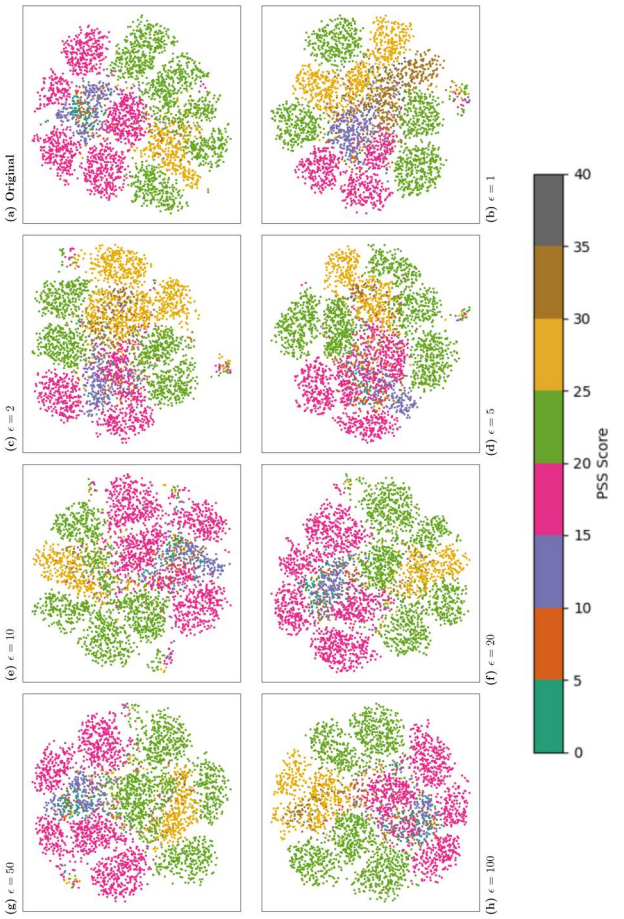


Figure 4. UMAP projections of the original and synthetic Survey datasets for different ε values. Vertical captions are placed beside each image, and the legend on the right shows the perceived stress score (PSS).

efficiency. Each attack was repeated with 200 synthetic draws per target to account for randomness.

The resulting ROC curves, shown in Figure 5, for all 10 targets fluctuate around the diagonal baseline, with no consistent deviation, suggesting unsuccessful attacks. In several cases, the curves fall slightly below the diagonal, indicating that the attack performed worse than random guessing. These results demonstrate that at $\varepsilon = 5$, the AIM-generated synthetic data for both Survey and Oura datasets provides strong resistance to membership inference, even when evaluated under the targeted threat model of TAPAS. More attack scenarios proving $\varepsilon = 5$ is a good candidate, can be found in the appendix, available as supplementary data at [*JAMIA Open*] online.

Discussion

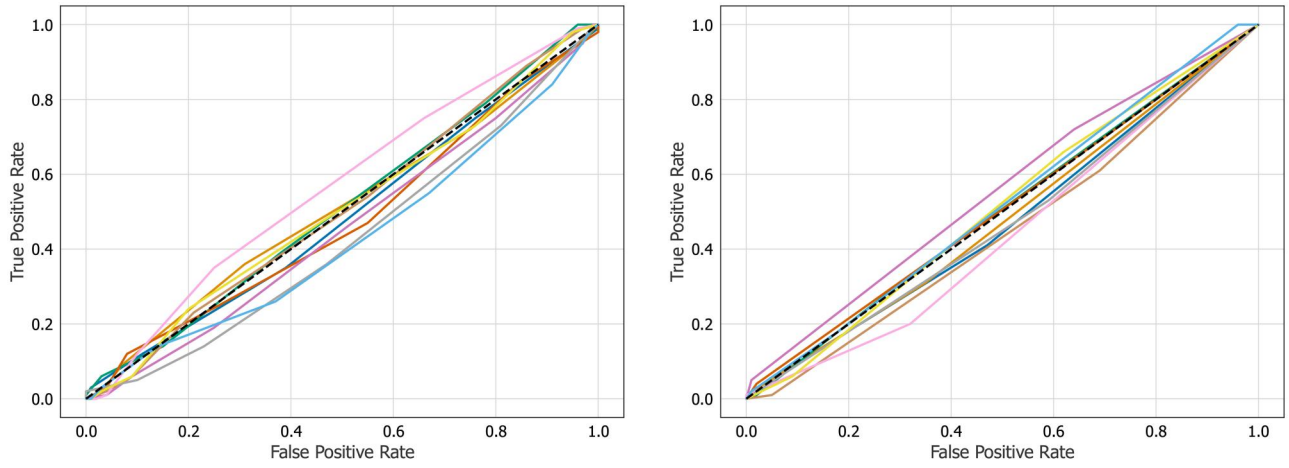
Our findings demonstrate that a mid-range privacy budget ($\varepsilon = 5$ or 10) offers a practical balance for DP synthetic data for LEMURS, preserving analytical utility while protecting against privacy attacks.

Tuning privacy parameters based on utility metrics can introduce information leakage that may compromise privacy. Notably, UMAP reveals the most information about the original dataset, which categorizes individuals based on their PSS. However, Yang et al.[30] demonstrated in their work that dimensionally reduced samples and clusters of UMAP reinforces heterogeneity and does not reveal the identity of any individual, and it provides valuable information on the data set, which aligns with the primary objective of this study: making the data set accessible to the public with meaningful information, thus preserving the privacy of the participants.

In this work, we adopt the "epistemic parity" and empirically investigate how varying ε impacts privacy-utility trade-off on downstream analyses in the context of health data. Utility loss is substantial at $\varepsilon = 1$ but plateaus after $\varepsilon = 5$ as synthesizers capture most structure and residual error reflects model limitations rather than DP noise.

Incorporating DP into real-world data workflows could transform how privacy policies are written and evaluated, enabling verification that data-sharing practices align with stated protections. This shift not only improves transparency but also strengthens public trust in institutions handling sensitive data.

The workload-aware nature of AIM, which focuses on specific marginals relevant to the analysis, contributes to its strong performance. By allocating the privacy budget to important relationships, it preserves meaningful patterns while minimizing



(a) Survey Dataset (b) Oura Dataset

Figure 5. ROC curves for the Closest-Distance membership inference attack on 10 randomly selected records across the Survey and Oura datasets at $\varepsilon = 5$ using AIM.

exposure of individual-level data. This makes AIM particularly suitable for health and behavioral datasets, where maintaining statistical trends is essential for the impact of research, but privacy risks are acute.

While our results highlight $\varepsilon = 5$ as a good balance for this dataset and use case, we do not recommend a one-size-fits-all value. Rather, we argue for the feasibility and importance of evaluating multiple privacy budgets depending on the analytic goals and sensitivity of the variables involved. Although such evaluations may require computational resources, especially for iterative or workload-aware mechanisms like AIM, they enable a more tailored and rigorous selection of ε values that align with real-world utility and risk trade-offs.

Limitations

While our study highlights the strengths of AIM, its ability to produce utility-preserving synthetic datasets, maintain statistical relationships under moderate privacy budgets, and support privacy claims with formal guarantees, it is not without limitations. Firstly, even a moderate privacy budget of $\varepsilon = 5$ may be deemed excessive in certain regulatory or clinical settings, particularly when conducting a critical life-dependent study whose outcome is intended to predict a dose of critical medication.[31]

Secondly, while AIM preserves some pairwise relationships, maintaining complex multivariate correlations remains challenging, especially for high-dimensional datasets. Additionally, AIM is designed exclusively to preserve numerical attributes and disregards qualitative values. Lastly, we did not assess re-identification risks beyond basic linkage attacks.

Thirdly, a significant consideration for this work is the distinction between event-level and user-level privacy. Because the LEMURS dataset is longitudinal, with each user contributing multiple rows, our application of AIM protects each row independently. This provides event-level differential privacy, meaning the synthesizer's output is not unduly influenced by any single daily entry. However, it does not provide formal user-level privacy, which would protect the entire contribution of a participant.

Conclusion and future work

In this work, we evaluated the performance of the AIM algorithm for generating differentially private synthetic health data using a real-world dataset on college students' physiological and mental health. We compared outputs across a range of ε values and found that $\varepsilon = 5$ offered the best balance of privacy and utility. This aligns with recent scholarship that emphasizes that privacy evaluation must consider the practical context and goals of data use.

Future work will explore alternative DP mechanisms such as GAN-based approaches and text-based DP algorithms for the qualitative columns, as well as utility preservation at lower privacy budgets (e.g., $\varepsilon < 1$). We also plan to assess the robustness of these mechanisms under more advanced privacy attacks and deploy them in practical research workflows involving underrepresented groups, where privacy risks are especially pronounced.

Acknowledgments

We are thankful to Peter Wilson and Anthony Barrows for their insightful comments. This study protocol was reviewed and approved by the University of Vermont Institutional Review Board (protocol number 00002126).

Author contributions

Mohsen Ghasemizade (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Juniper Lovato (Conceptualization, Data curation, Methodology, Project administration, Writing—original draft, Writing—review & editing), Chris Danforth (Conceptualization, Data curation, Methodology, Project administration, Writing—original draft,

Writing—review & editing), Peter Sheridan Dodds (Data curation, Project administration), Laura S.P. Bloomfield(Data curation, Project administration), and Matthew Price (Data curation, Project administration), Joseph Near (Conceptualization, Methodology, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Conflicts of interest

None declared.

Funding statement

This work was supported by the National Science Foundation award #2242829 and the MassMutual Center of Excellence in Complex Systems and Data Science. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the aforementioned financial supporters.

Data availability

The dataset and the codes for generating the dataset are available in our GitHub repository: https://github.com/mghasemizade/lemurs_dp.

References

1. Price M, Hidalgo JE, Bird YM, et al. A large clinical trial to improve well-being during the transition to college using wearables: the lived experiences measured using rings study. *Contemp Clin Trials*. 2023;133:107338.
2. Sweeney L. AboutMyInfo.org; 2024. Accessed August 26, 2024. https://aboutmyinfo.org/.
3. Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA l Rev*. 2009;57:1701.
4. contributors W. Netflix Prize; 2024. Accessed August 26, 2024. https://en.wikipedia.org/wiki/Netflix%5FPrize.
5. Tockar A. Riding with the stars: passenger privacy in the NYC taxicab dataset; 2014. Accessed August 26, 2024. https://agkn.wordpress.com/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.
6. Mason AE, Hecht FM, Davis SK, et al. Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict study. *Sci Rep*. 2022;12:3463.
7. Shiba SK, Temple CA, Krasnoff J, et al. Assessing adherence to multi-modal oura ring wearables from COVID-19 detection among healthcare workers. *Cureus*. 2023;15:e45362.
8. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006*, New York, NY, USA, March 4–7, 2006. Proceedings 3. Springer; 2006: 265–284.
9. Dwork C, Roth A. The algorithmic foundations of differential privacy. *Foundati Trends Theoret Comput Sci*. 2014; 9:211-487.
10. Center VCS. LEMURS: Lived Experiences Measured Using Rings Study; 2024. Accessed November 20, 2025. https://www.vermontcomplexsystems.org/projects/lemurs/.
11. Bloomfield L, Fudolig MI, Dodds PS, et al. Events and behaviors associated with symptoms of generalized anxiety disorder in first-year college students. PsyArXiv; 2023. https://dx.doi.org/10.31234/osf.io/278ey.
12. Bloomfield LSP, Fudolig MI, Kim J, et al. Predicting stress in first-year college students using sleep data from wearable devices. *PLOS Digit Health*. 2024;3:e0000473.
13. Fudolig MI, Bloomfield LSP, Price M, et al. Collective sleep and activity patterns of college students from wearable devices. *npj Complex*2. 2025;32. https://dx.doi.org/10.1038/s44260-025-00055-x
14. Dankar FK, El Emam K. The application of differential privacy to health data. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops; 2012:158–166.
15. Dwork C. The promise of differential privacy: a tutorial on algorithmic techniques. In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, D* (Oct. 2011). Citeseer; 2021:1–2.
16. Vu D, Slavkovic A. Differential privacy for clinical trial data: preliminary evaluations. In: *2009 IEEE International Conference on Data Mining Workshops*. IEEE; 2009:138–143.
17. Zhang J, Zhang Z, Xiao X, Yang Y, Winslett M. Functional mechanism: regression analysis under differential privacy. "Proceedings of the VLDB Endowment 5.11. 2012.
18. Houssiau F, Jordon J, Cohen SN, et al. TAPAS: a toolbox for adversarial privacy auditing of synthetic data. 2022. arXiv preprint arXiv : 221106550.
19. Rosenblatt L, Herman B, Holovenko A, et al. Epistemic parity: reproducibility as an evaluation metric for differential privacy. *ACM SIGMOD Record*. 2024;53:65–74.
20. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. 2018. arXiv preprint arXiv : 180203426.
21. Ghasemizade M, Onaolapo J. Developing a hierarchical model for unraveling conspiracy theories. *EPJ Data Sci*. 2024;13:31.
22. Record Linkage Development Team. Record linkage documentation; 2025. Accessed January 27, 2025. https://recordlinkage.readthedocs.io/en/latest/.
23. Balle B, Barthe G, Gaboardi M, Hsu J, Sato T. Hypothesis testing interpretations and renyi differential privacy. In: *International Conference on Artificial Intelligence and Statistics*. PMLR; 2020:2496–2506.
24. Near JP, Darais D, Lefkovitz N, Howarth G, et al. *Guidelines for Evaluating Differential Privacy Guarantees*. National Institute of Standards and Technology; 2023:800–226.
25. Wood A, Altman M, Bembenek A, et al. Differential privacy: a primer for a non-technical audience. *Vand J Ent Tech L*. 2018; 21:209.

26. McKenna R, Mullins B, Sheldon D, Miklau G. Aim: an adaptive and iterative mechanism for differentially private synthetic data. Proceedings of the VLDB Endowment. 2022;15 (11):2599–2612.
27. McKenna R, Miklau G, Sheldon D. Winning the nist contest: A scalable and general approach to differentially private synthetic data. 2021. arXiv preprint arXiv : 210804978.
28. Rosenblatt L, Liu X, Pouyanfar S, de Leon E, Desai A, Allen J. Differentially private synthetic data: applied evaluations and enhancements. 2020. arXiv preprint arXiv : 201105537.
29. Pilgram L, Dankar FK, Drechsler J, et al. A consensus privacy metrics framework for synthetic data. *Patterns*. 2025; 6:101320.
30. Yang Y, Sun H, Zhang Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep*. 2021;36:109442.
31. Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In: 23rd USENIX security symposium (USENIX Security 14); 2014:17–32.

JAMIA Open, 2026, 9, 1–10
https://doi.org/10.1093/jamiaopen/ooag066
Research and Applications