Received: 23 June 2024 | Accepted: 30 January 2025

DOI: 10.1111/bjet.13576

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Ensuring privacy through synthetic data generation in education

Qinyi Liu[1] | Ronas Shakya[1] | Jelena Jovanovic[2] | Mohammad Khalil[1] | Javier de la Hoz-Ruiz[3]

[1]Centre for the Science of Learning & Technology, University of Bergen, Bergen, Norway

[2]Faculty of Organisational Sciences, University of Belgrade, Serbia; Centre for the Science of Learning & Technology (SLATE), University of Bergen, Bergen, Norway

[3]Faculty of Education Sciences, University of Granada, Granada, Spain

Correspondence
Qinyi Liu, University of Bergen,
Christiesgate 12, 2nd floor, Bergen 5020,
Bergen, Norway.
Email: qinyi.liu@uib.no

Funding information
No funding was used for this study

Abstract: High-volume, high-quality and diverse datasets are crucial for advancing research in the education field. However, such datasets often contain sensitive information that poses significant privacy challenges. Traditional anonymisation techniques fail to meet the privacy standards required by regulations like GDPR, prompting the need for more robust solutions. Synthetic data have emerged as a promising privacy-preserving approach, allowing for the generation and sharing of datasets that mimic real data while ensuring privacy. Still, the application of synthetic data alone on educational datasets remains vulnerable to privacy threats such as linkage attacks. Therefore, this study explores for the first time the application of *private synthetic data*, which combines synthetic data with differential privacy mechanisms, in the education sector. By considering the dual needs of data utility and privacy, we investigate the performance of various synthetic data generation techniques in safeguarding sensitive educational information. Our research focuses on two key questions: the capability of these techniques to prevent privacy threats and their impact on the utility of synthetic educational datasets. Through this investigation, we aim to bridge the gap in understanding the balance between privacy and utility of advanced privacy-preserving techniques within educational contexts.

KEYWORDS

artificial intelligence for education, educational data mining, privacy, synthetic data

© 2025 British Educational Research Association.

Practitioner notes

What is already known about this topic

- Traditional privacy-preserving methods for educational datasets have not proven successful in ensuring a balance of data utility and privacy. Additionally, these methods often lack empirical evaluation and/or evidence of successful application in practice.
- Synthetic data generation is a state-of-the-art privacy-preserving method that has been increasingly used as a substitute for real datasets for data publishing and sharing. However, recent research has demonstrated that even synthetic data are vulnerable to privacy threats.
- Differential privacy (DP) is the gold standard for quantifying and mitigating privacy concerns. Its combination with synthetic data, often referred to as *private synthetic data,* is presently the best available approach to ensuring data privacy. However, private synthetic data have not been studied in the educational domain.

What this study contributes

- The study has applied synthetic data generation methods with DP mechanisms to educational data for the first time, provided a comprehensive report on the utility and privacy of the resulting synthetic data, and explored factors affecting the performance of synthetic data generators in the context of educational datasets.
- The experimental results of this study indicate that no synthetic data generator consistently outperforms others across all evaluation metrics in the examined educational datasets. Instead, different generators excel in their respective areas of proficiency, such as privacy or utility.
- Highlighting the potential of synthetic data generation techniques in the education sector, this work paves the way for future developments in the use of synthetic data generation for privacy-preserving educational research.

Implications for practice and/or policy

- Key takeaways for practical application include the importance of conducting case-specific evaluations, carefully balancing data privacy with utility and exercising caution when using private synthetic data generators for high-precision computational tasks, especially in resource-limited settings as highlighted in this study.
- Educational researchers and practitioners can leverage synthetic data to release data without compromising student privacy, thereby promoting the development of open science and contributing to the advancement of education research.
- The robust privacy performance of DP-synthetic data generators may help alleviate students' privacy concerns while fostering their trust in sharing personal information.
- By improving the transparency and security of data sharing, DP-synthetic data generators technologies can promote student-centred data governance practices while providing a strong technical foundation for developing responsible data usage policies.

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

INTRODUCTION

High-volume, high-quality and diverse datasets, along with advanced statistical and computational tools, can accelerate research and innovation in the education field. This synergy has given rise to new interdisciplinary and data-driven areas such as educational data mining and learning analytics. However, many large-scale educational datasets often contain highly sensitive data (eg, student grades, login data and IP addresses). Using, sharing or publishing sensitive data often violates increasingly strict privacy regulations, such as the General Data Protection Regulation (GDPR) (Jordon, Szpruch, et al., 2022; Prinsloo et al., 2019). Moreover, as the technologies evolve, maintaining the privacy of educational datasets becomes increasingly challenging (Hutt et al., 2023; Ladjal et al., 2022; Liu & Khalil, 2023). Educational research on ethics and privacy further highlights the complexity of this issue. For example, Yacobson et al. (2021) used unsupervised machine learning techniques to successfully identify sensitive student data that have been anonymised and published. This suggests that traditional means of anonymisation may be insufficient to face the latest advances in AI technology, thus resulting in further requirements for robust privacy-preserving techniques for educational datasets. Still, common privacy-preserving techniques previously used in the education sector often come at the expense of data utility (Khalil & Ebner, 2016). This means that as privacy-preserving measures are enhanced, the utility of the data—such as its ability to accurately predict outcomes—declines significantly.

The growing consensus on the need for ethical and responsible use of data, both legally and socially, has increased the demand for robust types of anonymisation. Furthermore, as more datasets are publicly released, attackers can link two or more public datasets to re-identify individuals within these datasets by overlaying information (ie, linkage attack). This is particularly concerning for educational datasets, which contain sensitive details like student academic records and personal identifiers. Such information is valuable to attackers who aim to exploit such data for malicious purposes, for example, by targeting institutions that could benefit from access to a ready-made database of potential clients or recruits. One recent example is the large-scale data breach experienced by Illuminate Education. Due to a cyberattack, the personal information of approximately 820,000 current and former students was exposed and subjected to linkage attacks, including sensitive data such as eligibility for special education services (Keierleber, 2022).

Against the above backdrop, synthetic data have emerged as a state-of-the-art privacy-preserving solution for highly sensitive information sectors, such as healthcare and finance (Koenecke & Varian, 2020; Yale et al., 2020). Synthetic data differs from traditional privacy-preserving techniques in that it allows data custodians to share and publish artificial datasets generated by algorithms mimicking real data to meet privacy requirements and promote open science. Given its purported good performance in ensuring both utility and privacy, the application of synthetic data has been slightly explored in the education sector, for example by Bautista and Inventado (2021) and Liu et al. (2024). However, recent studies indicate that synthetic data are not a panacea for privacy protection. In particular, the use of synthetic data as a privacy-preserving technique may not be sufficient in some high-privacy requirement cases. For instance, sensitive outlier data, such as students with different learning patterns than the cohort within synthetic datasets, are vulnerable to linkage attacks (Stadler et al., 2022).

Private synthetic data, which combines synthetic data with differential privacy (DP) mechanisms, are gradually coming into focus for researchers as a means of addressing some of the privacy shortcomings of using synthetic data alone (Stadler et al., 2022). Differential privacy achieves strong privacy guarantees by adding noise, and its combination with synthetic data is still a new area that requires further exploration (Rosenblatt et al., 2020). Research exploring the performance of private synthetic data in terms of privacy and utility has already begun in other areas such as healthcare (Kaabachi et al., 2023). However, no such attempts have yet been made with educational datasets, a stark contrast to the increasingly pressing challenges surrounding personal data in education. Additionally, as machine learning continues to expand in the education sector, private synthetic data demonstrate higher compatibility with machine learning compared to many encryption-based privacy protection methods (Li et al., 2017). This makes exploring both non-private and private synthetic data generators well-suited to addressing future needs in educational data analysis.

Another important reason to explore the impact of private synthetic data generators (SDGs) on data in the educational domain is that education datasets often include learning events, either as raw event data or as features derived from such data (Buu & Karunaratne, 2024). This further means that even though educational data bears similarity to data in other privacy-sensitive fields where private SDGs have been successfully applied (eg, healthcare or finance), there are also some notable differences. This implies that lessons learned and conclusions drawn from applying differential privacy in other fields cannot be directly applied to the educational context without first systematically examining the effect of differential privacy noise on student data. Finally, considering the multi-stakeholder nature of educational datasets, the evaluation of distinct SDGs needs to incorporate a comprehensive and representative selection of utility and privacy evaluation metrics. Educational data are not dedicated to a single purpose, but may serve school administrators for decision-making, education policymakers for trend analysis, teachers for evaluating teaching strategies and educational technology companies for optimising system design (Hutt et al., 2023). This multi-purpose characteristic imposes complex demands on both data privacy and utility, making it essential to evaluate both aspects simultaneously from a holistic perspective.

Based on above reasons, the current research study focuses on private synthetic tabular data (see the definition in 2.3) and addresses the following research questions (RQs):

RQ1: To what extent can distinct synthetic data generation techniques prevent potential privacy threats in tabular education datasets?
RQ2: How much do different synthetic data generation techniques affect the utility of the resulting synthetic educational dataset?

RELATED WORK

Privacy challenge in educational datasets

Privacy is a vast field that can be approached from many different perspectives. In this paper, we adopt the perspective on privacy within the context of machine learning as proposed by Jordon, Yoon, and van der Schaar (2022), highlighting that the essence of privacy hinges on individuals' consent to data collection and its subsequent information release should not inflict harm upon them. The potential harm arises from adversaries gaining access to individuals' information through the release of data or derived outputs (eg, algorithm output) (Jordon, Yoon, & van der Schaar, 2022). This definition of privacy is particularly relevant to the data protection issues that may arise when sharing and releasing educational datasets, which are the focus of this paper.

The current privacy status of educational datasets is far from desirable (Liu & Khalil, 2023). In the broader context, stringent data protection laws such as GDPR impose strict requirements on the publication and sharing of data involving personal information, mandating that publicly released datasets cannot be traced back to individuals (Jordon, Szpruch, et al., 2022). This underscores the necessity for proper handling of privacy compliance in educational datasets, as non-compliance constitutes a violation of the law. However, the

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

privacy landscape of educational datasets reveals dichotomous trends in the context of their publication and sharing: on one hand, there is an increasing richness and volume of data types, while on the other hand, traditional privacy-enhancing techniques that used to be applied to educational datasets are lagging behind. This dichotomy has been directly addressed by Hutt et al. (2022), particularly in the context of online courses where the volume of data collected is substantial, rendering traditional anonymisation insufficient. As the volume of data increases, the likelihood of data breaches such as linkage attacks becomes higher, necessitating more advanced privacy-preserving techniques. Educational datasets often contain subtle identifiers, such as demographic information and IP addresses, which necessitate further technological means to ensure privacy protection (Hutt et al., 2022). Additionally, previously employed techniques on student datasets, such as *k*-anonymity and l-diversity (Gursoy et al., 2016) are considered relatively outdated. Their usage as privacy-enhancing techniques has been in decline since the emergence of differential privacy (Appenzeller et al., 2022). Instead, they are nowadays applied as re-identification measures to evaluate privacy after applying privacy-preserving methods. The trade-off between privacy and data utility in educational datasets has been unsatisfactory with previous technical approaches, often resulting in significant sacrifices in data utility for privacy enhancements (Khalil & Ebner, 2016). Furthermore, many methods developed to protect privacy in educational datasets lack proper evaluation and/or evidence from actual application (Liu & Khalil, 2023).

The combined use of synthetic data generation and differential privacy has been proposed in the field of computer science as a novel approach to addressing the privacy challenges of data sharing and publication. Synthetic data generation as a method that balances privacy and data utility, especially when integrated with differential privacy, holds a prominent privacy-preserving potential. In the following two sections, we introduce differential privacy and synthetic data, including private synthetic data, respectively.

Differential privacy

Differential privacy (DP) was introduced by Cynthia Dwork in 2006 to establish a privacy guarantee independent of any background knowledge (Dwork, 2006). Through years of practice, DP has come to be regarded as a robust, meaningful and practical privacy guarantee (Jordon, Yoon, & van der Schaar, 2022; Liu et al., 2025). DP ensures that the output of any random algorithm should not differ significantly between datasets that differ in only one individual record. This means that the results of operations on a differentially private dataset should be similar or indistinguishable whether an individual record is included or not.

In DP, there are two important settings and parameters. First is the interactive/non-interactive setting. In a non-interactive setting, differentially private data are released once as a synthetic dataset, making it suitable for scenarios where data needs to be shared and published widely (Li et al., 2017). In contrast, an interactive setting answers queries individually with privacy guarantees each time, making it ideal for scenarios requiring dynamic, on-demand access to data while maintaining privacy (Li et al., 2017). As this paper focuses on investigating the privacy of publishing educational datasets, we have employed a non-interactive setting when configuring DP in our experiments. The parameter ϵ in DP quantifies and crucially determines the balance between data privacy and utility. Lower values of ϵ enhance privacy but reduce utility, whereas higher values decrease privacy while improving utility. In DP-synthetic data generation, the ϵ can be tuned to attain the desired privacy versus utility trade-off. The combination of DP with synthetic data, often referred to as *private synthetic data*, is described next. Mathematical definition of DP is given in the supplementary file (Section S1).

Synthetic data

According to Jordon, Szpruch, et al. (2022), synthetic data are “data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)” (p. 5). Types of synthetic data include text data, tabular data and media data, but since this paper targets educational tabular dataset, the focus is exclusively on synthetic tabular data. Tabular data refer to data that are organised into tables in which information is arranged in rows and columns (Krishnamurthi et al., 2017). Each row usually represents a record, while each column represents a variable associated with the record. As for the use of synthetic data, one of the primary uses of synthetic data is private data release (Jordon, Szpruch, et al., 2022). Synthetic data can serve as a substitute for authentic data that cannot be externally released due to privacy concerns, thus supporting open science and maintaining data quality while complying with privacy regulations such as GDPR (El Emam et al., 2020). This paper focuses on synthetic data use for the purpose of sharing/publishing datasets with privacy requirements.

Generation of synthetic tabular data

For the synthetic tabular data discussed in this paper, generation methods can be primarily divided into two categories: statistical methods and deep learning methods (Dankar et al., 2022; Figueira & Vaz, 2022). Statistical methods are advantageous due to their speed, ease of application, minimal computational resource requirements and manageable parameters. However, they may be unsuitable for handling large or complex datasets (Hernadez et al., 2023). Deep learning methods, particularly generative adversarial networks (GANs), are known for their efficiency and ability to learn underlying patterns in the data (Figueira & Vaz, 2022). This paper will examine representative methods from both categories: (1) Bayesian networks (BayNet) based on statistical methods, suitable for multiple data types including categorical and continuous data, and (2) CTGAN (Conditional Tabular Generative Adversarial Network) based on deep learning methods, which performs well in tabular datasets with complex relationships.

While synthetic data generation methods introduce a layer of privacy protection synthesising a substitute for each individual record, they may still leak sensitive information. For instance, if there is a strong outlier in the original data, like a student with a very idiosyncratic learning behaviour, replicating this value according to the statistical structure would certainly disclose the individual's personal information (Arnold et al., 2020). Overall, many synthetic data generation methods are vulnerable to attacks that can reveal whether an individual's data were used to train the model (ie, membership inference attacks) and attacks that can uncover sensitive attributes of individuals (ie, attribute inference attacks). These vulnerabilities can lead to sensitive information leakage and a decline in model performance (Lu et al., 2021). Thus, synthetic data still require privacy protection, and DP is proposed as a robust approach to ensure that the inclusion or exclusion of a single data point does not significantly affect the overall dataset, thereby safeguarding individual privacy (Jordon, Szpruch, et al., 2022). When synthetic data are combined with DP to create private synthetic data, the result is such that the similarity of a data point in the private synthetic data and its corresponding data point in the original data does not imply a privacy breach (Jordon, Szpruch, et al., 2022). Furthermore, the combination of synthetic data and DP is highly promising and is being used by tech companies such as Google in their attempts to apply it to their mobile devices (Kurakin & Ponomareva, 2023; Lin et al., 2024). Scholars have proposed several differentially private synthetic data generators (private SDGs). Pioneering examples include DPGAN, the first

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

private SDG (Xie et al., 2018), PrivBayes (Zhang et al., 2017) and PATEGAN (Jordon, Yoon, & van der Schaar, 2022). In this study, we select DPGAN (Differentially Private GAN) and PATEGAN (Private Aggregation of Teacher Ensembles GAN) because both have a record of generating high-quality private synthetic data using DP.

Synthetic data evaluation

Synthetic data are generally evaluated via the dimensions of utility and privacy.

Utility. Data utility, which measures the suitability of the data for processing and analysis, reflects the usefulness and reliability of the data (Dankar et al., 2022). Specifically, the utility of synthetic data is assessed by comparing them to the original dataset to check if they retain predictive capacity and important statistical properties, such as means or variances (Dankar et al., 2022). Broadly, utility includes different kinds of fidelity (ie, how closely the synthetic data distribution matches the real data) and prediction accuracy (ie, the performance of predictive algorithms on synthetic versus real data). Fidelity is categorised into: (i) attribute fidelity, which measures the basic structural similarity between the datasets; (ii) bivariate fidelity that measures correlations among pairs of variables in the dataset; (iii) population fidelity that measures the similarity of the entire distribution; and (iv) application fidelity, which evaluates the performance of synthetic data in a specific analysis, similar to prediction accuracy (Dankar et al., 2022). This study reports on selected representative metrics from each fidelity category.

Privacy. Methods for measuring the privacy of synthetic data can be divided into two categories: dataset evaluation and model evaluation (Kaabachi et al., 2023). Dataset evaluation methods focus on assessing the privacy protection provided by the synthetic data itself, examining how the synthetic dataset compares to the real dataset in terms of privacy. On the other hand, model evaluation methods assess the generative model's ability to protect the privacy of the original data (Kaabachi et al., 2023), often by calculating an upper bound on privacy risk and evaluating worst-case scenarios. These two approaches highlight the privacy of the data and the robustness of the generation models, respectively. This paper will use representative metrics from both method groups (see Section “Synthetic data evaluation methods”).

EXPERIMENTS

To answer the two RQs posed in this paper, we designed an experiment the details of which are presented in the following.

The experiment workflow (Figure 1) starts with the use of raw datasets to train SDG. PATEGAN and DPGAN methods are used as private SDGs, whereas CTGAN and BayNet are used as non-private SDGs. In the case of each SDG, the generated synthetic datasets are of the same dimensions as the original datasets. For private SDGs, to comprehensively assess the outcomes for different ϵ values, this study tests each private SDG for three distinct ϵ values (when ϵ=0.1, 1, 10). Finally, each SDG is evaluated by measuring utility and privacy of the synthesised data. Specifically, the evaluation of utility covers different aspects of fidelity. The evaluation of privacy is approached through model evaluation and dataset evaluation. To ensure robustness and reliability, the overall process of data generation and evaluation is repeated twice for each SDG, and evaluation measures averaged across the two runs.

The experiments in this study used the Synthcity python library (Qian et al., 2023), which allows for convenient generation and evaluation of synthetic tabular data. Google Colab was

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

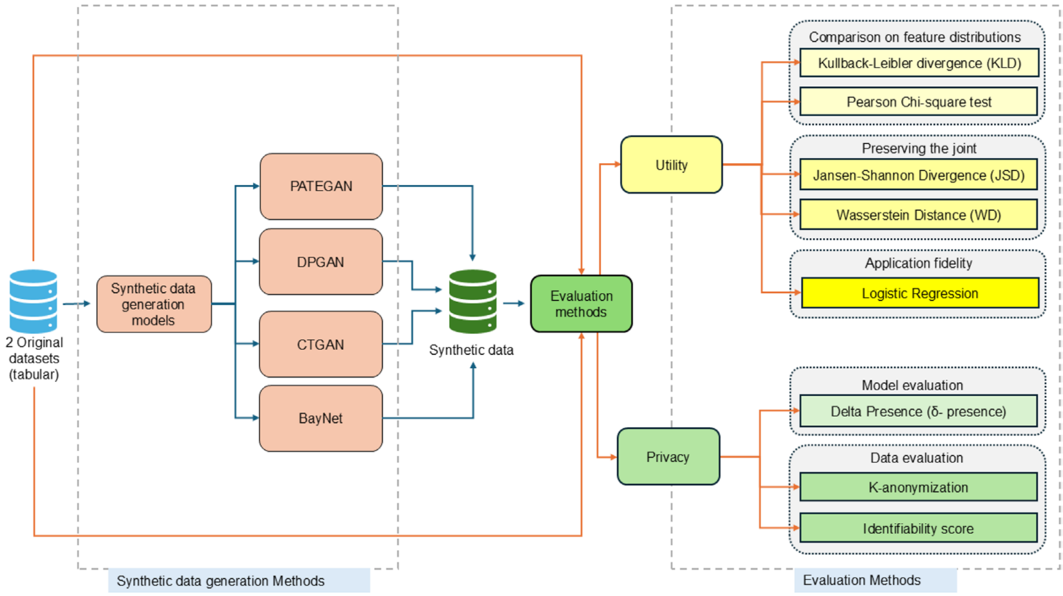
2 Original datasets (tabular)
Synthetic data generation models
PATEGAN
DPGAN
CTGAN
BayNet
Synthetic data
Evaluation methods
Utility
Privacy
Comparison on feature distributions
Kullback-Leibler divergence (KLD)
Pearson Chi-square test
Preserving the joint
Jansen-Shannon Divergence (JSD)
Wasserstein Distance (WD)
Application fidelity
Logistic Regression
Model evaluation
Delta Presence (δ- presence)
Data evaluation
K-anonymization
Identifiability score
Synthetic data generation Methods
Evaluation Methods

FIGURE 1 Pipeline of the proposed experiments.

TABLE 1 Brief description of the selected educational datasets (before pre-processing).

ID	Dataset name	Year	Number of attributes	Number of records	Target variable	#C	#B	#M	#Mi
A	Students performance in exams[a]	2018	8	1000	Continuous	3	3	2	0.05
B	Open University Learning Analytics Dataset (OULAD)—“studentInfo”[b]	2017	12	32,594	Multi-class	2	2	8	0.45

Note: The notations C, B and M represent the number of continuous, binary and multi-class categorical variables, respectively, whereas Mi stands for the imbalance ratio (ie, the ratio of minority and majority samples).

[a] Dataset retrieved from (https://www.kaggle.com/datasets/spscientist/students-performance-in-exams), licensed CC-BY.

[b] Dataset retrieved from (https://analyse.kmi.open.ac.uk/open_dataset) licensed CC-BY 4.0.

used as the computational environment, as it enables easy access to powerful computational resources, namely the Nvidia A100 GPU known for efficient handling of deep learning tasks.

Dataset selection

For our experiments, we chose two frequently used educational datasets (see Table 1) to evaluate the performance of SDG on different types and sizes of data (Hernadez et al., 2023). Both datasets are from Kaggle's publicly available dataset repositories. The first dataset is the test scores of American high school students (Dataset A). The other dataset is the

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017), which is considered one of the most comprehensive and benchmark datasets in the learning analytics domain (Alhakbani & Alnassar, 2022). Specifically, we selected the 'student info' dataset from OULAD, which includes various personal details about students (such as age, region, and disability) along with their grades. The first 10 rows (with all variables) of two used dataset are provided in the supplementary file (Tables S1 and S2). The two datasets differ significantly in size: Dataset A contains 1000 records, while Dataset B contains around 30,000 records. Regarding balance, Dataset A has a higher imbalance rate (Mi in Table 1). This indicates that Dataset A contains a significantly smaller proportion of minority class samples (less frequent category) compared to majority class samples (more frequent category), resulting in more uneven class distribution. No data pre-processing (eg removing missing values) was performed on either dataset, following the recommendation in the literature to use raw unprocessed datasets as input for an SDG (Dankar et al., 2022).

Synthetic data generation methods

We used four open-source data generation methods: (1) Bayesian Network (BayNet) by Ankan and Panda (2015), (2) Conditional Tabular Generative Adversarial Network (CTGAN) by Xu et al. (2019), (3) Differentially Private GAN (DPGAN) by Xie et al. (2018) and (4) Private Aggregation of Teacher Ensembles (PATE) GAN (PATEGAN) by Jordon, Yoon, and van der Schaar (2022). The former two were used for generating 'regular' non-private synthetic data, whereas the latter two were used for private synthetic data. Details of each method are described in the supplementary file (Section "Related work"). The rationale for their selection is outlined in Section "Generation of Synthetic tabular data". We instantiated each method with its default parameter values.

Synthetic data evaluation methods

We evaluate the generated synthetic data with respect to utility and privacy.

1. Utility—comparison on feature distributions: This kind of comparison estimates attribute fidelity (introduced in Section "Synthetic data evaluation"). It measures how well individual attributes (or columns) in the synthetic data match corresponding attributes in the original data. We use the commonly employed Kullback–Leibler divergence (KLD) (Dankar et al., 2022) and the Pearson chi-squared test, the latter especially for categorical variables. KLD measures how one probability distribution diverges from the expected probability distribution, with values ranging between [0–1], where a larger value indicates greater similarity to the original distribution. For the chi-squared test, the test statistic (chi-squared value) quantifies the difference between observed and expected frequencies. The statistical significance of this difference is evaluated based on the associated p-value. A small p-value (typically below 0.05) suggests rejecting the null hypothesis of identical distributions, indicating a significant difference.
2. Utility—preserving the joint[1]: This section corresponds to the earlier mentioned bivariate fidelity and population fidelity, assessing the fidelity of synthetic data in capturing both the inter-attribute relationships and the overall population characteristics compared to the original data. We employ two widely used metrics in the synthetic data community: Jensen-Shannon divergence (JSD), which measures the difference between the probability mass distributions of individual categorical variables, and the Wasserstein distance, which

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

measures the distance between two probability distributions, particularly for continuous variables (Yoon et al., 2020; Zhao et al., 2021). For JSD, we followed the recommendation of Qian et al. (2023) to take the square root, transforming it into Jensen-Shannon distance. The JSD value is between [0–1], and both larger JSD and WD indicate that the datasets are more dissimilar.

3. Utility—application fidelity: This is to evaluate the performance of synthetic data in downstream machine learning tasks. If the synthetic data performs similarly to the real data on the same machine learning task, its quality is considered to be high. Specifically, this paper adopts the widely used train-synthetic-test-real (TSTR) method (Jordon, Szpruch, et al., 2022), which evaluates the quality of synthetic data by training a model on synthetic data and then testing it on real data. The ratio used in this paper is 70/30, that is, 70% of the real dataset is used to generate synthetic data, whereas 30% of the held-out real dataset is used to evaluate the application fidelity of the synthetic data. The machine learning method used in this paper is logistic regression, which is commonly applied in education. To measure and report the gap between synthetic data and real data, this paper uses the utility loss metric. It is obtained by subtracting the AUCROC score of the synthetic data from the AUCROC score of the real data. A positive result indicates a performance gap between the synthetic and real data, with larger values denoting a wider gap and lower performance of SDG. A negative result means the synthetic data outperforms the real data.
4. Privacy—model evaluation: We employed the commonly used re-identification risk method, delta presence (δ-presence) (Google Cloud, 2023). Delta presence can detect member disclosure by computing δ: $\delta = 0.95$ means the attacker might learn that the target has 95% chance of being in the dataset (Qian et al., 2023). This method is especially used in dataset release scenarios, which is matched with our study context (Van den Bossche, 2023). A lower value shows a good privacy protection by the generative model and it means that the individual is not easy to be matched in the synthetic dataset, which reduces the possibility of linkage attack.
5. Privacy—dataset evaluation: We employ *k*-anonymisation and identifiability score measures. *k*-anonymisation is a privacy protection measure that ensures that each individual's information is indistinguishable from at least k-1 other individuals (Qian et al., 2023). If the synthetic dataset can maintain a high *k*-anonymity, it means that it has a stronger ability to resist linkage attacks. A higher score represents better privacy. The identifiability score returns the re-identification risk on the real dataset from the synthetic dataset (Qian et al., 2023). It is bounded between [0–1], a lower value denotes lower re-identification risk, indicating better privacy. All the mathematical definitions of evaluation metrics used in this paper can be seen in the supplementary file (Section "Synthetic data generation methods").

RESULTS

In this section, we report the experiment results.

Utility results

Tables 2 and 3 report the results of the four used SDGs across four utility metrics: KLD, chi-squared, JSD and WD. For KLD and chi-squared test results, higher values indicate better utility, whereas for JSD and WD results, lower values indicate better utility. For utility loss (Figure 3), smaller positive values indicate better utility, while negative values indicate

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Utility metrics: JSD, WD, chi-squared test and KLD for non-private synthetic data generators, averaged over the two experimental runs.

SDG	Dataset ID	Epsilon	JSD	WD	Chi-squared test	KLD
BayNet	A	/	0.029	0.167	0.604	0.823
	B	/	0.002	0.032	0.817	0.994
Average		/	0.015	0.100	0.711	0.909
CTGAN	A	/	0.010	0.057	0.681	0.837
	B	/	0.008	0.040	0.896	0.981
Average		/	**0.009**	**0.048**	**0.788**	**0.909**

Note: Bold indicates best performance in the table.

TABLE 3 Utility metrics: Average JSD, WD, chi-squared test and KLD for private synthetic data generators.

SDG	Dataset ID	Epsilon	JSD	WD	Chi-squared test	KLD
DPGAN	A	0.1	0.086 ± 0.018[a]	0.824 ± 0.246	0.243 ± 0.219	0.472 ± 0.159
		1	0.056 ± 0.003	0.418 ± 0.061	0.357 ± 0.025	0.603 ± 0.037
		10	0.056 ± 0.013	0.471 ± 0.133	0.27 ± 0.148	0.643 ± 0.09
	B	0.1	0.104 ± 0.011	1.416 ± 0.016	0.026 ± 0.026	0.276 ± 0.015
		1	0.087 ± 0.012	1.175 ± 0.236	0.34 ± 0.135	0.538 ± 0.083
		10	0.104 ± 0.009	1.893 ± 0.138	0.237 ± 0.146	0.41 ± 0.081
Average		0.1	0.095	1.120	0.135	0.374
		1	**0.072**	**0.797**	**0.349**	**0.571**
		10	0.080	1.182	0.254	0.527
PATEGAN	A	0.1	0.072 ± 0.009	0.794 ± 0.148	0.352 ± 0.106	0.543 ± 0.025
		1	0.032 ± 0.003	0.215 ± 0.016	0.748 ± 0.022	0.834 ± 0.042
		10	0.053 ± 0.016	0.396 ± 0.194	0.342 ± 0.209	0.579 ± 0.141
	B	0.1	0.01 ± 0.001	0.087 ± 0.008	0.891 ± 0.002	0.962 ± 0.001
		1	0.01 ± 0.001	0.087 ± 0.008	0.891 ± 0.002	0.962 ± 0.001
		10	0.013 ± 0.005	0.066 ± 0.011	0.876 ± 0.017	0.955 ± 0.02
Average		0.1	0.041	0.441	0.622	0.753
		1	**0.021**	**0.151**	**0.820**	**0.898**
		10	0.033	0.231	0.609	0.767

Note: Bold indicates best performance in one SDG.

[a] "±" notation is because private SDG include the level of inherent randomness by DP.

synthetic data have better utility performance than real data (as the score of real data is even smaller than synthetic data). Note that the results reported for the private SDG evaluation metrics contain "±" notation, whereas results for non-private SDGs do not have this notation. This is because non-private SDG methods do not typically include the same level of inherent randomness as DP methods. Private SDGs often produce inconsistent results across runs due to their use of random noise for ensuring privacy.

Comparing the utility metrics of private SDGs (ie, DPGAN and PATEGAN) to those of non-DP generators (ie, BayNet and CTGAN) reveals the trade-off impact of privacy-preserving techniques (see Figure 2 and Table 3). DPGAN shows improvement in utility as the epsilon (ε) increases, with JSD values decreasing from 0.095 at $\varepsilon = 0.1$ to 0.072 at

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

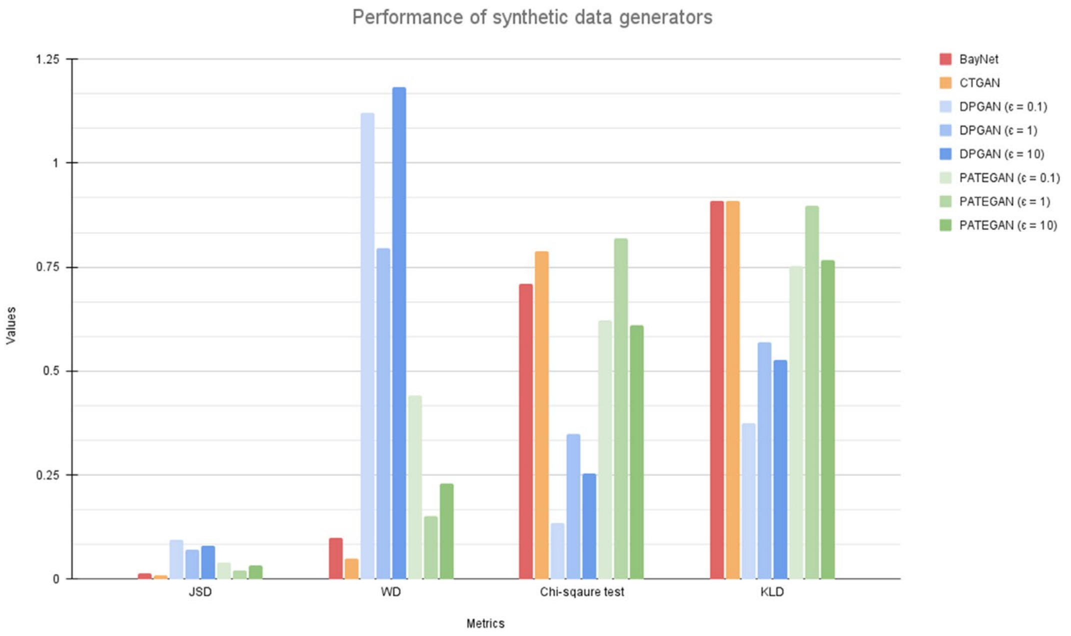

FIGURE 2 Utility metrics: Average JSD, WD, chi-squared test and KLD for four SDGs.

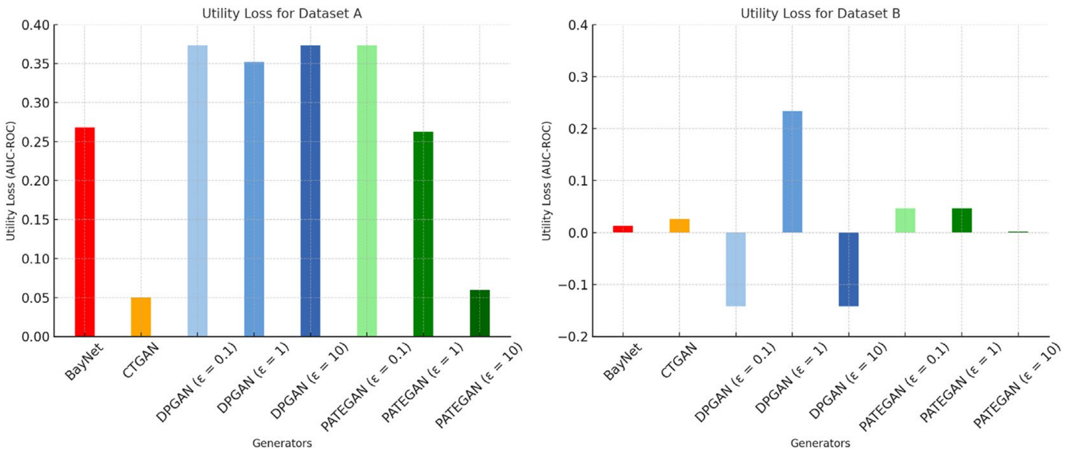

FIGURE 3 Utility metrics: Application fidelity (AUCROC difference between synthetic and real data) for the two datasets.

$\varepsilon = 1$, and slightly rising to 0.080 at $\varepsilon = 10$. However, the WD values, which are lowest at $\varepsilon = 1$ (0.797), still indicate less utility compared to CTGAN and even BayNet. Similarly, PATEGAN, while demonstrating better utility than DPGAN, with JSD values (0.041 at $\varepsilon = 0.1$, 0.021 at $\varepsilon = 1$ and 0.033 at $\varepsilon = 10$) and WD values (0.441, 0.151 and 0.231 respectively), falls short of the performance of CTGAN and BayNet in terms of JSD and WD. The general lower utility performance of private SDG compared to non-private SDG is expected. This is because DP limits the amount of information that can be extracted from the real dataset to compute the output (Jordon, Yoon, & van der Schaar, 2022). As a result, any finite sample drawn from a DP-based generator will contain less information from the original dataset compared to a

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

non-DP one, typically leading to lower utility. However, this does not mean that all private SDGs perform poorly in terms of utility, as the results for PATEGAN demonstrate.

When it comes to chi-squared test and KLD metrics, PATEGAN performs well even compared to non-private SDG, particularly at $\varepsilon = 1$, where it achieves values of 0.820 and 0.898, respectively. This indicates robust preservation of data utility. Based on the chi-squared test results, PATEGAN ($\varepsilon = 1$) even outperforms non-private SDGs. This suggests that while privacy-preserving techniques in DPGAN and PATEGAN do introduce some degradation in data utility compared to non-private SDGs, PATEGAN, especially, manages to maintain a relatively high level of utility. Overall, the comparison shows that while non-private generators like CTGAN offer the best utility, private generators like PATEGAN still show close utility performance, outperforming DPGAN and maintaining competitive utility levels. The difference between the non-private SDGs is very small, with the largest difference on the chi-squared test (0.078) and the smallest difference in KLD values (<0.0001). This goes in line with previous results on datasets from other domains (such as Stadler et al., 2022), where both BayNet and CTGAN are top-performing generators for tabular data generation.

Finally, Figure 3 reports the performance differences across the examined SDGs on a downstream machine learning task. As shown in Figure 3, overall, the utility loss of Dataset B is lower than that of Dataset A, particularly for private SDGs. This is because private SDGs tend to perform better on larger datasets (Jordon, Yoon, & van der Schaar, 2022). Additionally, the figure shows that DPGAN on Dataset B (when ε equals 1 and 10) shows negative values, which indicates that the machine learning performance of DPGAN generated synthetic data is even better than real dataset. While in general, the machine learning performance of synthetic data is expected to be inferior to that of real datasets (Jordon, Yoon, & van der Schaar, 2022). The reasons behind the phenomenon observed in Figure 3 will be discussed in more detail in the next section combined with privacy results. Additionally, both Figures 2 and 3 demonstrate that, in terms of utility, private SDGs generally underperform compared to non-private SDG. However, certain private SDGs, such as PATEGAN ($\varepsilon = 10$), perform on par with the best-performing non-private SDGs in both datasets. Additionally, both figures show that private SDGs do not follow the trend, documented in the literature, of utility decrease as ε increases. For example, in Figure 2, PATEGAN at ($\varepsilon = 1$) performs better in terms of JSD, WD, chi-squared test and KLD compared to $\varepsilon = 10$. Similarly, in Figure 3, the utility loss of DPGAN on Dataset A at $\varepsilon = 1$ is smaller than at $\varepsilon = 10$. This may be due to overfitting. A higher privacy budget ($\varepsilon = 10$) means less noise is added, allowing the model to learn more precise patterns from the original data, which can increase the risk of overfitting (Van Breugel et al., 2023). The model may capture irrelevant details along with true patterns that eventually lead to the decline of utility values. In contrast, a lower privacy budget (more noise) introduces randomness, making it harder for the model to overfit, as it struggles to learn specific irrelevant details.

Privacy results

Tables 4 and 5 report the results of the four used SDGs (including non-private SDG and private SDG) across three privacy metrics: Delta-presence, *k*-anonymity and identifiability score. For delta-presence and identifiability score, smaller values indicate better privacy. For *k*-anonymity, higher values are better.

The four SDGs show distinct performance in terms of the privacy metrics. Overall, private SDGs tend to perform better than non-private ones. For identifiability score, DPGAN with $\varepsilon = 0.1$ shows the best privacy, while for *k*-anonymity, PATEGAN with $\varepsilon = 0.1$ excels. However, some exceptions exist, such as CTGAN demonstrating superior privacy compared to private SDGs, with a delta-presence of 2.61, followed closely by PATEGAN ($\varepsilon = 0.1$) at

TABLE 4 Privacy metrics: Delta presence, *k*-anonymisation and identifiability score for different synthetic data generators.

SDG	Dataset ID	Epsilon	Delta-presence	*k*-anonymisation	Identifiability score
Real dataset	A	/	/	1.000	/
	B	/	/	28.000	/
BayNet	A	/	7.667	1.000	0.135
	B	/	1.232	25.000	0.4356
Average		/	4.449	**13.000**	**0.285**
CTGAN	A	/	3.000	2.000	0.315
	B	/	2.222	17.000	0.392
Average		/	**2.611**	9.500	0.353

Note: Bold values indicate best performance in the table.

TABLE 5 Privacy metrics: Delta presence, *k*-anonymisation and identifiability score for private synthetic data generators.

SDG	Dataset ID	Epsilon	Delta-presence	*k*-anonymisation	Identifiability score
DPGAN	A	0.1	27.000 ± 2.000[a]	2.500 ± 1.500	0.020 ± 0.020
		1	15.333 ± 7.667	5.000 ± 1.000	0.148 ± 0.007
		10	14.500 ± 1.500	3.000 ± 1.000	0.143 ± 0.028
	B	0.1	2409.500 ± 1558.500	2.500 ± 1.500	0.074 ± 0.001
		1	231.689 ± 68.144	35.000 ± 29.000	0.125 ± 0.011
		10	275.375 ± 174.375	11.500 ± 3.500	0.032 ± 0.009
Average		0.1	1218.250	2.500	**0.047**
		1	**123.511**	**20.000**	0.137
		10	144.9375	7.250	0.088
PATEGAN	A	0.1	5.759 ± 6.398	3.400 ± 2.728	0.004 ± 0.002
		1	17.240 ± 7.989	2.400 ± 0.490	0.086 ± 0.05
		10	12.867 ± 6.145	4.400 ± 0.800	0.185 ± 0.04
	B	0.1	2.005 ± 0.423	50.500 ± 9.500	0.378 ± 0.023
		1	2.005 ± 0.423	50.500 ± 9.500	0.378 ± 0.023
		10	6.036 ± 2.536	29.000 ± 7.000	0.352 ± 0.066
Average		0.1	**3.882**	**26.950**	**0.191**
		1	9.623	26.450	0.232
		10	9.452	16.700	0.269

Note: Bold values indicate best performance in one SDG.

[a] "±" notation is because private SDG include the level of inherent randomness by DP.

3.88. BayNet's performance in privacy is not as strong as CTGAN's. In summary, if users prioritise privacy above all else, PATEGAN should be considered first due to its stable and good privacy performance. CTGAN is also a good option; despite not being a private SDG, it rivals private SDGs in some privacy metrics. Considering CTGAN's excellent performance in utility, it is an optimal choice for users seeking a balance between privacy and utility.

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

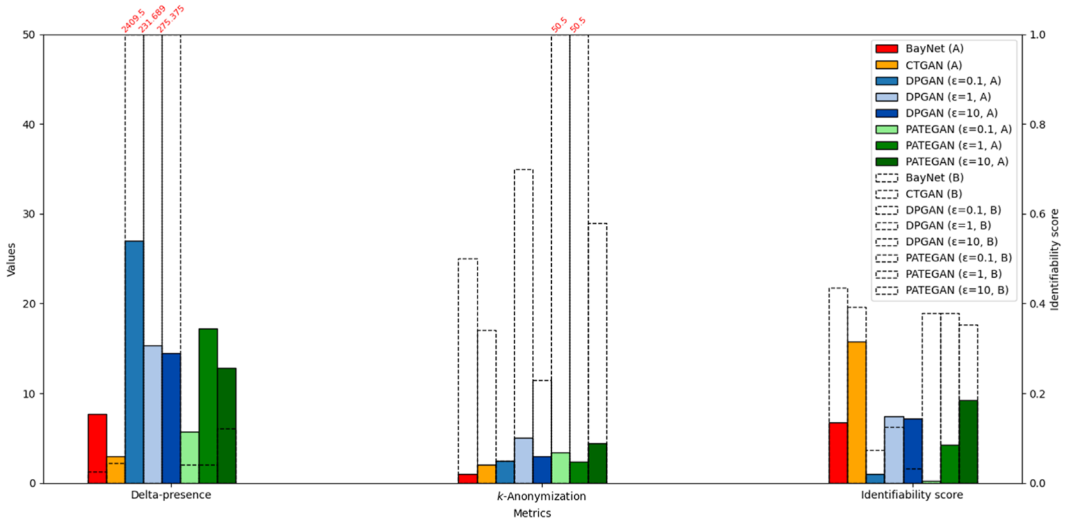

FIGURE 4 Privacy metrics for different synthetic data generators and each of the two datasets.

Another phenomenon of concern is that DPGAN displays anomalous behaviour in two aspects. First, as shown in Figure 4, *k*-anonymity and identifiability score suggest opposite conclusions about the privacy level ensured with DPGAN. This is in contrast to the more consistent patterns observed in PATEGAN, BayNet and CTGAN. Second, DPGAN's delta-presence values rise as the privacy budget epsilon decreases, which is counterintuitive. Normally, a lower epsilon should result in higher privacy and thus lower delta-presence. The reason for this can be model instability. The training process of DPGAN might be unstable, especially under strict privacy constraints, causing it to memorise and replicate specific data points rather than generalise from the data distribution (Cong et al., 2020). This instability, where DPGAN memorises and replicates specific data points, likely explains why the above machine learning performance of DPGAN-generated synthetic data in Figure 3 (Dataset B) surpasses that of the real data. The overfitting caused by this instability can lead to artificially improved performance due to excessive memorisation rather than proper generalisation.

Another interesting finding is that when comparing performance across different datasets, it is observed that the larger of the two datasets (Dataset B) performs better than the smaller one (Dataset A) in terms of overall *k*-anonymisation (larger value in Dataset B, more privacy) and delta-presence (smaller value in Dataset B, more privacy), but generally worse in terms of identifiability score (larger value in Dataset B, less privacy). This discrepancy can be attributed to factors such as dataset complexity, and how the metrics are calculated. Dataset complexity encompasses multiple dimensions, each of which increases overall complexity (Peek, 2023). The dimensions relevant to this discussion include data volume and data variety.

From the perspective of data volume and variety, the larger Dataset B contains more records and more diverse attribute combinations. This makes it easier to find multiple identical or similar records, thus satisfying the requirements of *k*-anonymisation and reducing the re-identification risk in delta-presence (Nergiz & Clifton, 2010). Additionally, considering that Dataset B had already undergone *k*-anonymisation at the time of release (Kuzilek et al., 2017), its better performance of *k*-anonymisation compared to Dataset A may be attributed to this, rather than the dataset's volume and variety. To rule out this factor, we report in the supplementary file (Section Table S3) the results of the privacy evaluation conducted on a synthetic dataset generated from a random 5% sample of Dataset B. The results show that when the volume of Dataset B is significantly reduced, both *k*-anonymisation and

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Delta-presence perform worse accordingly. This indicates that both initial *k*-anonymisation and the size of the dataset are important factors in preventing data re-identification.

From the perspective of the identifiability score, more data records and diverse set of attribute combinations in Dataset B can also lead to more matches between the synthetic dataset and the original data records. Consequently, the re-identification risk for certain records in the synthetic dataset increases, causing higher identifiability scores. To further explain this from the perspective of calculation methods, Delta-presence primarily focuses on the maximum re-identification risk, that is the maximum probability that a particular record appears in the synthetic dataset (Nergiz & Clifton, 2010). This means that as long as the maximum re-identification probability is low, delta-presence performs well. Identifiability score, on the other hand, is based on the matching level between the synthetic data and the real data, calculating how many records in the real dataset are uniquely matched (Nergiz & Clifton, 2010). This calculation method is more sensitive to the number of matches in the synthetic dataset. Even if the maximum re-identification risk is low, if the total number of uniquely matched records is high, the identifiability score will still be high.

DISCUSSION AND CONCLUSION

We now turn to the discussion of our findings and their implications for practice.

RQ1. First, regarding the capacity of SDG techniques to prevent privacy threats in tabular education datasets (RQ1), the results show that, overall, private SDGs outperform non-private SDGs. This finding aligns with prior results on benchmark datasets (Stadler et al., 2022). It suggests that when publishing or sharing educational datasets requires a higher level of privacy protection, private SDGs should be used.

While private SDGs generally outperform non-private ones in terms of privacy protection, they are not without issues. As noted before (Section “Synthetic data generation methods”), outlier data points are more vulnerable to linkage attacks (Stadler et al., 2022), and private SDGs are largely immune to linkage attacks due to their integration of DP. However, in some cases, the synthetic data generation process may lose information about outliers, especially when outliers are exceptionally uncommon in the dataset, causing the model to either overlook them or fail to accurately capture their characteristics (Jordon, Szpruch, et al., 2022). Therefore, although outliers in private SDGs are protected, there is the drawback that they may not be part of the synthetic dataset. This further indicates that the future development of synthetic data should focus on better capturing outliers distribution.

Dataset B had already undergone *k*-anonymisation, which involved some suppression of outliers (Kuzilek et al., 2017). This to some extent made it easier for the synthetic data to mimic the structure of the original dataset and may explain Dataset B's strong performance in terms of utility. However, *k*-anonymisation before synthetic data generation may have a negative impact on tasks requiring high-precision analysis. For tasks such as anomaly detection or precise individual behaviour modelling, the *k*-anonymised data had already lost some valuable information. When generating synthetic data from such a dataset, the model may struggle to capture the complex patterns necessary for these tasks, thereby reducing the overall utility of the data. For practitioners looking to release or use publicly available education datasets in the future, if their intended tasks involve high-precision analysis, they should be cautious when using datasets that have undergone private SDG or a combination of multiple privacy-preserving techniques.

Finally, our experimental results suggest that multiple privacy metrics, which evaluate privacy from various angles, should be employed to obtain a holistic evaluation of SDGs' ability to mitigate privacy threats. As evident in the reported privacy results, due to focus on

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

different aspects, k-anonymisation and delta-presence demonstrate stronger privacy protection for larger-scale datasets, whereas the identifiability score shows better performance in smaller-scale datasets. Similar discrepancies in privacy risk assessments across synthetic dataset sizes—stemming from differences in evaluation metrics—were also observed by Sundaram Muthu et al. (2024). From the perspective of inference attacks on synthetic data, they found that smaller synthetic datasets reduce the likelihood of successful inference attacks, thereby enhancing privacy. However, they also acknowledge that focusing solely on inference attacks is insufficient. Instead, a broader range of scenarios and privacy requirements should be considered for a more comprehensive and robust understanding of privacy leakage risks (Sundaram Muthu et al., 2024).

RQ2. Regarding the effect of SDGs on the utility of the synthetic educational dataset (RQ2), the results show that while non-private SDGs may be inferior to private SDGs in terms of privacy, they generally perform better in utility. Still, private SDGs, which tend to have a larger trade-off between privacy and utility, show relatively balanced results, as demonstrated in the performance of PATEGAN at ($\varepsilon = 10$). The relatively balanced performance of PATEGAN in terms of privacy and data utility is also consistent with the experimental results of Rosenblatt et al. (2020).

Overall, the results of RQ2 suggest that SDGs still face a privacy-utility trade-off, and no single SDG outperforms all others across all metrics. For practitioners, this finding suggests that SDGs should be evaluated based on specific educational scenarios and needs. In other words, practitioners should avoid relying on a single model to meet all objectives. Instead, they should thoroughly evaluate different models based on their specific requirements and priorities—whether privacy is the main focus or finding the best balance is more important—when selecting the most suitable SDG.

Implications

Taken together, the results of RQ1 and RQ2, have the following implications:

1. *Reducing students' fear of external data breaches and fostering their trust in sharing personal information*. Research has shown that students have low trust in external institutions, and their willingness to share personal information significantly decreases when it involves technology companies or third-party service providers (Slade et al., 2019). Tsai et al. (2020) also noted that students have the lowest trust in external parties and are most reluctant to share their personal data with them. The strong privacy performance of private SDGs, as shown in this paper, offers greater security for educational institutions when collaborating with external parties and may help ease students' privacy concerns and foster trust.
2. *Providing a secure educational analysis environment*. The privacy-preserving capabilities demonstrated by private SDGs in response to RQ1 allow for setting up a secure environment for educational analysis, reducing the risk of sensitive data breaches. This enables more educational institutions to conduct data-driven decision-making and analyses with less concern about exposing confidential information.
3. *Promoting data sharing and collaboration, and advancing educational technology*. Educational researchers and practitioners can use both non-private and private synthetic data to publish tabular datasets without compromising student privacy, while also enhancing robustness against potential attacks. This facilitates the growth of open science, making large-scale, high-quality and diverse datasets accessible to a wider audience, which promotes educational research. Furthermore, the development of educational technology relies heavily on student data for development and optimisation (Buu & Karunaratne, 2024).

The widespread adoption of SDGs allows practitioners to access publicly available high-quality datasets and generate their own synthetic data, reducing the need for sensitive data during the development phase and accelerating technological innovation.

4. *Promoting student-centred data governance frameworks*. By improving the transparency and security of data sharing, these technologies can promote student-centred data governance practices while providing a strong technical foundation for developing responsible data usage policies.

Limitations

The limitations of this study are as follows: First, we only explored two typical educational datasets; using a broader range of datasets would provide more comprehensive insights. Second, due to the study's primary focus on privacy and limited space, we evaluated downstream machine learning performance using logistic regression only. A wider variety of machine learning tasks would offer a more complete utility evaluation. Third, we explored only a limited number of SDGs (four in total), and those were tested with default settings (ie, without optimisation) and only a limited variety of epsilon values, which may not reflect the full potential of these models under different configurations. Fourth, the study relied on limited measures of utility, namely different measures of fidelity, with application fidelity limited to one prediction metric. Evaluating the utility of private SDGs using a variety of predictive accuracy metrics is an important direction for advancing this research forward. Finally, the computational cost of this study is relatively high, especially for the larger Dataset B. Generating and evaluating private SDGs required 6 hours of computation on an advanced A100 GPU. Such high computational demands may limit the practical application of these methods in academic institutions with limited computational resources.

Future directions

For future direction, this paper encourages more experimentation with synthetic data in the education field to overcome the aforementioned limitation. Additionally, we recommend that practitioners document their specific needs and the evaluation thresholds they set during these experiments to provide valuable references for others. Since the thresholds for evaluation metrics largely depend on the context, making a general recommendation is difficult. However, if documented and shared, practices of previous users can serve as useful, evidence-based guidance for those new to the task. This approach can help minimise situations where, despite good evaluation results, practical issues in specific contexts remain unresolved due to differences in use cases.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest related to this study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Students Performance in Exams at https://www.kaggle.com/datasets/spscientist/students-performance-in-exams.

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ethics Statement

This study did not involve human participants, animals, or personal data requiring ethical approval. All research procedures adhered to standard ethical guidelines in accordance with institutional and international best practices.

ORCID

Qinyi Liu https://orcid.org/0009-0003-4973-0901
Mohammad Khalil https://orcid.org/0000-0002-6860-4404
Javier de la Hoz-Ruiz https://orcid.org/0000-0001-7670-5662

Endnote

[1] A term often used in the literature on synthetic data, which refers to the evaluation of the overall statistical distribution of synthetic data.

REFERENCES

Alhakbani, H. A., & Alnassar, F. M. (2022). Open learning analytics: A systematic review of benchmark studies using Open University learning analytics dataset (OULAD). In *2022 7th International Conference on Machine Learning Technologies (ICMLT) (ICMLT 2022), March 11–13, 2022, Rome, Italy* (p. 6). ACM. https://doi.org/10.1145/3529399.3529413

Ankan, A., & Panda, A. (2015). pgmpy: Probabilistic graphical models using python. In *Proceedings of the Python in Science Conferences (SciPy 2015), July 6-12, 2015, Austin, Texas, USA*. https://doi.org/10.25080/majora-7b98e3ed-001

Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., & Beyerer, J. (2022). Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, *12*(23), 12320. https://doi.org/10.3390/app122312320

Arnold, C., Liu, X., Pouyanfar, S., de Leon, E., Desai, A., & Allen, J. (2020). *Really useful synthetic data a framework to evaluate the quality of differentially private synthetic data summary for EcoPaDL at ICML 2020*. https://arxiv.org/pdf/2004.07740

Bautista, P., & Inventado, P. S. (2021). *Protecting student privacy with synthetic data from generative adversarial networks* (pp. 66–70). Springer EBooks. https://doi.org/10.1007/978-3-030-78270-2_11

Buu, N., & Karunaratne, T. (2024). Learning analytics with small datasets—State of the art and beyond. *Education Sciences*, *14*(6), 608. https://doi.org/10.3390/educsci14060608

Cong, Y., Zhao, M., Li, J., Wang, S., & Carin, L. (2020). GAN memory with no forgetting. *ArXiv.org*. https://doi.org/10.48550/arXiv.2006.07543

Dankar, F. K., Ibrahim, M. K., & Ismail, L. (2022). A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, *10*(11147), 11158. https://doi.org/10.1109/access.2022.3144765

Dwork, C. (2006). Differential privacy. *International Colloquium on Automata, Languages, and Programming*, *4052*, 1–12. https://doi.org/10.1007/11787006_1

El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation* [Book]. In www.oreilly.com. O'Reilly Media, Inc. https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733. https://doi.org/10.3390/math10152733

Google Cloud. (2023). *Computing δ-presence for a dataset | sensitive data protection documentation*. Google Cloud. https://cloud.google.com/sensitive-data-protection/docs/compute-d-presence

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2016). Privacy-preserving learning analytics: Challenges and techniques. *IEEE Transactions on Learning Technologies*, *1*, 68–81. https://research.sabanciuniv.edu/id/eprint/29682/1/TLT-CameraReady-Saygin.pdf

Hernadez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2023). Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods of Information in Medicine*, *62 Suppl 1*, e19–e38. https://doi.org/10.1055/s-0042-1760247

Hutt, S., Das, S., & Baker, R. (2023). *The right to be forgotten and educational data mining: challenges and paths forward*. https://learninganalytics.upenn.edu/ryanbaker/EDM23_RightToBeForgotten_Camera.pdf

Hutt, S., Baker, R. S., Ashenafi, M. M., Andres-Bray, J. M., & Brooks, C. (2022). Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology*, *53*(4), 756–775. https://doi.org/10.1111/bjet.13231

Jordon, J., Yoon, J., & van der Schaar, M. (2022). *PATE-GAN: Generating synthetic data with differential privacy guarantees*. Openreview.net. https://openreview.net/forum?id=S1zk9iRqF7

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S., & Weller, A. (2022). *Synthetic data—What, why and how?* https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf

Kaabachi, B., Espraz, J., Meurers, T., Otte, K., Halilovic, M., Prasser, F., & Raisaro, J. L. (2023). *Can we trust synthetic data in medicine? A scoping review of privacy and utility metrics*. *MedRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.11.28.23299124

Keierleber, M. (2022). *After huge illuminate data breach, Ed Tech's "student privacy pledge" under fire*. https://www.the74million.org/article/after-huge-illuminate-data-breach-ed-techs-student-privacy-pledge-under-fire/

Khalil, M., & Ebner, M. (2016). De-identification in learning analytics. *Journal of Learning Analytics*, *3*(1), 129–138. https://doi.org/10.18608/jla.2016.31.8

Koenecke, A., & Varian, H. (2020). *Synthetic data generation for economists*. https://arxiv.org/pdf/2011.01374.pdf

Krishnamurthi, S., Lerner, B. S., & Politz, J. G. (2017). *4 Introduction to Tabular Data*. Papl.cs.brown.edu. https://papl.cs.brown.edu/2016/intro-tabular-data.html

Kurakin, A., & Ponomareva, N. (2023). *Protecting users with differentially private synthetic training data*. Research.google. https://research.google/blog/protecting-users-with-differentially-private-synthetic-training-data/

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University learning analytics dataset. *Scientific Data*, *4*, 170171. https://doi.org/10.1038/sdata.2017.171

Ladjal, D., Joksimović, S., Rakotoarivelo, T., & Zhan, C. (2022). Technological frameworks on ethical and trustworthy learning analytics. *British Journal of Educational Technology*, *53*(4), 733–736. https://doi.org/10.1111/bjet.13236

Li, N., Lyu, M., Su, D., & Yang, W. (2017). Differential privacy. In *Synthesis lectures on information security, privacy, and trust*. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02350-7

Lin, Z., Gopi, S., Kulkarni, J., Nori, H., & Yekhanin, S. (2024). *Differentially private synthetic data via foundation model APIS 1: Images*. https://arxiv.org/pdf/2305.15560

Liu, Q., & Khalil, M. (2023). Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology*, *54*(1715), 1747. https://doi.org/10.1111/bjet.13388

Liu, Q., Khalil, M., Jovanovic, J., & Shakya, R. (2024). Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK'24), March 18–22, 2024, Kyoto, Japan* (p. 12). ACM. https://doi.org/10.1145/3636555.3636921

Liu, Q., Shakya, R., Khalil, M., & Jovanovic, J. (2025). Advancing privacy in learning analytics using differential privacy. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK 2025), Dublin, Ireland, March 3–7, 2025* (pp. 1–11). ACM. https://doi.org/10.1145/3706468.3706493

Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2021). *Machine learning for synthetic data generation: A review*. *14*(8). https://arxiv.org/pdf/2302.04062

Nergiz, M. E., & Clifton, C. (2010). δ-Presence without complete world knowledge. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 868–883. https://doi.org/10.1109/tkde.2009.125

Peek, S. (2023). *Data complexity rise means for business*. Business.com. https://www.business.com/articles/what-the-rise-in-data-complexity-means-for-business-departments/

Prinsloo, P., Slade, S., & Khalil, M. (2019). Student data privacy in MOOCs: A sentiment analysis. *Distance Education*, *40*(3), 395–413. https://doi.org/10.1080/01587919.2019.1632171

Qian, Z., Cebere, B.-C., & van der Mihaela, S. (2023). Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2301.07573

Rosenblatt, L., Liu, X., Pouyanfar, S., de Leon, E., Desai, A., & Allen, J. (2020). Differentially private synthetic data: Applied evaluations and enhancements. *ArXiv.org*. https://doi.org/10.48550/arXiv.2011.05537

Slade, S., Prinsloo, P., & Khalil, M. (2019). Learning analytics at the intersections of student trust, disclosure and benefit. In The 9th International Learning Analytics & Knowledge Conference (LAK19), March, 2019, Tempe, AZ, USA. ACM, New York, NY, USA. Article 4, 10 pages https://doi.org/10.1145/3303772.3303796

Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data—Anonymisation groundhog day. In Proceedings of the 31st USENIX Security Symposium (pp. 1450–1464). USENIX Association, August 10–12, 2022, Boston, MA, USA. https://www.usenix.org/system/files/sec22-stadler.pdf

Sundaram Muthu, M., Annamalai, S., Gadotti, A., Rocher, L., Sundaram, M., & Annamalai, M. (2024). *A linear reconstruction approach for attribute inference attacks against synthetic data*. https://www.usenix.org/system/files/usenixsecurity24-annamalai-linear.pdf

Tsai, Y.-. S., Whitelock-Wainwright, A., & Gasevic, D. (2020). The privacy paradox and its implications for learning analytics. In LAK '20: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (pp. 230–239), March 23–27, 2020, Frankfurt, Germany. https://doi.org/10.1145/3375462.3375536

Van Breugel, B., Sun, H., Qian, Z., & Van Der Schaar, M. (2023). *Membership inference attacks against synthetic data through overfitting detection*. *206*. https://arxiv.org/pdf/2302.12580

Van den Bossche, T. V. (2023). *Methods for dataset de-identification Thomas Van den Bossche*. https://www.msec.be/GDPR/presentaties/UG1-MSEC.pdf

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially private generative adversarial network. *ArXiv.org*. https://doi.org/10.48550/arXiv.1802.06739

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *ArXiv:1907.00503 [Cs, Stat]*. https://arxiv.org/abs/1907.00503

Yacobson, E., Fuhrman, O., Hershkovitz, S., & Alexandron, G. (2021). De-identification is insufficient to protect student privacy, or—What can a field trip reveal? *Journal of Learning Analytics*, *8*(2), 83–92. https://doi.org/10.18608/jla.2021.7353

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244–255. https://doi.org/10.1016/j.neucom.2019.12.136

Yoon, J., Drumright, L. N., & van der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, *24*(8), 2378–2388. https://doi.org/10.1109/jbhi.2020.2980262

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). PrivBayes. *ACM Transactions on Database Systems*, *42*(4), 1–41. https://doi.org/10.1145/3134428

Zhao, Z., Kunar, A., Birke, R., & Chen, L. (2021). CTAB-GAN: Effective table data synthesizing. *Proceedings of Machine Learning Research*, *157*, 97–112. https://proceedings.mlr.press/v157/zhao21a/zhao21a.pdf

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, Q., Shakya, R., Jovanovic, J., Khalil, M., & de la Hoz-Ruiz, J. (2025). Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology*, *00*, 1–21. https://doi.org/10.1111/bjet.13576

14678535, 0, Downloaded from https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13576 by Humboldt-Universitat Zu Berlin, Wiley Online Library on [04/04/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License