Apr, 2026
GRINS DISCUSSION PAPER SERIES DP N° 99/2026 ISSN 3035-5567

Synthetic data for academic research: more opportunities than challenges

DP N° 99/2026

Authors:
Vincenzo Atella, Piera Bello, Cristina Davino, Rosa Fabbricatore, Margherita Fort, Goncalo Da Silva Lima, Costanza Marconi, Federica Origo, Elena Pisanelli, Veronica Rattini, Daniela Vuri

Apr, 2026
GRINS DISCUSSION PAPER SERIES DP N° 99/2026 ISSN 3035-5567

Synthetic data for academic research: more opportunities than challenges

Vincenzo Atella, Piera Bello, Cristina Davino, Rosa Fabbricatore, Margherita Fort, Goncalo Da Silva Lima, Costanza Marconi, Federica Origo, Elena Pisanelli, Veronica Rattini, Daniela Vuri

KEYWORDS

Privacy-preserving technologies | Synthetic Data | Peers | Drop- out

JEL CODE

C80, I21, C50, C60

ACKNOWLEDGEMENTS

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CITE THIS WORK

Author(s): Vincenzo Atella, Piera Bello, Cristina Davino, Rosa Fabbricatore, Margherita Fort, Goncalo Da Silva Lima, Costanza Marconi, Federica Origo, Elena Pisanelli, Veronica Rattini, Daniela Vuri. Title: Synthetic data for academic research: more opportunities than challenges. Publication Date: 2026.

A long-standing problem among researchers is the possibility of performing analyses on integrated granular data from different sources and to collaborate using them.This limitation comes from the implementation of GDPR-compliant procedures and the necessity of combining data based on the principles of transparency, fairness, and limited purpose set by GDPR.
This report presents results of pilot studies that rely on recently developed privacy-preserving technology (synthetic data) that can potentially overcome some of these limitations.
We discuss applications in education economics, comparing results of the same analysis across different case studies that can be scaled up.
We then performed a meta-analysis and illustrated how much the general public could learn from gaining access to such data in a common format.
We show that this technology can offer interesting opportunities but, at the same time, some challenges remain to be solved.

Synthetic data for academic research: more opportunities than challenges *

V. Atella* P. Bello § C. Davino† R. Fabbricatore†
M. Fort‡ G. Lima ‡ C. Marconi § F. Origo §
E. Pisanelli § V. Rattini ‡ D.Vuri*

April 13, 2026

Abstract

A long-standing problem among researchers is the possibility of performing analyses on integrated granular data from different sources and to collaborate using them. This limitation comes from the implementation of GDPR compliant procedures and the necessity of combining data based on the principles of transparency, fairness, and limited purpose set by GDPR. This report presents results of pilot studies that rely on recently developed privacy-preserving technology (synthetic data) that can potentially overcome some of these limitations. We discuss applications in education economics, comparing results of the same analysis across different case studies that can be scaled-up. We then performed a meta-analysis and illustrated how much the general public could learn from gaining access to such data in a common format. We show that this technology can offer interesting opportunities but, at the same time, some challenges remain to be solved. JEL: C80, I21, C50, C60

Keywords: Privacy-preserving technologies, Synthetic Data, Peers, Dropout.

*Funding by the European Union - NextGenerationEU - GRINS project Mission 4, Component 2, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP J33C22002910001) is gratefully ackwnoledged. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. We also wish to thank Lorenzo Rocco, Fabio Bravo, Sebastiano Saccani, for valuable comments and suggestions on earlier versions of this work. All errors are ours. Author Contributions, in alphabetic order:(i) project conceptualization and project work administration: V.A., M.F, F.O., V.R., D.V.; (ii) conceptualization of the empirical examples: C. D., R.F., M.F, G.L., F.O., V.R., ; (iii) data management and data analysis: R.F, G.L., C.M., F.O., E.P., V.R., D.V; (iv) draft preparation: writing, review and editing: V.A., P.B., R.F., C.D., M.F, G.L, C.M., F.O., E.P., V.R, D.V.; (v) funding acquisition: GRINS PE00000018 – CUP J33C22002910001. Spoke 0, Spoke 3.

* University of Roma Tor Vergata, Dept. of Economics
† University Federico II, Dept. of Economics
‡ University of Bologna, Dept. of Economics
§ University of Bergamo, Dept. of Economics

1 Introduction

Data have become a fundamental resource in the modern world, playing an essential role in business, scientific research, and decision-making (Scassola et al., 2025). Across high-impact sectors—such as healthcare (Appenzeller et al., 2022; Gonzales et al., 2023; Giuffrè and Shung, 2023), finance (Assefa et al., 2020), and education (Bonnéry et al., 2019)—granular, individual-level data underpin the analyses that inform policy choices, guide investment decisions, and advance scientific knowledge.

However, a long-standing challenge among European researchers is the limited ability to perform analyses on integrated granular data from multiple sources and to collaborate using such data. Privacy concerns often restrict access to these data: sensitive information about individuals, households, or firms cannot be shared openly (Abdelhameed et al., 2018), and even when statistical disclosure limitation methods are applied, the data may still be vulnerable to re-identification attacks (Cheng et al., 2017) or suffer from significantly reduced analytical utility (O'Keefe and Rubin, 2015). In Europe, this limitation is also due to the need to comply with the General Data Protection Regulation (GDPR), which mandates strict principles of transparency, fairness, and delimited purpose when handling personal data.

In this context, we present the results of pilot studies conducted using the synthetic data technology developed by AINDO. This privacy-preserving approach addresses key GDPR compliance challenges by enabling secure, utility-preserving data sharing. Synthetic data offers a principled solution to this tension. Formally, synthetic data are data that have been generated using a purpose-built mathematical model or algorithm - the generator - with the aim of solving one or more data-science tasks (Jordon et al., 2022). A generator is trained on a real (confidential) dataset and learns its statistical structure; it then produces an entirely new dataset that preserves the distributional and relational properties of the original one, but it contains no actual records from it. When properly generated, synthetic data preserve the analytic utility of real data while mitigating the risk that any specific individual can be identified (Plasencia Palacios et al., 2025). This makes synthetic data an increasingly important *privacy-enhancing technology* (PET) (El Emam, 2020; Jordon et al., 2022; Stadler et al., 2022).

The practical value of synthetic data extends well beyond privacy protection. Synthetic datasets can fill gaps in data availability for counterfactual research and agent-based simulations (Fagiolo et al., 2019), support synthetic control methods in causal inference (Abadie et al., 2015), augment scarce data in domains where collection is expensive (Appenzeller et al., 2022), and enable the construction of predictive models using machine learning on data that would otherwise remain locked away (Shwartz-Ziv and Armon, 2022; Goncalves et al.,

2020). In development economics and official statistics, where microdata from surveys and censuses often contain sensitive personal information, synthetic data can unlock the information content of administrative records for researchers and policymakers without compromising respondent confidentiality (Solatorio and Dupriez, 2023).

The field of synthetic data generation has seen rapid methodological development in recent years (Hernandez et al., 2022; Figueira and Vaz, 2022; Gupta et al., 2016). Early approaches relied on Bayesian networks (Zhang et al., 2017), factor graphs (McKenna et al., 2019), and sequential regression models such as the CART-based methods implemented in the widely used `synthpop` package (Nowok et al., 2016). These classical methods remain popular in official statistics and survey applications, but they impose strong parametric assumptions that limit their ability to capture complex nonlinear dependencies.

The breakthroughs of deep learning have opened new avenues. Generative adversarial networks (GANs) were among the first architectures applied to tabular data (Park et al., 2018; Xu et al., 2019), followed by variational autoencoders (VAEs) (Kingma and Welling, 2014; Xu et al., 2019) and, more recently, by language-model-style autoregressive transformers (Borisov et al., 2022b; Solatorio and Dupriez, 2023) and diffusion models (Kotelnikov et al., 2023; Ho et al., 2020). These approaches have considerably expanded the expressiveness of synthetic data generators, though a comprehensive survey by Borisov et al. (2022a) shows that the relative performance of different architectures remains highly task- and dataset-specific.

However, most real-world data are not stored as a single flat table but rather as *relational databases* - collections of multiple tables linked through foreign-key relationships (Solatorio and Dupriez, 2023; Scassola et al., 2025). Household surveys with separate person- and household-level files, employer–employee linked administrative records, and transactional databases with customer, order, and product tables are all examples of relational data. Despite the prevalence of this format, the vast majority of synthetic data research has focused on generating a single independent table (Borisov et al., 2022a; Hudovernik et al., 2024).

Modeling relational data is substantially more challenging because a generator must capture not only the marginal and joint distributions within each table, but also the dependencies and cardinality constraints *across* tables - for example, the number of child records per parent, or the correlation between attributes in linked rows (Solatorio and Dupriez, 2023). The foundations of relational synthetic data generation were laid by the Synthetic Data Vault (Patki et al., 2016), which uses Gaussian copulas to model multi-table structures. More recently, a variety of deep learning methods have been proposed: GAN-based approaches such as Row Conditional-TGAN (Gueye et al., 2023) and the Incremental Relational Generator (Li and Tay, 2023); transformer-based methods,

most notably REaLTabFormer, which employs a GPT-2 autoregressive backbone for parent tables and a sequence-to-sequence architecture for child tables (Solatorio and Dupriez, 2023); graph-variational autoencoders (Mami et al., 2022); and diffusion-based approaches such as ClavaDDPM (Pang et al., 2024). A particularly recent advance is the graph-conditional flow matching framework of Scassola et al. (2025), which recasts relational data generation as a continuous normalizing flow problem and uses graph neural networks (Scarselli et al., 2009; Battaglia et al., 2018) to propagate information across arbitrarily connected records, achieving state-of-the-art fidelity on several benchmark datasets.

Despite this rapid progress, a systematic empirical comparison by Hudovernik et al. (2024) - the first comprehensive benchmark of relational data synthesis methods - reveals that no current method is able to produce synthetic relational data that is indistinguishable from the original. For downstream predictive utility, the correlation between models trained on real versus synthetic data remains only moderate. These findings underscore that relational data synthesis, while rapidly advancing, remains a substantially harder problem than single-table generation, and that standardized evaluation frameworks are essential for tracking progress (Hudovernik et al., 2024).

On the privacy side, synthetic data represent a good solution, yet the concept of "data privacy" remains elusive and difficult to quantify (Plasencia Palacios et al., 2025). Classical synthesis methods do not, by themselves, provide formal guarantees against re-identification or attribute disclosure. Different privacy mechanisms (Ji et al., 2014) offer mathematically rigorous bounds on privacy loss, but the noise they inject typically degrades analytical utility (Stadler et al., 2022). Generative models trained on real data may also memorize and reproduce individual records - a phenomenon known as data copying - which must be explicitly prevented through techniques such as target masking (Solatorio and Dupriez, 2023) or monitored through distance-to-closest-record metrics (Meehan et al., 2020).

A key challenge for practitioners is the absence of a universal standard for quantifying privacy risk. Privacy is an interdisciplinary concept that spans legal, technical, and ethical dimensions (Plasencia Palacios et al., 2025), and the practical difficulties of meeting regulatory requirements - such as those imposed by the European Union's General Data Protection Regulation (GDPR) - can limit innovation in data-driven fields (Bellomarini et al., 2022; Giomi et al., 2022). The GDPR's reasonableness test stipulates that data may be considered anonymous only when re-identification is infeasible given all means "reasonably likely" to be employed, a standard that requires continuous re-evaluation as technology evolves (Plasencia Palacios et al., 2025).

Recent work has begun to address this gap. Plasencia Palacios et al. (2025) propose a benchmarking framework for empirically assessing privacy quantifi-

cation methods through the controlled insertion of risk into synthetic datasets. Their experiments show that most commonly used privacy metrics consistently reflect controlled risks in both idealized and practical settings, though data characteristics - such as a high proportion of categorical attributes - can amplify measured privacy risk. By linking quantitative privacy metrics to GDPR-derived legal standards, this framework contributes to the standardization of privacy assessment that the field urgently needs (Boudewijn et al., 2023; Hittmeir et al., 2020; Ganev et al., 2024).

This report lies at the intersection of the developments described above. We present the results of pilot studies conducted using the synthetic data technology developed by AINDO to generate synthetic versions of university administrative datasets. Our goal is to demonstrate that privacy-preserving synthetic data can enable granular econometric analysis - using regression methods on both original and synthetic datasets - while complying with GDPR requirements and allowing cross-institutional collaboration.

We replicate the analysis on data from four different universities that differ along several dimensions relevant for these applications (e.g., underlying raw data distribution and features of the hardware used to generate synthetic data), and we perform a meta-analysis to illustrate the value that the general public could gain from access to such data in a common format.

We consider a cross-institutional study that enables a meta-analytic assessment of fidelity and utility of the synthetic data privacy preserving technology in a frontier setting, i.e. relational synthetic data aimed at capturing intra-table and cross-table depdendencies within a single modelling pipeline, and with applications considering different empirical analysis tools that are conventionally adopted in economics and statistics. In doing this, we depart from the literature that typically considers single tables and single case studies. We find that that synthetic data generally replicate very well marginal distributions, across all settings. When we look into specific applications, synthetic data generated starting from integrated data base replicate results obtained on original data when applications do not focus on tail behaviours/rare events, with limited role played by differences in hardware settings.[1] The performance of synthetic data generated through federated learning relative to original data shows larger (in absolute value) deviations.

A notable feature of the case studies considered in this report, is that these can later be eventually scaled-up.

In what follows, the paper is organized in 5 sections. Section 2 presents a comprehensive review of the literature aimed at discussing recent evolution in a highly dynamic field of research. Section 3 describes how synthetic data were

[1] The latter statement should be interpreted with caution as the variability in hardware we have is limited to the distinct number of institutions participating to the study.

generated relative to technical specifications (type of hardware and software used) and discusses the indicators used by AINDO for its reporting. In section 4 we present original empirical contributions based on real and synthetic data, including within- and between-university comparisons across four institutions from the GRINS network, that cover different geographic areas (North, Center, South) in Italy, as well as overall number of enrolled students per year. For each application, we provide a description of the dataset, descriptive statistics comparing the original and synthetic samples, and validation of the synthetic data on marginal distributions. The main results are illustrated through a set of tables and figures, and a meta-analysis perspective is adopted to assess the consistency of the findings between universities. In Section 5 we acknowledge that this study is subject to a number of limitations and we conclude by highlighting that the use of synthetic data is a viable alternative to real data in empirical research, while preserving the key statistical properties of the original samples.

2 Literature Review

The growing reliance of empirical economics on granular micro-data - covering individuals, households, firms, and linked administrative records - has intensified the tension between data access and confidentiality protection. This section provides a comprehensive review of the literature on synthetic data as a statistical disclosure limitation strategy, examining its foundations, methods, evaluation frameworks, and implications for economic research. We organize the discussion around eight interconnected themes: the theoretical and methodological foundations of synthetic micro-data (Section 2.1); inferential validity (Section 2.2); predictive utility (Section 2.3); distributional fidelity (Section 2.4); policy evaluation and micro-simulation (Section 2.5); privacy guaranties and differential privacy (Section 2.6); complex survey design considerations (Section 2.7); the generation of relational and multi-table synthetic data (Section 2.8); and the evolution of generative methods from classical approaches to modern deep learning architectures (Section 2.9). We conclude by identifying the main gaps in the current evidence base and lining directions for future research (Section 2.10).

2.1 Theoretical and Methodological Foundations

The conceptual architecture of synthetic microdata is based on a distinction between two fundamental paradigms. In *fully synthetic* data generation, all records and variables in the released data set are drawn from statistical models fitted on the confidential microdata, so that no original observation appears in the public-use file (Rubin, 1993; Little, 1993). In *partially synthetic* approaches,

the original sample structure is retained, but sensitive variables are replaced with values imputed from predictive models conditioned on non-sensitive covariates (Reiter, 2003). This distinction is not merely taxonomic: it determines the choice of variance estimators, the interpretation of inferential results, and the applicability of disclosure control strategies.

The theoretical basis for fully synthetic data was laid by Rubin (1993), who proposed treating synthetic datasets as a form of multiple imputation applied to the entire population rather than to missing observations alone. Rubin's framework exploited the established inferential machinery of multiple imputation (Rubin, 1996), but required important adaptations. Raghunathan et al. (2001) developed the sequential regression multivariate imputation (SRMI) framework, which provided a computationally tractable approach to generate synthetic records by fitting a sequence of conditional models, one for each variable given all the others. This approach became the workhorse of many subsequent implementations.

Reiter (2003) provided the foundational inference framework for partially synthetic data, combining rules for point estimates and variance estimators that account for the additional uncertainty introduced by the synthesis process. A key result was that standard multiple imputation rules that combine (Rubin, 1996) cannot be applied directly to fully synthetic datasets; instead, specialized estimators, designated as the T_s family, are required for valid inference under large-sample conditions (Reiter, 2005). These theoretical advances were empirically tested on household survey data, including the Survey of Income and Program Participation (SIPP) and the American Community Survey (ACS), and established the feasibility of releasing synthetic public-use files that approximate the analytical results of confidential originals (Abowd et al., 2006).

Drechsler (2011) consolidated much of this foundational work, providing a unified treatment of statistical theory, practical implementation strategies, and evaluation protocols for fully and partially synthetic data. Drechsler's contribution was instrumental in framing synthetic data as a legitimate and practically viable tool for statistical disclosure control, shifting the field from purely theoretical proposals toward operational implementations. The complementary work of Drechsler and Reiter (2010) on sampling-with-synthesis further demonstrated that hybrid strategies - combining synthetic data generation with complex sampling designs - can enhance both utility and protection.

2.2 Inferential Validity

For economists, the credibility of synthetic data ultimately hinges on whether standard analytical outputs - regression coefficients, standard errors, confidence intervals, and hypothesis tests - derived from synthetic datasets approximate those obtained from the confidential originals with acceptable accuracy.

The most extensive body of evidence on inferential validity comes from the *SIPP Synthetic Beta* (SSB), a multiply-imputed synthetic version of the SIPP linked to Social Security Administration earnings records. Comparisons between the SSB and the *SIPP Gold Standard File* (GSF) have been conducted in multiple studies (Abowd and Stinson, 2013; Benedetto et al., 2013; Abowd et al., 2006). The general finding is that the regression coefficients estimated from the SSB replicate the direction and approximate magnitude of the original estimates, with modest discrepancies for descriptive statistics and regression-based measures of central tendency. However, discrepancies become material for estimands that depend on higher-order moments, tail behavior, or dynamic structures - such as income volatility, transitory earnings components, and inequality decompositions (Abowd and Vilhuber, 2005).

ACS-based evaluations confirm these patterns in a cross-sectional setting: many regression coefficients estimated from synthetic data fall within one original standard error of confidential estimates (Drechsler and Reiter, 2010). However, estimates for small subgroups, rare events, or high-order interaction terms can diverge substantially, underscoring the sensitivity of inferential quality to both the specification of the synthesis model and the nature of the target estimand.

In establishment and business datasets, edit-imputation methods combined with synthesis yield regression coefficients broadly consistent with those from confidential data (Kinney et al., 2011; Drechsler and Vilhuber, 2014). The Synthetic Longitudinal Business Database (SynLBD), developed for the U.S. Census Bureau, demonstrated that synthetic firm-level data could support a range of standard empirical analyses (Kinney et al., 2011). Nevertheless, synthesis tends to attenuate coefficients associated with rare categories or extreme values, a limitation of particular concern for research on productivity heterogeneity, firm size distributions, and market concentration.

Overall, the literature suggests that inferential utility is strongest for large-sample, central-tendency measures and weakest for tail-sensitive or dynamically structured estimands - precisely the types of outcomes often central to contemporary economic analysis of inequality, volatility, and intergenerational mobility.

2.3 Predictive Utility

Beyond regression-based inference, synthetic data is increasingly evaluated through the lens of predictive performance, a dimension of particular relevance for economic applications in credit scoring, labor market transition forecasting, firm growth prediction, and machine learning–augmented policy evaluation.

Classical synthesis approaches, including CART-based sequential methods and SRMI, generally reproduce predictive patterns with reasonable fidelity.[2]

[2] CART (Classification and Regression Trees) is a binary decision tree algorithm that

The `synthpop` package developed by Nowok et al. (2016) provides a flexible implementation of CART-based synthesis in R and has been widely adopted for both research and official statistics applications. Systematic evaluations by Snoke et al. (2018) demonstrated that CART-based synthetic data preserves the predictive structure of the original data for standard classification and regression tasks, with utility measures such as propensity score mean-squared error (pMSE) indicating high global similarity.

Deep learning–based generative models, including the Conditional Tabular GAN (CTGAN; Xu et al., 2019), the Tabular Variational Autoencoder (TVAE; Xu et al., 2019), and the Synthetic Data Vault (SDV; Patki et al., 2016), show promise in capturing complex nonlinear dependence structures. The broader landscape of deep generative models for tabular data has been surveyed comprehensively by Borisov et al. (2022a), who identify GANs, VAEs, and diffusion-based approaches as the three dominant architectural families. More recent contributions include diffusion-based models for tabular data (Kotelnikov et al., 2023), which have shown competitive performance on specific benchmark tasks. Transformer-based autoregressive models have also emerged as a competitive alternative: Solatorio and Dupriez (2023) introduce REaLTabFormer, which employs a GPT-2 backbone and achieves state-of-the-art predictive performance on large non-relational tabular datasets without requiring task-specific fine-tuning.

However, head-to-head comparisons reveal a mixed picture. Although deep generative models sometimes achieve higher predictive accuracy in narrowly defined tasks, they often underperform classical methods in maintaining broad distributional properties or inference-relevant variance structures (Stadler et al., 2020; Jordon et al., 2018). Bowen and Liu (2020) conducted a comparative evaluation of differentially private synthetic data algorithms in the context of the NIST 2018 challenge, finding that no single method dominated in all metrics and that the generator choice was highly dependent on the data set.

A persistent and consequential gap in the literature is the absence of systematic train-on-synthetic, test-on-original (TOSO) evaluation protocols. Few studies rigorously assess whether predictive models fitted on synthetic data generalize to holdout samples from the original confidential data with comparable accuracy. A recent exception is the benchmarking study by Hudovernik et al. (2024), which systematically evaluates the utility of synthetic relational data us-

recursively partitions data into homogeneous subsets by learning splitting rules from the data. In the context of synthetic data generation, CART is applied *sequentially* to model the conditional distribution of each variable given the others, thereby producing synthetic samples that preserve the statistical structure of the original data. SRMI (Sequential Regression Multiple Imputation) is a related iterative procedure in which a sequence of regression models - each of which may be based on CART - is used to impute missing values variable by variable, cycling through the variables multiple times until convergence. Together, CART-based sequential methods and SRMI provide a flexible, non-parametric framework for handling complex data structures, accommodating both continuous and categorical variables without requiring strong distributional assumptions.

ing train-on-synthetic, evaluate-on-real protocols and reports only moderate correlation between predictive performance on real and synthetic data. Their findings reinforce the conclusion that predictive utility cannot be taken for granted - particularly for complex, multi-table datasets. This omission limits the evidence base for economists and policymakers seeking to deploy synthetic data in operationally predictive contexts - for instance, targeting social programs using models trained on publicly available synthetic survey data.

2.4 Distributional Fidelity

Distributional similarity between synthetic and original data is a critical concern in economics, where tail behavior often determines conclusions about inequality, poverty incidence, wealth concentration, and extreme event risk.

The methodological toolkit for assessing distributional fidelity has expanded considerably. Snoke et al. (2018) proposed a framework distinguishing between *general utility* measures - which assess global distributional similarity (e.g., pMSE, Kolmogorov-Smirnov statistics, correlation matrix comparisons) - and *specific utility* measures - which evaluate the quality of particular analytical outputs (e.g., regression coefficients, quantile estimates). This taxonomy has become a standard reference in the field.

Empirical results are generally encouraging for the central mass of distributions: means, medians, and interquartile ranges are typically well preserved across a range of synthesis methods and datasets (Nowok et al., 2016; Drechsler and Reiter, 2010). However, tails and rare events are consistently poorly replicated. The SSB, for instance, reproduces the central earnings distribution with reasonable accuracy but substantially overstates short-run earnings volatility and transitory inequality components (Benedetto et al., 2013; Abowd and Vilhuber, 2005). ACS-based evaluations similarly demonstrate good replication of common demographic distributions but degradation in the tails and for rare subgroups (Drechsler and Reiter, 2010).

For deep generative models, the evidence is mixed. Xu et al. (2019) report that CTGAN captures multimodal distributions more effectively than purely parametric approaches, but this advantage does not consistently extend to tail preservation.[3] Patki et al. (2016) introduce the Synthetic Data Vault (SDV) framework with automated evaluation metrics, though the emphasis is primarily on overall distributional similarity rather than tail-specific fidelity. Raab

[3] CTGAN (Conditional Tabular Generative Adversarial Network) is a state-of-the-art deep learning library for generating realistic synthetic data from single tables, effectively handling mixed-type variables (discrete and continuous) and imbalanced datasets. Developed for NeurIPS 2019, it employs a conditional generator to preserve complex statistical relationships, making it particularly well-suited for sharing sensitive data. CTGAN is part of the SDV (Synthetic Data Vault) ecosystem and is available as a Python package (`ctgan`) or as an R interface.

et al. (2020) provide a practical assessment of synthesis for large samples, noting that even well-calibrated synthesis methods produce discrepancies in joint distributions that accumulate as sample sizes grow.

Recent work on relational data synthesis has extended the distributional fidelity toolkit to multi-table settings. Hudovernik et al. (2024) combine statistical, distance-based, and detection-based fidelity metrics into an open-source benchmarking tool (SyntheRela) and show that no current method produces synthetic relational data that is indistinguishable from the original. Scassola et al. (2025) propose a graph-conditional flow matching approach that leverages graph neural networks to propagate information across related records, achieving state-of-the-art fidelity on several relational benchmarks. These results suggest that while single-table fidelity has matured, faithfully reproducing the joint distribution across linked tables remains an open challenge.

These distributional shortcomings are consequential for policy-relevant economic research. Studies of income inequality, poverty rates, wealth concentration, and tax incidence - where outcomes are highly sensitive to the behavior of extreme values - may yield misleading conclusions when based on synthetic data that systematically compresses or distorts the tails. Researchers must therefore exercise caution and, where possible, validate tail-sensitive findings against the confidential data through verification mechanisms.

2.5 Policy Evaluation

One of the most consequential applications of synthetic data in economics is tax-benefit microsimulation and policy evaluation. Microsimulation models, which apply policy rules to individual - or household - level data to estimate the distributional effects of reforms, are central to fiscal policy analysis in virtually every advanced economy. The availability of high-quality microdata is essential for these exercises, making synthetic public-use files an attractive complement to restricted-access administrative records.

Synthetic public-use files for tax data have demonstrated the ability to replicate aggregate policy outcomes while preserving confidentiality (Abowd and Schmutte, 2019). The SSB has been employed to study earnings dynamics, mobility, and inequality, with results that broadly reproduce the patterns observed in the confidential data (Benedetto et al., 2013). Evaluation evidence indicates that synthetic datasets replicate broad policy - relevant patterns - such as average tax burdens, benefit eligibility rates, and aggregate redistribution measures - with reasonable accuracy.

However, biases emerge in second-order outcomes that are critical for nuanced policy analysis. Volatility decompositions, inequality indices (particularly those sensitive to the tails, such as the Gini coefficient at high income levels or the P90/P10 ratio), and subgroup-specific eligibility estimates are subject to

distortion (Abowd and Vilhuber, 2005; Benedetto et al., 2013). These discrepancies are not trivial: a tax reform microsimulation that underestimates the concentration of income at the top will produce misleading estimates of revenue yield and distributional incidence.

An important institutional innovation that addresses these limitations is the *verification server* model, pioneered by Karr et al. (2006); Reiter (2005). In this framework, researchers conduct their analyses on the synthetic public-use data, then submit their regression specifications or statistical queries for validation against the confidential data. Differentially private responses provide calibrated feedback on the accuracy of the synthetic-data-based estimates, enhancing both user trust and analytical rigor. This hybrid approach - synthetic data for exploration and model development, complemented by secure verification mechanisms for policy-critical inference - represents a promising paradigm for integrating synthetic data into the policy analysis workflow.

A notable absence in the literature, however, is the systematic application of *policy-microsimulation parity tests*, in which identical tax-benefit simulation pipelines are run on both original and synthetic datasets and the resulting policy estimates are compared across the full distribution. Such tests would provide a direct and policy-relevant assessment of synthetic data quality that goes beyond the regression-centric evaluations that currently dominate the field.

2.6 Privacy Guarantees and Differential Privacy

Synthetic data is primarily motivated by the imperative of privacy protection, but classical synthesis methods alone do not provide formal guarantees against re-identification or attribute disclosure. This limitation has prompted growing adoption of *differential privacy* (DP) frameworks, which offer mathematically rigorous bounds on the additional information that an adversary can extract from the released data about any individual record (Abowd and Schmutte, 2019; Abowd et al., 2020).

The integration of differential privacy into synthetic data generation introduces a fundamental trade-off between privacy protection and analytical utility. Studies comparing DP-based generators to classical synthesis methods consistently document that stronger formal privacy guarantees come at a substantial cost to inferential and distributional quality (Stadler et al., 2020; Jordon et al., 2018; Bowen and Snoke, 2021). For instance, differentially private implementations of CTGAN produce datasets that provide robust protection against re-identification attacks but introduce noise that degrades regression coefficients, attenuates correlations, and reduces distributional fidelity - particularly in the tails.

Bowen and Liu (2020) conducted a systematic comparison of differentially private synthetic data algorithms submitted to the NIST 2018 challenge, find-

ing wide variation in the utility-privacy trade-off across methods and datasets. Their results underscore that the choice of privacy budget ε - which governs the permissible level of noise injection - has dramatic consequences for downstream analytical utility, with small ε values (strong privacy) rendering synthetic data nearly unusable for quantitative economic research.

Stadler et al. (2020) provide a particularly cautionary assessment, demonstrating that several purportedly privacy-preserving synthetic data methods fail to prevent attribute inference attacks in practice, a phenomenon they term "anonymisation groundhog day." Their findings highlight the gap between formal DP guarantees and the practical security of synthetic data releases, suggesting that reliance on synthesis alone - without rigorous adversarial evaluation - may provide a false sense of confidentiality protection.

More recently, Plasencia Palacios et al. (2025) propose a comprehensive benchmarking framework for empirically assessing the efficacy of privacy quantification methods applied to synthetic tabular data. Their approach is based on the controlled, deliberate insertion of risk - both in idealized settings (directly leaking proportions of real records) and in practical scenarios (generator overfitting and varying privacy budgets). They find that most commonly used privacy metrics consistently reflect controlled risks, although data characteristics such as a high proportion of categorical attributes can amplify measured privacy risk. Importantly, their framework connects quantitative privacy metrics to the legal standards established by the GDPR, contributing to the standardization of privacy assessment for synthetic data as a viable privacy-enhancing technology.

A more nuanced evaluation paradigm has been proposed under the rubric of *epistemic parity*. Rosenblatt and Howe (2022) argue that the appropriate benchmark for differentially private synthetic data is not exact numerical reproduction of original estimates, but rather the ability to reproduce the qualitative conclusions of published empirical findings - that is, to preserve the sign, significance, and approximate magnitude of key coefficients. While epistemic parity may be adequate for exploratory research and teaching applications, it remains insufficient for policy-relevant economics that requires quantitative precision in effect sizes, confidence intervals, and distributional outcomes.

Abowd and Schmutte (2019) frame the privacy-utility trade-off in economic terms, conceptualizing the choice between statistical accuracy and privacy protection as a social welfare optimization problem. Their framework provides a rigorous foundation for evaluating the costs and benefits of different disclosure limitation strategies, including synthesis with and without differential privacy, and highlights the role of institutional design in mediating this trade-off.

2.7 Complex Survey Design Considerations

Many economic datasets of primary importance - including national labor force surveys, household expenditure surveys, health interview surveys, and consumer finance surveys - are based on complex probability sampling designs involving stratification, clustering, and survey weights. These design features are integral to producing valid population-level inferences and must be taken into account properly in the synthesis process.

The literature documents that ignoring the features of the survey design during synthesis can introduce systematic biases in the point estimates and, more critically, in the estimation of variance (Raab et al., 2020; Drechsler, 2011). Incorporating stratification and clustering variables as conditioning variables in the synthesis models and preserving the original survey weight structure has been shown to improve the analytic validity of synthetic data for design-based inference.

Zhang et al. (2025) provide a recent contribution to this area by comparing multiple synthesis approaches - including CART-based methods and deep learning generators - applied to the NCHS Research and Development Survey (RANDS), a complex survey dataset. Their study evaluates synthetic data quality through design-consistent estimators and highlights the additional difficulties that stratification and complex weighting introduce for synthetic data generators.

However, the current literature has important gaps. Replicate-weight variance estimation - the standard approach for variance estimation in complex surveys, involving techniques such as balanced repeated replication (BRR), jackknife, and successive difference replication - has rarely been evaluated in the context of synthetic data. The creation of synthetic replicated weights that preserve the variance properties of the original design remains an open methodological challenge. Shokri et al. (2017); Stadler et al. (2022) discuss disclosure risk assessment in the context of computational models, but the literature lacks systematic treatments of how synthesis interacts with replicate-weight variance estimation.

This gap undermines confidence in using synthetic data for design-based inference in surveys widely used in economics, such as the Current Population Survey (CPS), the Consumer Expenditure Survey (CE), the Panel Study of Income Dynamics (PSID), and the Survey of Consumer Finances (SCF). Addressing this shortcoming - by developing synthesis methods that jointly generate synthetic data and coherent replicate weights - is essential for broader adoption of synthetic data in empirical economics.

Our case studies use administrative data from the universe of enrolled (or graduated) students and survey data on the population of graduates. As such, we do not suffer issues related to complex survey design in the application

considered in the report.

2.8 Relational and Multi-Table Synthetic Data

While the majority of the synthetic data literature has focused on single-table generation, real-world datasets in economics and official statistics are frequently organized as relational databases - comprising multiple tables linked by foreign-key relationships. Examples include household surveys with separate person- and household-level files, employer–employee linked datasets, and administrative records joining tax, benefit, and demographic tables. Faithfully synthesizing such relational structures requires modeling not only within-table distributions but also across-table dependencies and cardinality constraints (e.g., the number of children per parent record).

The foundations for relational synthetic data generation were laid by the Synthetic Data Vault (Patki et al., 2016), which uses Gaussian copulas and predefined distributions to model multi-table data. Subsequent methods have drawn on a variety of deep learning architectures. GAN-based approaches include Row Conditional-TGAN (Gueye et al., 2023) and the Incremental Relational Generator (Li and Tay, 2023). Transformer-based methods have gained prominence with REaLTabFormer (Solatorio and Dupriez, 2023), which employs a GPT-2 backbone for parent-table generation and a sequence-to-sequence model for child tables conditioned on the parent. REaLTabFormer introduces target masking to prevent data copying and proposes the Q_δ statistic with statistical bootstrapping for overfitting detection. Diffusion-based approaches are represented by ClavaDDPM (Pang et al., 2024), which uses classifier-guided diffusion for relational data.

A significant recent advance is the graph-conditional flow matching framework of Scassola et al. (2025), which recasts relational data generation as a continuous normalizing flow problem over the entire database. By training a denoising neural network that incorporates a graph neural network (GNN), information can propagate across arbitrarily connected records within the same component. This approach is distinguished by its flexibility - it naturally handles complex relational structures including tables with multiple parents and multiple relationship types between the same pair of tables - and its scalability to large databases. Experimental evaluations demonstrate state-of-the-art fidelity on several benchmark datasets.

Despite these advances, the systematic evaluation of relational synthesis methods has lagged behind single-table benchmarks. Hudovernik et al. (2024) address this gap with the first comprehensive empirical comparison of six relational data synthesis methods - including two commercial tools - across multiple datasets. Their benchmarking framework combines statistical fidelity metrics (Kolmogorov–Smirnov and chi-squared tests), distance-based metrics, and

a novel robust detection approach. A key finding is that no current method is able to produce synthetic relational data that is indistinguishable from the original. For utility, measured through train-on-synthetic, evaluate-on-real protocols, they observe only moderate correlation between real and synthetic predictive performance and feature importance. These results underscore that relational data synthesis - while rapidly advancing - remains substantially more challenging than single-table generation, and that standardized benchmarks are essential for tracking progress.

2.9 The Evolution of Generative Methods

The methodological landscape of synthetic data generation has undergone three broad phases of development.

Foundational phase (2001–2011). The initial period was characterized by the development of theoretical frameworks for fully and partially synthetic data, the derivation of appropriate variance estimators, and the first empirical demonstrations of feasibility. Key contributions include the theoretical proposals of Rubin (1993) and Little (1993); the inference frameworks of Reiter (2003, 2005); the SRMI methodology of Raghunathan et al. (2001); and the consolidating monograph of Drechsler (2011). The first generation of synthetic public-use files, including the SIPP Synthetic Beta, was developed during this period (Abowd et al., 2006; Abowd and Woodcock, 2001). The appropriate variance estimators for synthetic data stem from two distinct frameworks, depending on whether the data are fully or partially synthetic:

1. ***Fully Synthetic Data*** (Reiter, 2003). When all records in the released dataset are synthetic (i.e., no real individual's data is directly released), the variance estimator combines: i) The within-imputation variance, capturing variability within each synthetic dataset; ii) the between-imputation variance, capturing variability across the multiple synthetic datasets generated. The combining rules are analogous to - but different from - those used in missing data multiple imputation.

2. ***Partially Synthetic Data*** (Reiter, 2005) When only some sensitive variables or records are replaced with synthetic values, a different set of combining rules applies, as the estimator must account for the fact that part of the data is still real. The variance estimator is adjusted accordingly to avoid underestimating uncertainty.

3. ***SRMI Framework*** (Raghunathan et al., 2001) The Sequential Regression Multiple Imputation approach provides a practical procedure for generating the synthetic datasets themselves, and the associated variance es-

timation follows the combining rules above depending on whether full or partial synthesis is adopted.

In all these cases, releasing multiple synthetic datasets (rather than just one) is essential, because it is the variability across synthetic replicates that allows proper uncertainty quantification. A single synthetic dataset would not permit valid variance estimation.

Consolidation phase (2012–2017). The second phase saw the maturation of synthesis methods through empirical application to a wider range of surveys and business datasets. Key developments include the Synthetic Longitudinal Business Database (Kinney et al., 2011); the introduction of the verification server model (Karr et al., 2006; Reiter, 2005); the development of the `synthpop` toolkit (Nowok et al., 2016); refined CART-based synthesis methods (Nowok et al., 2016; Raab et al., 2020); and utility evaluation frameworks (Snoke et al., 2018). During this phase, the practical limitations of classical synthesis - particularly in tail preservation and dynamic processes - became increasingly well documented.

Expansion phase (2018–present). The current phase is defined by three concurrent developments: the emergence of deep generative models for tabular data (Xu et al., 2019; Patki et al., 2016; Kotelnikov et al., 2023; Borisov et al., 2022a); the integration of differential privacy into synthetic data generation (Jordon et al., 2018; Stadler et al., 2020; Bowen and Snoke, 2021); and the application of synthetic data methods to large-scale official statistical products, including census microdata (Abowd et al., 2020) and tax public-use files (Abowd and Schmutte, 2019). Deep generative approaches - including GANs, VAEs, normalizing flows, and diffusion models - offer the promise of capturing arbitrarily complex dependence structures without explicit parametric assumptions. However, the empirical evidence on their performance relative to classical methods remains equivocal, with advantages that are highly task- and dataset-specific (Bowen and Liu, 2020). A notable expansion of this phase is the extension from single-table to relational data synthesis, driven by transformer-based models (Solatorio and Dupriez, 2023), flow matching on graphs (Scassola et al., 2025), and the development of systematic evaluation frameworks for multi-table fidelity and utility (Hudovernik et al., 2024).

A particularly important recent development is the work of Decruyenaere et al. (2024a,b); Ghalebikesabi et al. (2022), who propose post-hoc correction methods for statistical inference conducted with synthetic data generated by deep learning models. Their approach acknowledges that deep generators typically do not produce data that is compatible with the specialized variance estimators developed for classical synthesis, and provides a practical pathway

for recovering valid inference from deep-learning-generated synthetic data.

Throughout all three phases, the SIPP has served as a crucial benchmark dataset, providing repeated within-study comparisons between synthetic and confidential versions (Abowd and Stinson, 2013; Benedetto et al., 2013; Abowd and Vilhuber, 2005). These comparisons have set the empirical standard for evaluating both classical and modern synthesis approaches and have documented the persistent challenges of tail preservation, dynamic process replication, and variance calibration.

2.10 Current Gaps and Future Directions

Despite the substantial progress documented above, several critical challenges remain unresolved.

First, deep generative models have rarely undergone rigorous side-by-side evaluations against confidential data using the inferential criteria standard in economics - coefficient accuracy, confidence interval coverage, and hypothesis test validity. The majority of evaluations in the machine learning literature rely on predictive accuracy metrics or distributional distance measures that are insufficient for assessing the validity of economic inference.

Second, predictive evaluation protocols are inconsistently applied and often lack the rigor necessary for deployment in consequential contexts. The absence of systematic TOSO protocols - in which models trained on synthetic data are evaluated on holdout samples from the original confidential data - limits the evidentiary basis for using synthetic data in operationally predictive economic applications.

Third, survey design considerations remain inadequately addressed. The creation of synthetic datasets that jointly preserve the informational content of the original data and the variance properties of the complex sampling design - including replicate weights and design effects - is an open methodological problem with significant practical implications.

Fourth, the literature lacks systematic *policy-microsimulation parity tests* in which identical tax-benefit simulation pipelines are applied to both original and synthetic datasets. Such tests would provide a direct and policy-relevant measure of synthetic data quality that goes beyond regression-centric evaluations.

Fifth, the relationship between the privacy-utility trade-off and the specific requirements of economic inference deserves further theoretical and empirical investigation. Current calibrations of ε in differentially private synthesis are largely driven by technical considerations rather than by the inferential demands of the downstream economic application.

Sixth, the generation of synthetic relational data - while rapidly advancing - remains substantially more difficult than single-table synthesis. No current method produces multi-table data that is indistinguishable from the original

(Hudovernik et al., 2024), and long-range dependencies across tables with complex foreign-key structures continue to pose challenges (Scassola et al., 2025). Furthermore, standardized benchmarks for relational data quality are only beginning to emerge.

Seventh, the standardization of privacy quantification for synthetic data remains incomplete. While Plasencia Palacios et al. (2025) provide a promising benchmarking framework that connects empirical privacy metrics to GDPR-derived legal standards, the field still lacks consensus on which metrics best capture operationally meaningful privacy risk across different data types and threat models.

Future research should prioritize: (i) hybrid synthesis approaches that combine the parametric discipline of classical sequential models with the representational flexibility of deep generative architectures; (ii) design-aware synthesis methods that fully integrate survey weights, clustering, and stratification into the generative process; (iii) verification server and DP mechanisms that provide calibrated, task-specific feedback on the accuracy of synthetic-data-based analyses; (iv) scalable relational synthesis methods - including graph-based and transformer-based approaches - that can handle multi-table databases with complex foreign-key structures; (v) standardized privacy benchmarking frameworks that link quantitative metrics to regulatory standards; and (vi) standardized evaluation benchmarks - including TOSO protocols, microsimulation parity tests, and tail-specific fidelity measures - that align assessment criteria with the inferential standards of empirical economics.

2.11 How and where do we improve with respect to the existing literature

The above review identifies several open challenges that this report directly confronts. The *first* and *sixth* gaps are arguably the most relevant for our purposes: deep generative models have rarely been evaluated using the inferential criteria standard in empirical economics, and the synthesis of relational, multi-table data – the format in which most administrative micro-data actually exist – remains substantially harder and less systematically evaluated than single-table generation (Hudovernik et al., 2024). The *fifth* gap is also directly relevant: we operate under binding GDPR requirements but without access to formal differential privacy mechanisms, and our results therefore contribute empirical evidence on where the privacy-utility frontier lies for European administrative micro-data in a non-differentially private setting.

Against this backdrop, the present report makes three interconnected contributions. First, it provides one of the first applications of a *relational* synthetic data engine to large-scale university administrative records: the input data span multiple linked tables (personal details, careers, enrollment, exam

records, and graduate-survey responses), and the generating model captures intra-table distributions and cross-table dependencies within a single modeling pipeline. Second, the evaluation is conducted using the inferential criteria standard in empirical economics – regression coefficients, classification performance, and cluster structure – applied to both original and synthetic data and replicated across four institutions from the GRINS network that differ in dataset size, geographic location, and synthesis hardware. This cross-institutional design enables a meta-analytic assessment of fidelity and utility that goes beyond the single-dataset evaluations that dominate the existing literature. Third, by operating under strict GDPR constraints without noise injection, the study contributes to the evidence base on the practical viability of generation-based privacy protection for institutional micro-data. The technical architecture of the synthetic data engine and its precise positioning within the landscape of generative methods reviewed in Section 2.8 and Section 2.9 are described in Section 3.

3 From real to synthetic data

The synthetic data technology used in this project belongs to the *expansion phase* of generative methods identified in Section 2.9 and is specifically designed for relational, multi-table data — the open challenge highlighted in Section 2.8 (Hudovernik et al., 2024). The following subsection describes the key components of AINDO's model and its positioning relative to the literature reviewed above; subsequent subsections describe the two generation approaches and the quality metrics used to evaluate their outputs.

3.1 AINDO's synthetic data model

AINDO's model is a proprietary autoregressive generative model for structured tabular and relational data (AINDO, 2025). In the autoregressive framework, the joint probability distribution of all variables is factorized as a chain of conditional probabilities, where each variable is generated sequentially conditioned on all previously generated ones.

$$P(X_0, X_1, X_2, \ldots, X_n) = P(X_0) \cdot P(X_1|X_0) \cdot P(X_2|X_1, X_0) \cdot \ldots \cdot P(X_n|X_{n-1}, \ldots, X_1, X_0).$$

The model is trained from scratch on each dataset using cross-entropy loss and stochastic gradient descent, with no pretraining on external corpora and no injection of domain-specific prior knowledge.

The architecture consists of four components:

1. *Attention*: the model uses a linear (rather than quadratic) attention mechanism with key-value caching, inspired by the Attention Free Transformer

(Zhai et al., 2021), which reduces the computational and memory cost of processing the long token sequences that arise when flattening multi-table databases into a single input.

2. *Tokenization*: each column is tokenized according to its data type. Categorical variables are assigned vocabulary tokens drawn from the observed categories, so that only valid categories can be generated. Continuous variables are decomposed digit by digit in base-10 scientific notation with position-dependent token identifiers, so that the same digit in different decimal positions receives a distinct token; a fixed-width encoding ensures that every value in a given column uses the same number of tokens. The datetime variables are decomposed into year, month, day, and hour components, each tokenized separately. The token sequences of all columns are concatenated in a fixed order to produce a complete row representation.

3. *Positional embeddings*: beyond a standard value-based embedding, each token also encodes the column and table from which it originates, enabling the model to distinguish tokens with similar values but different structural roles (e.g. a date in a "BirthDate" column versus a date in a "TransactionDate" column); an additional "siblings embedding" propagates information about rows sharing a common foreign key — for instance, all exam entries belonging to the same enrollment — allowing the model to capture longitudinal dependencies within an individual's record.

4. *Output heads*: a dedicated output head is assigned to each column; at each generation step, the head masks all tokens that are structurally invalid for that column, ensuring that generated records always conform to the schema constraints of the original data.

AINDO's model handles *relational* data by exploiting the tree structure that characterizes most privacy-sensitive administrative databases. A relational database is represented as a graph in which tables are vertices and foreign-key relationships are directed edges. When each table contains at most one foreign key - in the tree-structured case - each connected component of the row-wise graph corresponds exactly to one individual's complete record across all linked tables. This one-to-one correspondence is the mathematical foundation for privacy-preserving generation: each component constitutes one independent sample of the data distribution, and synthesis consists of drawing new such samples without referencing any original individual. To linearize each individual's relational subgraph into a sequence suitable for the autoregressive model, a *depth-first search* (DFS) traversal is applied starting from the unique root-table row, visiting all linked child rows in a fixed order; the tokenized rows are then concatenated (*flattened*) into a single contiguous input sequence. The relational structure is encoded implicitly in this ordering: the parent row of any

child record is always the last row of its parent table in the preceding part of the sequence.

AINDO's model does not implement formal differential privacy. Privacy is instead embedded in the generation process itself: the model learns a statistical approximation of the data distribution and draws entirely new records from it, so that no original observation appears in the released dataset. This approach avoids the utility degradation associated with the injection of ε-noise documented in Section 2.6 (Stadler et al., 2022; Bowen and Liu, 2020), while targeting GDPR compliance through distribution anonymity (Plasencia Palacios et al., 2025).

3.2 Data generation

The generation of synthetic data in this project aims at reproducing the statistical properties of administrative student career data and AlmaLaurea datasets while ensuring strong privacy guaranties.

Synthetic data generation is based on generative artificial intelligence techniques, which aim to learn the underlying probability distribution of a data set and generate new samples from it. The process follows three main steps. First, a training data set is provided as input. Second, the generative model learns the patterns and relationships present in the data by approximating their joint probability distribution. Finally, new data points are generated by sampling the learned distribution, producing a synthetic dataset that mimics the statistical properties of the original data such as marginal distributions, correlations, and relational structure.

In the present project, the synthetic data generation process relies on two main data sources, administrative student career data and AlmaLaurea data, that are inherently relational, as they consist of multiple interconnected tables linked through keys. Table 1 reports the name and size of the original data for each table used by each university. A detailed description of the data tables is provided in Appendix B. In these pilot cases, each University operates as *data controller* and can thus have access to both administrative and survey data. However, the typical situation observed in practice is that the data relevant for a research project come from several distinct sources with distinct institutions playing the role of data controllers. Thus, we generate synthetic data with two distinct approaches. First, we generate data by assuming that both administrative and survey data have the same *data controller*. In this context, all administrative and survey data can be used to generate the synthetic data. The resulting synthetic data bases are indicated as CSD in all tables and figures in this report. Second, we design the synthetic data generation process assuming that the data controllers differ for administrative and survey data. As a consequence, the data generation process differ substantially in this second case.

Table 1. Pre-processed input data - sample size across tables by University

Relational data-base table	Reference University code 1	2	3	4
	Registry data from University archives^{+}			
Personal details	70,633	275,605	85,034	114,107
College career	83,767	332,356	103,798	134,178
Enrollment	218,648	887,772	103,798	331,011
Entry test	72,611	50,150	–	35,449
Exams	973,731	4,999,606	1,772,461	1,681,124
	Survey data^{++}			
	Graduates' Profile, at graduation $^{++}$			
Graduates' Profile	27,544	143,332	103,773	55,760
	Graduates' Employment Status, post graduation^{++}			
in 1 Year	22,754	123,763	77,980	42,006
in 3 Years	5,216	50,606	26,770	14,232
in 5 Years	3,004	33,142	16,115	9,835

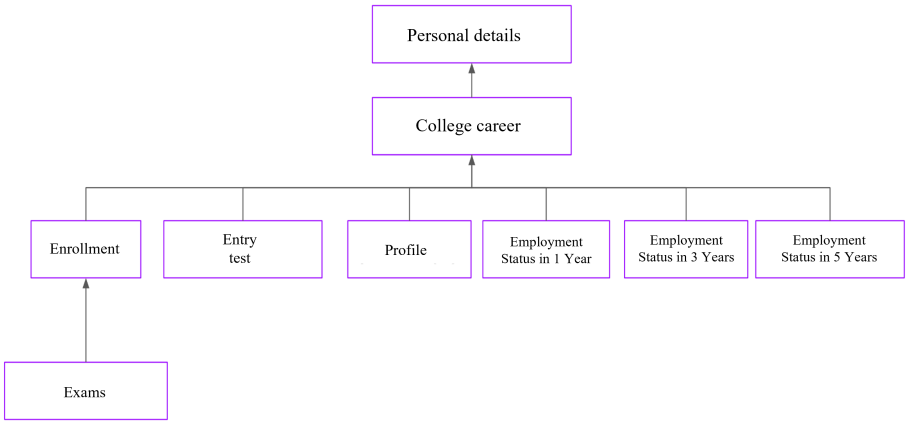
Note: $^{+}$ Registry data out of administrative data from the universe of enrolled students in the period 2012-2025.
$^{++}$ Survey data on graduates' profiles (at graduation) and employment status (post-graduation) come from AlmaLaurea. Specifically, data on graduates' profiles are available for the period 2012–2024, while data on employment status cover the period 2013–2024.

Synthetic data corresponding to administrative data only are generated, and, independently, synthetic data corresponding to survey data are generated. In both cases, each block relies on a distinct relational data base consistent with the original data structure. An encrypted representation of the data bases is used to retrieve the structure data mapping between the administrative and survey data (*linking model*). The two steps are used to link synthetic data that are independently generated. The resulting synthetic data bases are indicated as FSD in all tables and figures in this report. We discuss the two approaches in more detail below.

Centralized Synthetic Data Generation (CSD). Synthetic data are generated using the AINDO's model 3.1 and a centralized approach, in which all available data are integrated into a single modeling pipeline (see Figure 1).

A large-scale model configuration (using 16 transformer layers, using a multi head attention with 18 heads each of dimension 20, referred to as "XXXL") was adopted in order to cope with the high dimensionality and complexity of the data. This choice proved crucial for accurately reproducing not only

Figure 1. Relational Structure in Centralized Synthetic Data Generation (CSD)



marginal distributions, but also correlations and multivariate patterns across tables. The generative model is trained to capture both intra-table distributions and inter-table dependencies, enabling the generation of coherent synthetic relational datasets. Table 2 summarizes the technical details of the autoregressive model implemented and the hardware used at each university. Further technical details are provided in Appendix A.

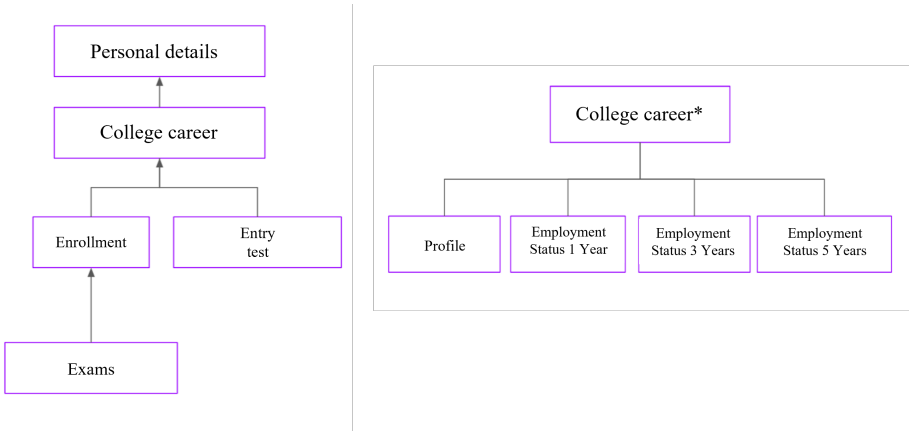
Table 2. Comparison of models trained at different universities.

	No. of parameters	Training time	Convergence	Synthesis time	Hardware
University 1	44M	28 days	No	72h	Intel(R) Xeon(R) Platinum 8460Y+
University 2	64M	11 days	Yes	104h	1x A4000 (16GB)
University 3	51M	30 days	No	33h	Intel(R) Xeon(R) Platinum 8370C (62GB)
University 4	48M	7 days	Yes	17h	NVIDIA RTX 500 Ada Generation (32GB)

Federated Synthetic Data Generation (FSD). The federated synthetic data generation extends the synthesis framework by introducing a federated approach designed to improve privacy protection. In this setting, the different data sources (i.e., university administrative data and AlmaLaurea data) are processed independently rather than merged into a single data set (see Figure 2).

The methodology is carried out in three stages. In the first stage, synthetic

Figure 2. Relational Structure in Federated Synthetic Data Generation (FSD)



data are generated independently for each data source using AINDO's model 3.1: one instance is trained on the administrative records (personal details, careers, enrollment, exams) and a separate instance is trained on the AlmaLaurea survey data. Each instance generates a fully synthetic relational database for its respective source, with no information exchanged between the two. In the second stage, a *linking model* is trained to reconstruct the foreign-key relationship that connects the two independently generated databases - that is, to learn which synthetic graduate in the administrative database corresponds to which record in the AlmaLaurea database. In the third stage, the trained linking model is applied to the synthetic outputs of both instances to produce a single coherent linked synthetic dataset.

The linking model operates as follows. Each row of the administrative (parent) table and each row of the AlmaLaurea (child) table are independently encoded into a real-valued embedding vector. The encoding uses the same tokenization scheme as AINDO's generative model, optionally augmented with information from ancestor rows in the relational hierarchy. A transformer-like encoder then maps each tokenized row to a fixed-dimensional embedding. A *similarity matrix* is then computed as the matrix of dot products between all child embeddings and all parent embeddings: each entry measures how likely a given AlmaLaurea record is to be linked to a given administrative record. The model is trained on the original data by minimizing a cross-entropy loss in which the actual foreign key provides the target. At inference time, the similarity matrix computed from the synthetic embeddings is used to sample the foreign key assignments - i.e. to decide which synthetic graduate record is linked to which synthetic AlmaLaurea record.

A naive row-by-row sampling from the similarity matrix can distort the distribution of the number of AlmaLaurea records per graduate (the number of children per parent row). To correct this, the linking model incorporates a dedicated module that predicts the number of children for each parent row, and the foreign key assignment is drawn using a Metropolis–Hastings Monte Carlo algorithm. At each step of the simulation, two child rows are chosen at random, and their parent assignments are swapped; the proposed swap is accepted if it increases the joint probability under the similarity matrix and otherwise accepted with probability equal to the probability ratio. This constrained sampling ensures that the number of children per parent remains consistent with the predicted distribution, preserving the relational structure of the original data in the linked synthetic output.

Due to differences in administrative procedures across Universities, hardware and human resources availability, within the short time-span of the project, only two Universities (University 2 and 4) were able to perform the creation of synthetic data using the two distinct designs. In contrast, all universities obtained the CSD. As described in Appendix B, the original data for University 3 comprise only graduates (2013–2025) who enrolled in degree programs under DM 270 between 2012 and 2022, while data for the other Universities refer to all enrolled students over the academic years 2012-2025. As a consequence, the assessment of synthetic data across Universities based on indices of synthetic data quality refer exclusively to CSD and two out of three empirical applications considered in this report use only data from University 1, 2 and 4.

3.3 Illustration of indices of synthetic data quality

In this section, we present indices that aim to evaluate the quality of synthetic data. These indices are conventionally used in the computer science literature and are appropriate to assess levels of privacy protection and statistical fidelity not related to specific application, i.e., in a way that is agnostic with respect to the actual use of original or synthetic data. We present some applications to assess the performance of synthetic data relative to the original data in specific applications in Section 4 of the report. The evaluation of synthetic data quality follows a dual objective framework: *(i)* privacy protection, ensuring that synthetic records do not disclose information about real individuals; *(ii)* statistical fidelity, ensuring that synthetic data preserve the analytical structure of the original data set (El Emam et al., 2020). Among the evaluation metrics proposed in the context of synthetic data generation, AINDO adopts the following measures:

Privacy metric. The privacy metric is based on nearest-neighbor distances. The underlying principle is that privacy risk arises when synthetic observations

are statistically too close to the real observations used for training, potentially leading to information leakage.

Let R denote the real training data set and S the synthetic data set generated by the model. The real data set R is randomly split into two disjoint subsets (R_1, R_2) and, for each observation $x \in R_1$, the following quantities are defined:

- $d(x, R_2)$: the distance from x to its nearest neighbor in R_2,
- $d_2(x, R_1)$: the distance from x to its second nearest neighbour in R_1 (the first nearest neighbour would be x itself).

The distribution of the ratio between these quantities is called *Train-to-Train Proximity Ratio* (TTPR):

$$\mathrm{TTPR}(x) = \frac{d(x, R_2)}{d_2(x, R_1)}. \tag{1}$$

The empirical distribution of $\mathrm{TTPR}(x)$ across $x \in R_1$ represents the baseline proximity structure of the real data. Low values of TTPR indicate that $d(x, R_2)$ is small relative to the internal spacing $d_2(x, R_1)$.

An analogous ratio is computed replacing R_2 with the synthetic data set S, obtaining the *Train to Synthetic Proximity Ratio (TSPR)* distribution:

$$\mathrm{TSPR}(x) = \frac{d(x, S)}{d_2(x, R_1)}, \tag{2}$$

where $d(x, S)$ denotes the distance from x to its nearest neighbor in S. The empirical distribution of $\mathrm{TSPR}(x)$ reflects how close synthetic records are to real training observations.

Then a quantile-based risk threshold is computed. Let $\alpha \in (0, 1)$ denote a fixed parameter (by default $\alpha = 0.1$), and let q_α be the α-quantile of the TTPR distribution such that $P(\mathrm{TTPR} \leq q_\alpha) = \alpha$. This threshold identifies the lower tail of the baseline proximity distribution; for example, when $\alpha = 0.1$, q_α corresponds to the lowest 10% % of the TTPR values. Since low TTPR values occur when $d(x, R_2)$ is small relative to the internal spacing $d_2(x, R_1)$, observations below q_α are those that are comparatively closer to independent real samples in R_2 than would be expected given the internal spacing of R_1.

The same threshold q_α is then applied to the TSPR distribution and the proportion of synthetic proximity ratios falling below this threshold is calculated as $\beta = P(\mathrm{TSPR} \leq q_\alpha)$. This step compares how many real observations have a synthetic neighbor that is at least as close as the most similar real-to-real neighbors in the lower tail. If synthetic data behaves similarly to real data, then β should be approximately equal to α. If, instead, β is substantially larger than α, this indicates that synthetic records are systematically closer to real

records than expected under the real-data baseline, suggesting an increased risk of privacy.

Finally, the *Privacy* score is defined as follows:

$$\text{Privacy} = \begin{cases} 100, & \text{if } \beta \leq \alpha, \\ 100 \times \dfrac{\alpha}{\beta}, & \text{if } \beta > \alpha. \end{cases} \tag{3}$$

If $\beta \leq \alpha$, the score is equal to 100, indicating that synthetic observations are not closer to real data than expected under the real data baseline scenario. If $\beta > \alpha$, the score decreases proportionally, reflecting the increasing risk of proximity-based disclosure. The metric compares the lower tail of the synthetic proximity distribution with the baseline lower tail of real-to-real distances. A score of 100 indicates the absence of detectable privacy risks in this distance-based diagnostic. In contrast, near-zero values indicate that most real records are at risk, as synthetic records are too similar to them.

Similarity Score. The Similarity Score provides an aggregate measure of the statistical fidelity between real and synthetic datasets. The score (ranging from 0 to 100) aggregates information from distributional and bivariate comparisons and provides a summary measure of how closely the synthetic data replicate the original structure.

The Similarity Score for a given data table is computed as the mean of all Bivariate Similarity scores S_{XY} across all distinct pairs of variables (X, Y) within the table:

$$\text{Similarity Score} = \frac{1}{K} \sum_{(X,Y)} S_{XY}, \tag{4}$$

where K denotes the total number of variable pairs considered.

The Bivariate Similarity Score is defined as follows:

$$S_{XY} = 1 - \text{TVD}(P_{XY}, \tilde{P}_{XY}), \tag{5}$$

where P_{XY} and $\tilde{P}_{XY}$ denote the empirical joint distributions of (X, Y) in the real and synthetic datasets, respectively, and TVD is the Total Variation Distance between the two distributions. The latter is defined as one half of the sum of absolute differences between the observed frequencies of each bin in histograms generated from the real and synthetic data. The TVD ranges between 0 and 1, with 0 indicating identical distributions. Consequently, the Bivariate Similarity Score also ranges between 0 and 1, with values closer to 1 indicating stronger agreement between real and synthetic joint distributions. Accordingly, the Similarity Score - rescaled to a 0–100 scale for ease of inter-

pretation - approaches 100 when the synthetic data faithfully reproduce the joint distributional structure of the original data. Conversely, lower values suggest discrepancies in one or more pairwise relationships, potentially reflecting distortions introduced during the generative process.

Similarity matrix. The full set of pairwise similarity measures is summarized in a Similarity Matrix. Each entry of the matrix corresponds to the similarity between the joint distributions of two variables in the real and synthetic datasets. Values closer to 1 indicate a high degree of alignment between the real and synthetic joint behavior, while lower values highlight discrepancies in the reproduced relationships. By examining the similarity matrix, it is possible to identify specific pairs of variables for which the synthetic data may fail to capture the underlying joint behavior.

Univariate and bivariate distributions. Univariate fidelity is evaluated by comparing the marginal distributions of each variable in the real and synthetic datasets. For numerical variables, the mean and standard deviation are examined for both datasets, whereas for categorical variables, the frequency distributions across categories are compared. Additionally, the proportions of missing values are explicitly reported for both datasets, together with a visual comparison of the empirical distributions. These analyses allow us to assess whether the synthetic data accurately reproduces the distributional properties of the original dataset at the univariate level.

Beyond marginal distributions, the evaluation considers pairwise (bivariate) relationships between variables. Bivariate fidelity is assessed by examining the joint distribution of pairs of variables in the real and synthetic datasets. Bivariate empirical distributions are typically visualized using heatmaps. In these plots, the possible combinations of values of the two variables are divided into a grid of intervals (bins). For each cell of the grid, the colour intensity represents the number or proportion of observations that fall within that specific combination of value ranges. This graphical representation enables a direct visual comparison of the joint distributions in the real and synthetic datasets.

PhiK Correlation. Dependency structures are further evaluated using the PhiK correlation coefficient (Baak et al., 2020). PhiK is specifically designed to measure associations in mixed-type datasets, accommodating categorical, ordinal, and numerical variables within a unified framework. It is based on a chi-square contingency framework and extends Pearson's correlation concept, allowing it to capture both linear and non-linear relationships. PhiK values range from 0 (no association) to 1 (perfect association), with higher values indicating stronger relationships.

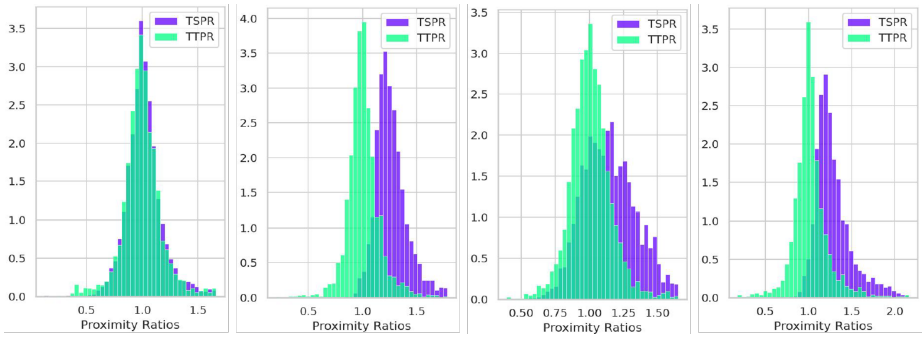
PhiK correlation matrices are computed separately for the real and synthetic datasets. A difference matrix, obtained by subtracting the synthetic correlation matrix from the real one, is then reported to highlight deviations in pairwise associations. This visualization facilitates the identification of attenuated, amplified, or distorted dependency patterns in the synthetic data. Close alignment between the real and synthetic correlation matrices suggests that the synthetic data preserves the multivariate dependency structure of the original dataset.

3.4 An assessment of synthetic data across Universities, based on indices of synthetic data quality (CSD)

The evaluation of synthetic data generated for the four participating universities reveals a highly consistent pattern across institutions and data tables. Both privacy protection and statistical fidelity metrics show stable and robust performance.

Privacy protection. Across all universities and all data tables considered (see Table 1 for the complete list) , the Privacy Score consistently reaches the maximum value of 100/100. This uniform result indicates that synthetic records are not statistically closer to real training observations than expected under the real-to-real baseline scenario defined by the Train-to-Train Proximity Ratio (TTPR). In practical terms, the lower tail of the Train-to-Synthetic Proximity Ratio (TSPR) distribution never exceeds the empirical proximity threshold derived from the real data (see Figure 3 for an example related to "Enrollment" data table). This result confirms that no detectable proximity-based disclosure risk emerges in any institutional context.

Figure 3. Train-to-Synthetic Proximity Ratio (TSPR) and Train-to-Train Proximity Ratio (TTPR) distributions for the "Enrollment" data table. The panels represent (from left to right) the results for University 1 to University 4.



Statistical fidelity. The *Global Similarity Score* is consistently high across all universities and data tables, ranging between 95 and 97 on a 0-100 scale. The narrow dispersion of these values indicates that the synthetic datasets replicate the joint distributional structure of the real data with a high degree of accuracy.

The *similarity matrices* further confirm that most pairwise bivariate similarity values are close to 1. As shown in Table 3, even the 5th percentile remains above 0.90 in all cases, indicating that even the weakest preserved relationships maintain a high degree of fidelity.

Table 3. Average percentiles (5th, median, 95th) of the similarity matrices computed across all data tables within each university

University	5th Percentile	Median	95th Percentile
University 1	0.93	0.97	0.99
University 2	0.94	0.98	0.99
University 3	0.93	0.97	0.99
University 4	0.92	0.97	0.99

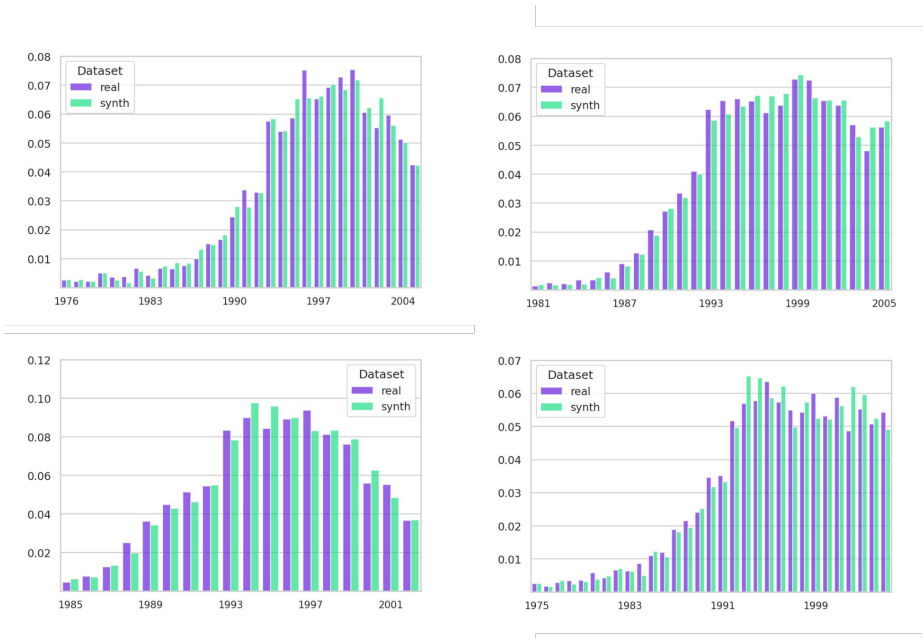
The inspection of marginal *univariate distributions* confirms the strong alignment between real and synthetic data already suggested by the similarity metrics. In the *Personal details* data table, the distributions of gender, year of birth, and month of birth show almost perfect overlap between real and synthetic data. Geographic variables such as region of birth and province of birth reproduce the concentration in dominant categories while maintaining the long tail of smaller regions. An illustrative example is shown in Figure 4 for the variable year of birth.

In the *College career* data table, the distributions of structural academic variables such as the type of degree programme are highly consistent between real and synthetic data. Numeric variables, such as the year of birth of the professor (where available), show preserved means and standard deviations. An illustrative example is shown in Figure 5 for the variable type of degree programme.

In the *Enrollment* data table, variables such as high school final grade and final degree grade maintain both distributional shape and dispersion. Even in variables with substantial missing values (e.g., degree programme class of the previous qualification), the synthetic data reproduce the missingness pattern. An illustrative example is shown in Figure 6 for the variable high school final grade.

In the *Exams* data table, exam-level data preserve the distribution of exam grades, credit values, instructor characteristics, and timing variables. The dispersion of numeric variables remains consistent, as do the distributions of categorical variables. An illustrative example is shown in Figure 7 for the variable

Figure 4. Distribution of the variable "year of birth" across universities. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).

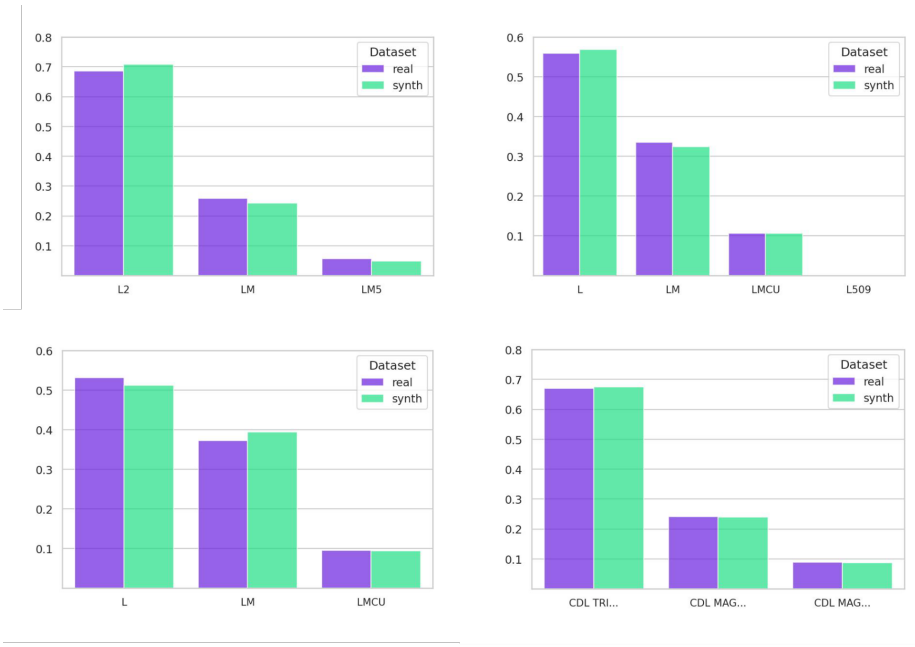


exam grade. Note that this variable represents the distribution of all exam grades recorded for all students, with all observations stacked together.

The *Entry test* data table preserves the distribution of admission scores, including subject-specific components such as mathematics, Italian language, and logic, as well as the overall TOLC score. The TOLC (Test OnLine CISIA) is a standardized online assessment widely used by Italian universities for admission purposes, designed to evaluate candidates' skills in core subject areas depending on the degree program. The synthetic data accurately reproduce the distributional shape and dispersion of both subject-level and total scores. An illustrative example is shown in Figure 8 for the variable overall TOLC score.

In the *Graduates' Profile* data table, graduate profile variables, including degree class, final grade, academic progression indicators, and work experience, exhibit strong overlap between real and synthetic distributions. For example, the distribution of the final grades preserves its central tendency and upper-tail concentration. Categorical variables related to study experience replicate dominant response categories without distortion. An illustrative example is shown in Figure 9 for the variable work experience.

Figure 5. Distribution of the variable "type of degree programme" across universities. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).

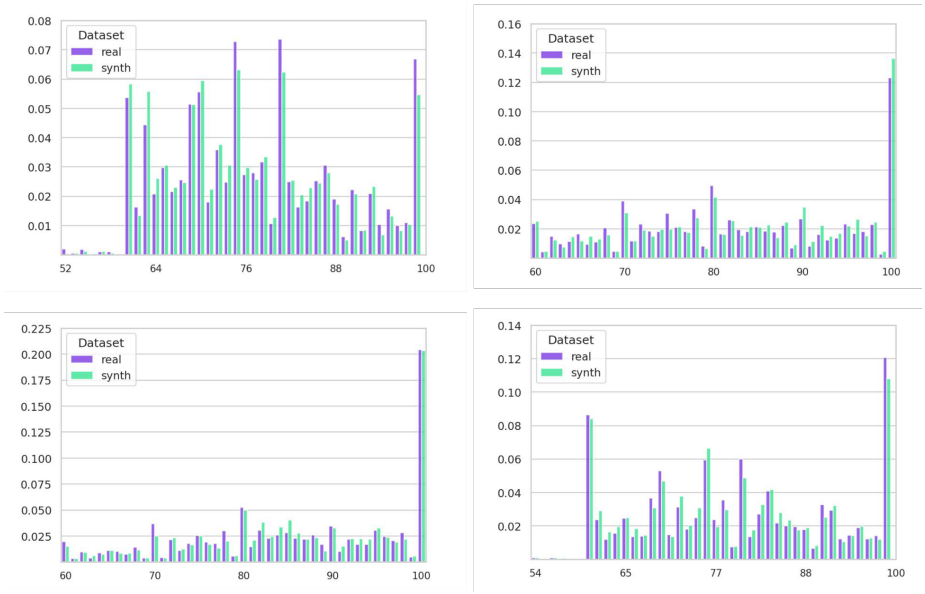


In the *Graduates' Employment Status, post graduation* data tables, employment status, contract type, and sector of activity show highly consistent proportions. The synthetic data successfully capture the dynamic evolution of graduates' employment trajectories across short-, medium-, and long-term horizons, preserving both overall proportions and structural transitions. For example, the proportion of employed individuals increases consistently over the years, and the synthetic data accurately reflects this trend (see Figure 10).

Visual inspection of *bivariate distributions* further supports the alignment between real and synthetic data. Joint distributions reproduce the main density patterns observed in the real data. Differences are generally confined to low-frequency combinations. Illustrative examples of bivariate heatmaps are shown in Figure 11 (for variables citizenship - gender) and Figure 12 (for variables university credits - exam grade).

A more granular assessment of multivariate dependency preservation is obtained by inspecting the difference between the real and synthetic *PhiK correlation* matrices. For each dataset and university, the minimum and maximum observed pairwise absolute differences are reported, along with the variable pairs

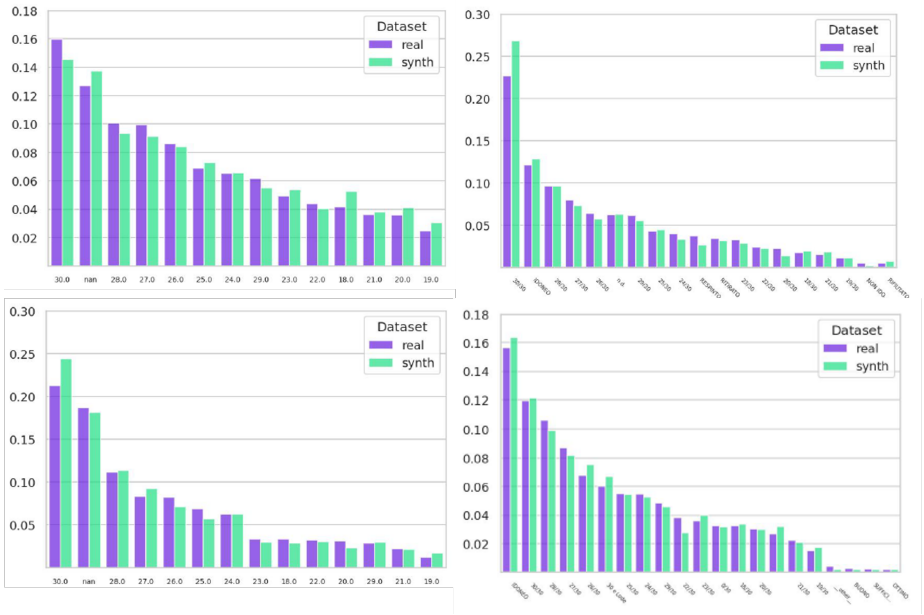
Figure 6. Distribution of the variable "high school final grade" across universities. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).



contributing to the largest deviations. Across all institutions, minimum differences are effectively zero, indicating that several dependency structures are reproduced without measurable distortion. Maximum deviations remain localized, typically concentrated in high-cardinality geographic variables, sparse categorical fields, or variables characterized by high missingness rates. A notable example is observed in the *Enrollment* table for University 4, where the largest discrepancies are associated with transfer-related variables (e.g., transfers between degree programs or from other universities), which likely correspond to relatively rare events.

Differences across universities are nonetheless evident in terms of the magnitude and distribution of the maximum deviations. University 4 exhibits the most pronounced discrepancies overall, with peak values close to 1 observed in the *College career*, *Enrollment*, and *Exams* data tables, indicating substantial distortions in specific dependency structures. University 3 also shows relatively high variability, particularly in the *College career* dataset (Max = 0.71) and in *Personal details* (Max = 0.53). In contrast, Universities 1 and 2 generally display more contained deviations across most datasets, with maximum differences typically below 0.4, although some notable exceptions emerge in the post-graduation employment datasets. In particular, the *Graduates' Employ-*

Figure 7. Distribution of the variable "exam grade" across universities. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).

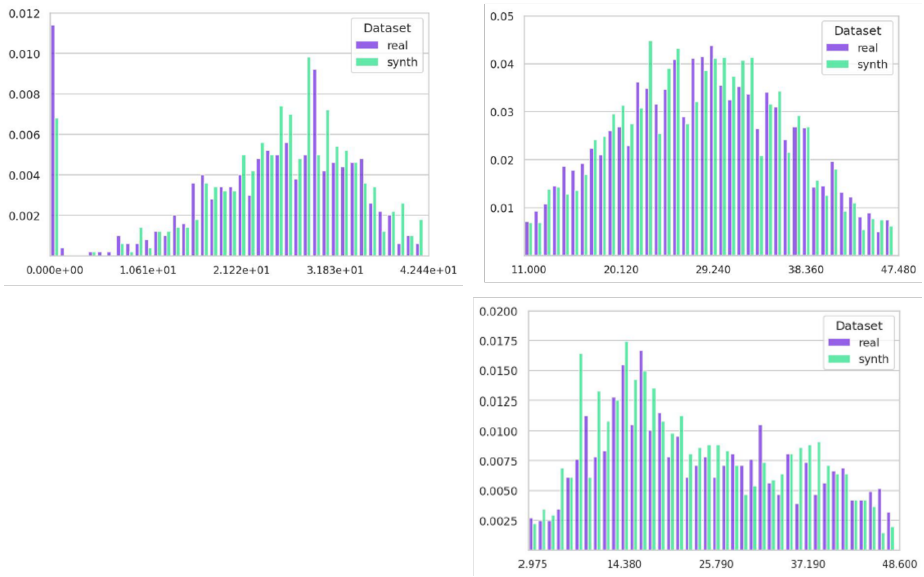


ment Status tables - especially *in 3 Years* and *in 5 Years* - show consistently higher maximum deviations across all universities, with peaks reaching 0.84 for University 1 and 0.60 for University 2.

A complementary assessment of dependency preservation can be obtained by counting the number of bivariate relationships exceeding predefined PhiK difference thresholds within each table. The tables considered include 9 variables for *Personal details* (36 possible pairs), 16 variables for *Entry test* (120 pairs), and 20 variables for all other datasets (190 pairs).

For the *Personal details* table, only a limited number of pairs exceed the 0.10 threshold across all universities, ranging from 2 (Universities 1 and 4) to 3 (University 2) and 7 (University 3). In the *College career* table, the number of pairs above 0.30–0.40 remains low for Universities 1 and 2 (2 pairs each), increases moderately for University 3 (4 pairs above 0.40), and becomes more substantial for University 4 (14 pairs above 0.40). For the *Enrollment* table, Universities 1 and 3 show very limited deviations (3 and 1 pairs above 0.15, respectively), while University 2 presents a higher number of cases (9 pairs above 0.15). A markedly different behavior is observed for University 4, where 52 pairs exceed 0.40. In the *Entry test* table, deviations are minimal: 2 pairs above 0.40

Figure 8. Distribution of the variable "overall TOLC score" across universities. The panels correspond to University 1 (upper left), University 2 (upper right), and University 4 (bottom right). Data are not available for University 3



for University 1, none above 0.40 (and only 1 above 0.10) for University 2, 1 above 0.40 for University 4, while data are not available for University 3. The *Graduates' Profile* table shows consistently low values across all institutions, with at most 2 pairs above 0.25 for University 1, 1 above 0.20 for University 2, 2 above 0.15 for University 3, and 3 above 0.25 for University 4. For the *Exams* table, the number of pairs exceeding 0.15 ranges from 1 (Universities 2 and 3) to 7 (University 1) and 8 above 0.40 for University 4. Regarding the post-graduation employment datasets, in the *1 year* table, counts range from 1 (University 4) to 2 (University 2), 4 (University 1), and 6 (University 3) above 0.30. In the *3 years* dataset, 4 (University 1), 12 (University 2), 3 (University 3), and 5 (University 4) pairs exceed 0.40. Finally, in the*5 years* dataset, 4 pairs exceed 0.50 for University 1, while no pairs exceed 0.50 for the other institutions; however, 6 pairs exceed 0.30 for University 2, 6 exceed 0.20 for University 3, and 13 exceed 0.30 for University 4.

Further details are provided in Table C.1 in Appendix C, which reports, for all nine data tables across the four universities, the number of considered variables, the observed ranges, and the corresponding most critical variables.

Figure 9. Distribution of the variable "work experience" across universities. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).

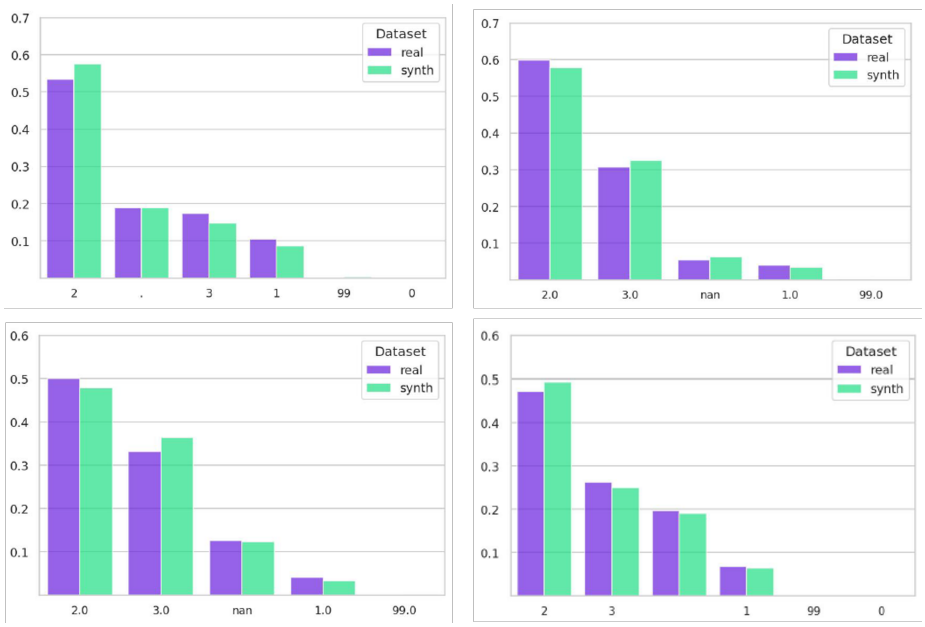


Figure 10. Distribution of the variable "employment status" across universities. Rows correspond to University 1–4 (top to bottom), while columns represent employment outcomes at 1, 3, and 5 years after graduation (left to right).

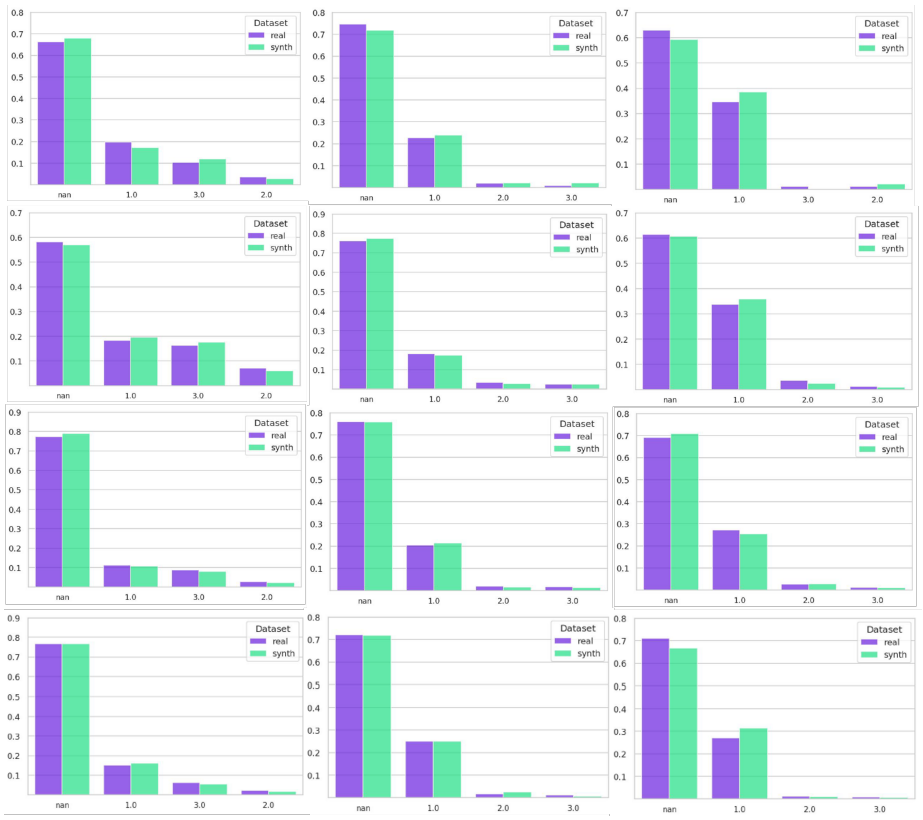


Figure 11. Bivariate heatmap for the variables citizenship - gender across universities. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).

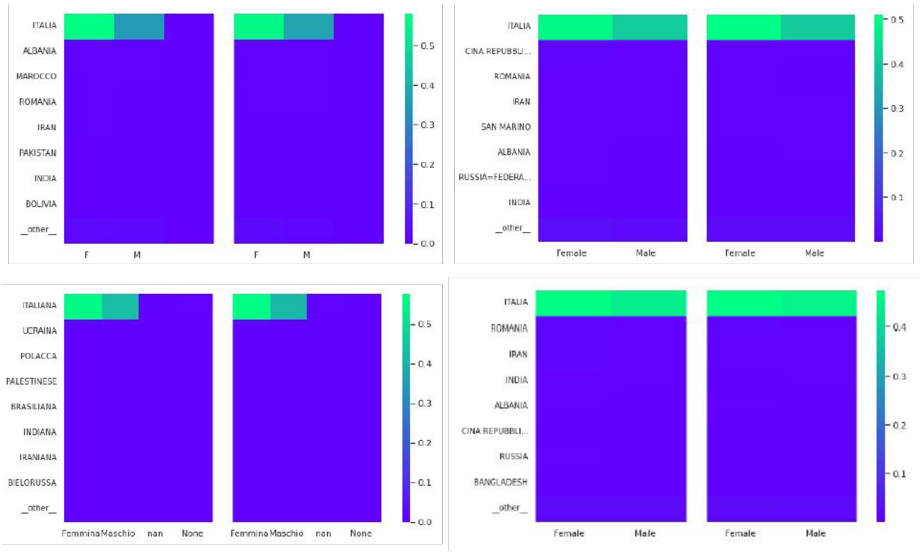
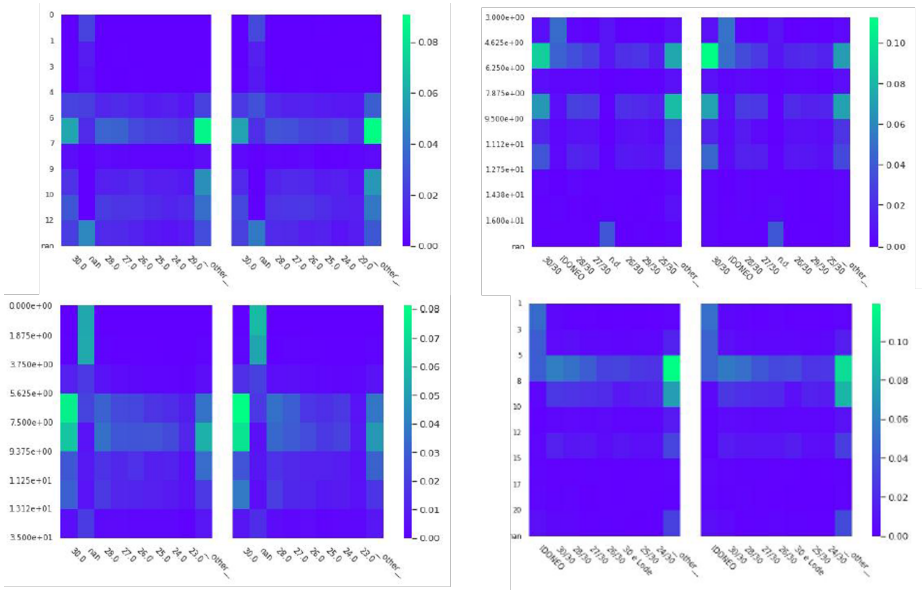


Figure 12. Bivariate heatmap for the variables university credits - exam grade across universities. The y-axis represents university credits, whereas the x-axis represents exam grades. The panels correspond to University 1 (upper left), University 2 (upper right), University 3 (bottom left), and University 4 (bottom right).



4 Syntetic data for academic research

This section aims at illustrating the performance of synthetic data with respect to real data in academic research applications. This differs from the assessment of synthetic data illustrated in Section 3.3. The previous assessment relied on indexes, while now we aim at checking the performance of synthetic data when used to carry out different types of empirical analysis, although the synthetic data were generated without targeting any specific research question and/or statistical or econometric model. We will consider three applications that differ with respect to key aspects, namely: i) data pre-processing and sample selection; ii) object of interest; iii) estimation and inferential procedures. In each application, the analysis is based on a subset of variables and records. The implemented sample selection is generally not random and is replicated using the same code on the original and synthetic data. This allows us to compare the sample size of the resulting final data set for the analysis and potentially detect deviations between the original and synthetic data.

The application discussed in Section 4.1 aims at assessing the peer effects among college students using linear regression models. It focuses on conditional expectations of a subset of outcomes imposing a specific parametric structure. It aims at detecting inter-dependencies between outcomes of different students who have in common the same set of *peers* (*causal inference on a single specific parameter of the regression*). It uses method of moments estimators and it relies on asymptotic theory for inference, conditional on the realized sample for both original and synthetic data.

The application discussed in Section 4.2 aims at examining predictors of drop-out in the early stages of students' academic careers. It aims at assessing prediction accuracy and identifying what are the variables with highest predictive power. The analysis relies on non-linear regression models (logit) and lasso (least absolute shrinkage and selection operator). The model is estimated on a subset of the dataset (the training dataset) and then used to predict the drop-out probability for the reminaing part (the test dataset).

While these two applications both exploit administrative data, the application discusses in Section 4.3 focuses on survey (Almalaurea) data, with the aim to identify distinct profiles of graduates based on their job aspirations. A clustering procedure is applied separately to real and synthetic datasets using variables related to job preferences. Clusters obtained with real and synthetic data are then compared in terms of centroid similarity, cluster size, and substantive interpretation.

While these three applications are far from comprehensively accounting for all possible applications, we argue they offer a reasonably large span of applications as they cover different objectives, with different estimation and inferential procedures.

Each application is describe in a separate subsection, where we summarize the objective of the analysis, the data pre-processing (cleaning and sample selection) and estimation and inference. For each application, we start by comparing real and synthetic data. We first explore whether there are issues in replicating features of the selected sample used in each application. We find that synthetic data generally reproduce marginal distributions, consistently we what is shown in section 3.3 of the report. We then compare results of each application within University - between real and synthetic data - and between Universities, whenever possible. We discuss whether differences can be attributed to hard features of the synthetic data generating process and what are the lessons learned.

4.1 Peer effects among college students

Exploiting within-degree, across-cohort variation—a standard approach in the literature—we study how having relatively higher-quality peers in a cohort affects a range of individual outcomes. We define a degree (d) as a specific first-cycle program offered in multiple years (t) (e.g., "Economics"). A cohort is a degree–year pair (dt).

Data Cleaning We apply identical data-cleaning procedures and sample restrictions to each data version and across universities. From the exam records, we construct a credit-weighted average of each student's exam grades over their career in while enrolled in a degree. We then merge these measures to the enrollment files at the student level. From the enrollment data, we collect the first year of enrollment, the cycle of study (e.g., bachelor's or master's), the degree type (*classe di laurea*), and high-school grade (*voto di maturità*). We also retrieve demographic characteristics, including gender, scholarship receipt, and migrant status. At the cohort level, we compute cohort size and the shares of different student types (e.g., female, migrant). Finally, for each cohort we compute the leave-one-out mean of peers' high-school grades:

$$\text{PeerAbility}_{idt} = \frac{\sum_{j \in dt, j \neq i} \text{Ability}_j}{N_{dt} - 1}, \tag{6}$$

where N_{dt} denotes cohort size.

We impose the following sample restrictions. First, we retain only students enrolled in first-cycle curricular programs, including bachelor's degrees and integrated master's degrees. Second, to allow outcomes to be observed, we restrict attention to students first enrolled between 2012 and 2020 for three-year bachelor's programs, and between 2012 and 2018 for five- or six-year integrated master's programs. Third, we keep only degrees that admit students in at least two years within the analysis window. We then construct completion outcomes—whether the student ever graduates, graduates on time, and final

graduation grade. Finally, we drop students with missing high-school grades.

Main Specification We estimate the effect of cohort peer quality on individual outcomes using the following OLS specification:

$$Y_{idt} = \alpha + \lambda_d + \theta_t + \beta \text{PeerAbility}_{idt} + \gamma \text{Ability}_{idt} + \mathbf{X}_{idt}\delta' + \varepsilon_{idt}. \quad (7)$$

Y_{idt} denotes the outcome of student i enrolled in degree d and cohort year t. The term PeerAbility_{idt} is the leave-one-out average of peers' pre-college ability in cohort dt, constructed from high-school grades as described above. To aid interpretation, we include a standardized version of this variable.

We include degree fixed effects, λ_d, which absorb all time-invariant differences across degrees. Year fixed effects, θ_t, capture common shocks that affect all degrees in a given enrollment year. The coefficient of interest is β, which is identified from variation in peer quality across cohorts within the same degree. Under the standard assumption that, conditional on degree and year fixed effects, cohort-to-cohort differences in peer ability within a degree are as good as random, β can be interpreted as causal.

To account for the mechanical correlation between peer ability and own ability, and to improve precision, we control for the student's own pre-college ability, Ability_{idt} (high-school GPA). We further include a vector of baseline covariates $\mathbf{X}_{idt}$—such as gender, scholarship status, migrant status, and, depending on the specification, other predetermined cohort-level characteristics available at the time of enrollment, so that identification comes from residual variation in peer quality across cohorts conditional on observable student composition.

For inference, we cluster standard errors at the cohort level (dt) to allow for arbitrary correlation in unobservables among students who enter the same degree in the same year.

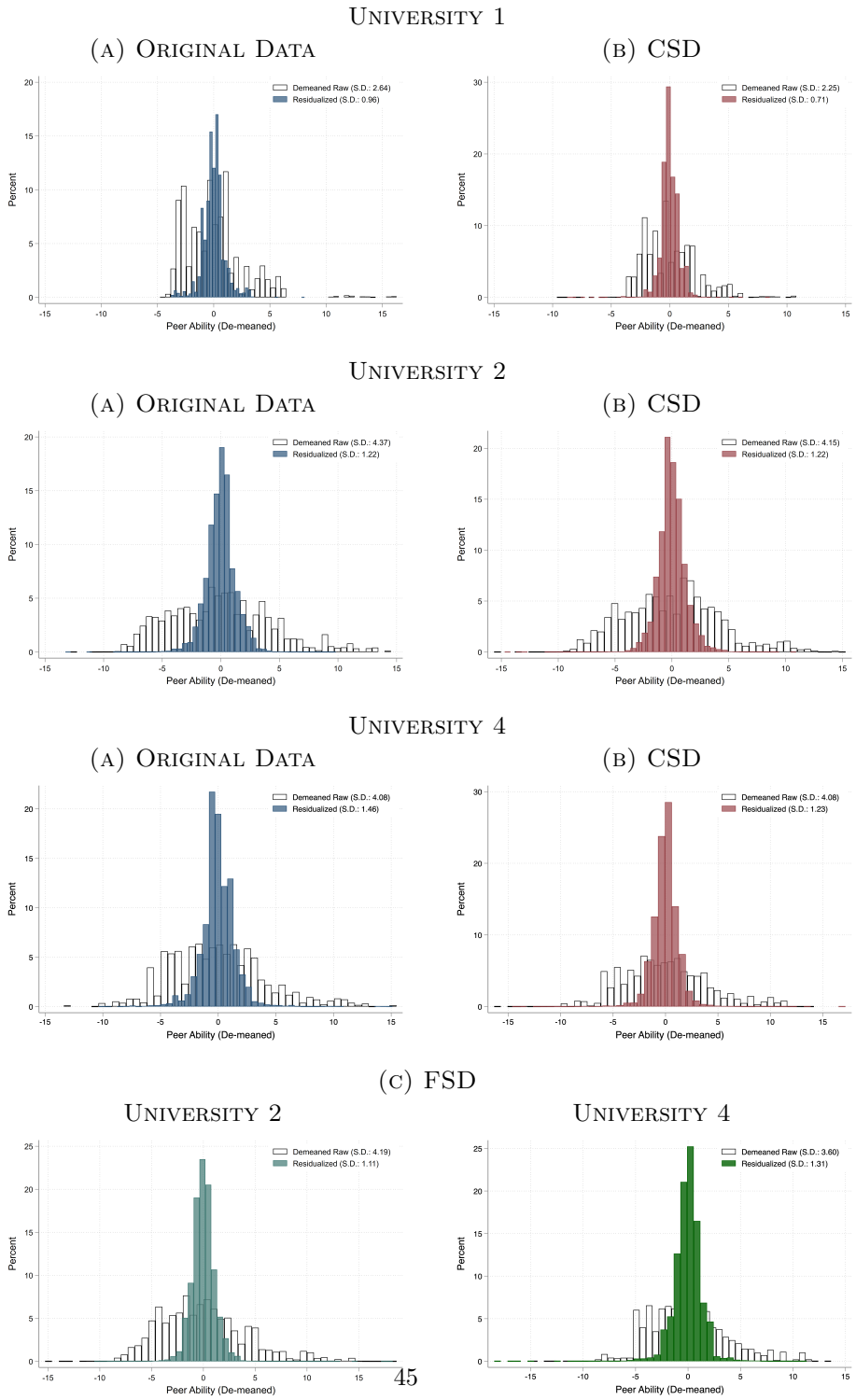
Summary Statistics Summary statistics for each University are reported in Appendix D. Specifically, Table D.2, D.3 and D.4 report summary statistics for the original administrative data and the synthetic ones (CSD and FSD, when available). In each of these three case studies, the original data and CSD are very similar in sample size. Some departures are observed in exam grades, which is available for a smaller number of observations in the synthetic data. Location and scale of marginal distribution of the selected variables are similar, we no sizable difference, regardless the considered variable is binary, discrete or has continuous support. By contrast, FSD is substantially different from the original data, particularly in the outcomes and overall ability levels. Graduation and on-time graduation are substantially lower in the FSD compared to original data in the two cases studies, while dropout is higher, suggesting that this

synthetic version does not preserve the outcome distribution as closely. Ability is also shifted down – both for own and peer ability average –, and exam grades are slightly lower in one of the two case studies. There are also differences in demographic composition: the share of female students is lower, migrants are somewhat higher, and prior degrees are lower. Cohort-level covariates remain broadly comparable in levels. The large differences in outcomes indicate that results relying on FSD may be less directly comparable to the original-data estimates than those based on CSD. It is notable that the deviations between original and FSD are generally consistent in the two cases, hinging on systematic biases induced by this alternative synthetic data generation process. Whether this differences can lead to substantial biases in the empirical analysis should be further explored.

Figure 13 provides a complementary view of how closely the original and synthetic datasets match not only in their unconditional moments, but also in the underlying identifying variation of the key regressor of interest. It depicts variation in cohort peer ability by comparing, across datasets, the distribution of the demeaned leave-one-out peer-ability measure (white bars) to the distribution of the same variable after residualizing it on own ability, degree fixed effects, and year fixed effects (colored bars), as in Equation (7). As identification of the peers effects in Equation (7) requires variation in peer ability conditional on own ability, and within degree and year, these figures allow us to check if there sufficient source of variation in the data, once we condition on the same covariates required for identification in our model, and if such "residual" variation is similar between original and synthetic data. Figure 13 includes 8 panels over four rows. In the top part of the figure (row 1- row 3), we compare original data to CSD for the three available case studies for University 1, 2 and 4. The last row presents FSD for University 2 and 4.

In all panels, residualization sharply compresses the distribution, indicating that most of the raw dispersion in peer ability is absorbed by systematic differences across degrees, cohorts, and students' own pre-college ability. The extent of residual variation is very similar in the original data (Panel A) and CSD (Panel B), with an almost identical residualized standard deviation, while FSD (Panel C) exhibits a slightly tighter residual distribution. Overall, the figure confirms that the synthetic datasets replicate not only the unconditional dispersion of peer ability, but also the amount of within-degree, across-cohort residual variation used to estimate the effects.

Figure 13. Residual Variation in Peer Ability Across Datasets



Notes. The white bars depict the distribution of the demeaned version of the individual leave-out mean of peer ability as defined in Equation 6. The colored bars represent the distribution of the same peer ability variable residualized of own ability, degree fixed effects, and year fixed effects, as in the specification in Equation 7. Panel A depicts the distributions for the original data. Panels B and C depict the distributions for the synthetic data, Centralized Synthetic Data and Federated Synthetic Data, respectively.

Results Tables D.5, D.6 and D.7 in the Appendix report the estimated effect of peer ability on student's academic performance for University 1, 2 and 4, respectively. We consider measured with four main outcomes: the probability to graduate, the probability to graduate on time, the probability to drop-out and the average exam grades (GPA). The results are summarized in Figure 14 where we present the normalized coefficients of the main regressor of interest (*peer ability*) across the different available data set and across the three available case studies. Across datasets in University 2 and 4, figures deliver a consistent qualitative interpretation: higher peer ability improves educational attainment and reduces dropout. Crucially, the magnitude of the effects differs in ways that reflect the differences in unconditional moments presented in Table D.3, and D.4. For University 1, we observe significant departures between the analysis based on the original data and the synthetic data. Synthetic data mirror the results obtained in University 2 and 4. As synthetic data underperform in replicating outliers while conditional averages are sensitive to outliers, these difference might be justified with the presence of outliers in the original data that cannot be replicated in the synthetic data.

In the original data (Panel A) for University 2 and University 4, a one-standard-deviation increase in cohort peer ability raises the probability of graduating (graduating on time) by about 5% (7%) relative to outcome mean, while reducing dropout by approximately 10% relative to mean drop-out levels. With few exceptions, effects are precisely estimated and the confidence intervals exclude zero. CSD (Panel B) closely tracks these patterns and, if anything, slightly amplifies them. Taken together, the synthetic version that best matches the original moments also produces estimates that are very similar in sign and broadly comparable in magnitude, suggesting that the qualitative interpretation of the peer effect remains unchanged.

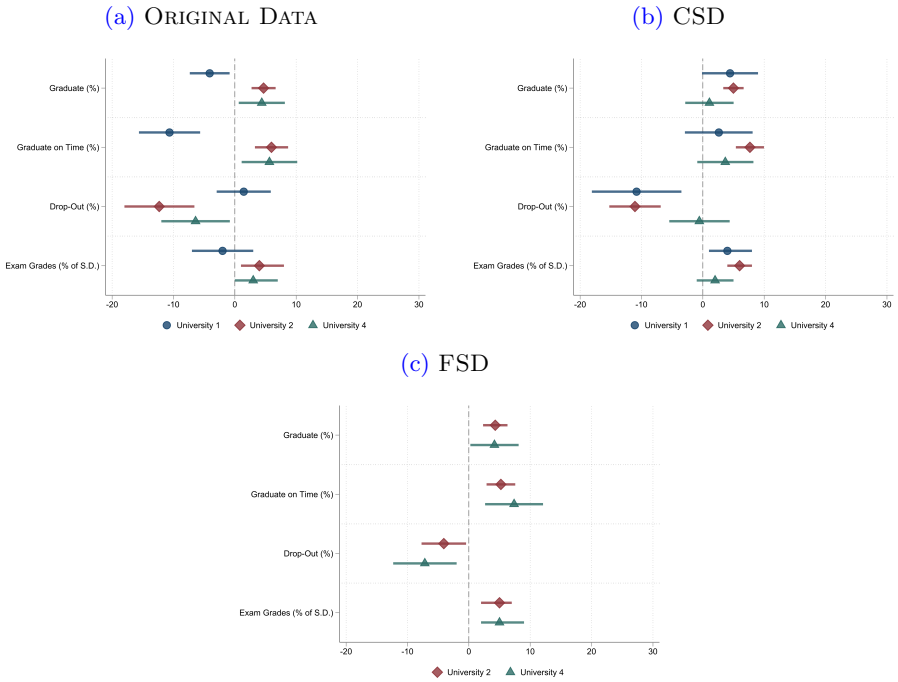
FSD (Panel C) is the outlier, again in line with its weaker match to the original outcome distributions. Although the effects remain in the expected direction, they are attenuated: the estimated impacts on graduation and on-time graduation fall slightly for University 2 while they move slightly up for University 4, and the dropout reduction shrinks (still statistically different from zero, but much smaller). This pattern of smaller attainment and dropout effects alongside similar or slightly altered grade effect, suggests that Synthetic 2 preserves some of the within-degree, across-cohort variation in peer ability (as also seen in Figure 13), but does not replicate the joint distribution linking peer quality to the extensive-margin outcomes as tightly as Synthetic 1 does.

University 1 is characterized by larger cohort size compared to University 2 and 4 and a smaller residual variation in peer ability both in the original and the synthetic data. The regression results based on original and synthetic data differ substantially. Estimated peer effects are not statistically significant on original

data (see Table D.5) when one looks at the peer effects on the probability of drop-out or on standardized exam grades. Furthermore, an increase in peer ability is associated with a reduction in the probability to graduate and to graduated on time.

Observed differences between the results based on original data between University 1 and University 2 and 4 might be related to institutional features and the relationship between these features and the empirical analysis of this section. Indeed, note that we define the relevant peer group based on first year enrollment. In University 2 and 4, students enrolled simultaneously in the first year tend to attend compulsory/chore classes together. Conversely, in University 1, classes are split starting from first year into curriculum-specific classes. As our data do not include curriculum-specific identifiers within the same degree, we are unable to define peers at the curricula-level. As a consequence, it is possible that the type of interactions individuals have during college in University 1 differ from those documented in University 2 and 4 and this is reflected in parameters estimates.

Figure 14. Normalized Coefficients Peer Analysis - Comparison Across Universities and Datasets



Notes. Each panel plots the normalized peer effect coefficients with 95% confidence intervals, estimated separately for three universities. For Graduate (%), Graduate on Time (%), and Drop-Out (%), the normalized coefficient is computed as the estimated coefficient divided by the mean of the dependent variable, expressed in percentage terms. For Exam Grades, the coefficient is expressed as a percentage of a standard deviation. The dependent variable means used for normalization are university- and dataset-specific. Panel A uses original administrative data, panel B uses CSD (Centralized Synthetic Data), and panel C uses FSD (Federated Synthetic Data). University 1 is not included in panel C as FSD is not available for this university. All specifications include individual covariates, year fixed effects, and degree fixed effects, and are estimated according to the specification in Equation 7.

4.2 Predicting drop-out

In this application, we analyze the predictors of drop-out in early stages of students' academic careers. Understanding the mechanisms that shape early academic trajectories is essential at the higher-education level, particularly in identifying the pre-existing factors and initial academic characteristics that mostly influence students' initial achievement and potential academic difficulties. An accurate prediction of the drop-out risk is essential to provide targeted services and support well before the event occurs, with the aim to prevent such risk. Our analysis is conducted at the student level and focuses on first-year students enrolled in either a bachelor's degree or a five-year or 6-year degree program, when the drop-out risk is the highest.

Data Cleaning We apply the same routine of data cleaning and sample restrictions to both original and synthetic datasets, and across universities. From student enrollment data, we retrieve demographics characteristics, including gender, country of birth, place of residence, high school type and final high school grade, as well as scholarship status at the University. From exam records, we build a credit-weighted average of students' grades over the course of their first year and first semester; similarly, we also construct the total number of credits acquired over the course of the first year and the first semester. As outcomes, we construct three indicators of dropout or student inactivity. The first one, *Dropout*, is a binary indicator taking value equal to 1 if the student's enrollment status is listed as "Withdrawn from studies" or "Transferred out".[4] While this is the standard measure of explicit drop-out, we should consider also that some students do not take (or pass) any exam in their first academic year. To take into account of this "implicit" drop-out, we compute also a broader indicator, labelled *BroadDropout*, which is equal to 1 if the student has either dropped out or has not registered credits during the course of the first academic year. Finally, we consider a third outcome, labelled *LowActivity*, taking value equal to 1 if the student has passed exams for less than 12 credits at the end of the first year of study.

Main Specification We estimate the effect of individual and academic characteristics on the probability that the student drops out or shows signs of low activity before the end of the first year of study. We use the following specification:

$$Y_{idt} = \alpha + \mathbf{X}_{idt}\beta' + \mathbf{Z}_{idt}\beta'' + \varepsilon_{idt}. \tag{8}$$

[4] We exclude from this definition all early dropouts, i.e., students who have dropped out or transferred out before the end of the first semester.

Y_{idt} denotes one of the three drop-out outcomes discussed above (i.e., dropout, broad dropout or low activity) of student i enrolled in degree d in year t. The vector $\mathbf{X}_{idt}$ includes a series of individual covariates, e.g., gender, high school performance, country of birth (Italy or other), and scholarship status, whereas the vector $\mathbf{Z}_{idt}$ controls for academic performance in the first semester of the degree, i.e., the mean score and the total number of credits acquired by the end of the first semester. We split the sample into a training and a test set. Individuals belonging to cohort years up to a.y. 2023/2024 are assigned to the training set, while students enrolled in a.y. 2024/2025 are included in the test set. For each dropout outcome, we train the model on the training sample and evaluate predictive accuracy using the test sample. We first employ a logit model and then estimate a lasso logit model to perform variable selection and identify the most relevant explanatory variables. We compare the results of the two models in terms of prediction accuracy across the original and synthetic datasets. Additionally, we compare the set of selected covariates and their rankings obtained from the lasso estimation in both datasets.

Summary Statistics Tables E.8, E.9, and E.10 in the Appendix report the main summary statistics for the original and the synthetic datasets (Synthetic 1 and Synthetic 2, when available). The number of observations is slightly different in synthetic data compared to the original one, but most statistics are very similar between the original and the CSD.[5] Among the outcomes, the only exception is the share of students with a low number of credits at the end of the first year: this is consistently higher in the CSD across all case studies, especially in University 1. Some covariates also show minor differences between the original and synthetic data, including the share of female students, non-resident students, and students from the academic track (*liceo*). However, these differences do not exhibit a consistent pattern across the three case studies.

Predictive power and main predictors Table 4 compares the predictive power of both logit and lasso logit models estimated with the three outcomes and the two datasets. For each University, dataset and model, the Table reports two indicators: the Area Under the ROC Curve (AUC) and Accuracy. Within each dataset and across all outcomes, there are no substantial differences in predictive performance between the logit and the lasso model, regardless of the metric considered. Comparing results across datasets, differences in classification accuracy are generally limited. The main exception concerns the lasso model for the prediction of low activity, for which all three Universities exhibit higher ac-

[5] Compared to original data, total number of observations is higher in synthetic data for University 1 and 2, lower for University 4. The lower number of observations for the mean grade (score) is because it is computed only for students who passed at least one exam.

curacy when using the original data. By contrast, the AUC is systematically higher for models estimated on the original data, with sizable differences, often between 8 and 10 percentage points, particularly for the *Dropout* and *Broad Dropout* outcomes.

Across all Universities and datasets, predictive performance is substantially higher for the low activity outcome, with AUC values close to one. This suggests that low activity is considerably easier to predict than dropout.

Taken together, these results indicate that, while classification performance at the chosen threshold (0.5) is broadly similar across datasets, models estimated on the original data exhibit a superior ability to rank individuals by risk. This pattern is consistent with a lower dispersion or a reduced presence of extreme observations in the synthetic data, which may limit its ability to distinguish between high- and low-risk students.

Table 4. Model Performance by Outcome

Panel A: University 1								
Outcome	Model	AUC (Orig)	Acc (Orig)	AUC (CSD)	Acc (CSD)	AUC (FSD)	Acc (FSD)	N Test
Broad Dropout	Logit	0.6905	0.9798	0.5913	0.9627			4108
Broad Dropout	Lasso	0.6996	0.9798	0.5874	0.9629			4108
Dropout	Logit	0.6905	0.7583	0.5912	0.7378			4108
Dropout	Lasso	0.6996	0.7583	0.5873	0.7380			4108
Low Activity	Logit	0.9896	0.7072	0.9711	0.6487			4108
Low Activity	Lasso	0.9905	0.7079	0.9739	0.6470			4108
Panel B: University 2								
Outcome	Model	AUC (Orig)	Acc (Orig)	AUC (CSD)	Acc (CSD)	AUC (FSD)	Acc (FSD)	N Test
Broad Dropout	Logit	0.7231	0.9266	0.6223	0.9142	0.5872	0.9067	16618
Broad Dropout	Lasso	0.7173	0.9266	0.6279	0.9142	0.5915	0.9067	16618
Dropout	Logit	0.7231	0.8042	0.6223	0.8011	0.5872	0.7797	16618
Dropout	Lasso	0.7173	0.8042	0.6279	0.8011	0.5915	0.7797	16618
Low Activity	Logit	0.9617	0.7868	0.9348	0.7289	0.9223	0.7214	16618
Low Activity	Lasso	0.9612	0.7867	0.9341	0.7292	0.9212	0.7216	16618
Panel C: University 4								
Outcome	Model	AUC (Orig)	Acc (Orig)	AUC (CSD)	Acc (CSD)	AUC (FSD)	Acc (FSD)	N Test
Broad Dropout	Logit	0.5925	0.7255	0.5769	0.7088	0.6181	0.7162	8379
Broad Dropout	Lasso	0.6200	0.8555	0.5695	0.8277	0.6213	0.8416	8379
Dropout	Logit	0.5925	0.5497	0.5765	0.5481	0.6175	0.5510	8379
Dropout	Lasso	0.6200	0.6797	0.5692	0.6671	0.6211	0.6764	8379
Enrolled w Low Credits	Logit	0.9428	0.7096	0.9258	0.6865	0.9205	0.6906	8379
Enrolled w Low Credits	Lasso	0.9664	0.7839	0.9270	0.7570	0.9287	0.7506	8379

Notes: The table reports predictive performance for logit and lasso logit models across three outcomes: *Dropout*, *Broad Dropout*, and *Low Activity*. Models are trained on the training sample (cohorts up to a.y. 2023/2024) and evaluated on the test sample (a.y. 2024/2025). Performance is measured using the Area Under the ROC Curve (AUC) and classification accuracy (Acc). Results are shown for the original dataset (Orig) and for two synthetic datasets (Centralized Synthetic Data (CSD) and Federated Synthetic Data (FSD), when available). The set of covariates includes individual characteristics (gender, high school performance, country of birth, scholarship status) and first-semester academic performance (average grade and total credits earned). N Test denotes the number of observations in the test sample.

As a further check, we compared the most relevant variables selected by Lasso model in the two datasets. For two of the three outcomes considered, Tables 5 and 6 report the list of the most important variables and the corresponding estimated coefficients for two of the outcomes considered [6]. Overall, the Lasso model selects the same variables in the two datasets, and the estimated standardized coefficients are very similar across datasets. A few exceptions emerge for University 1 when the outcome variable is *Dropout*. Interestingly, while early academic performance (especially the average grade at the end of the first semester) is selected as an important predictor across all the Universities, the selected socio-demographic characteristics are more heterogeneous. For example, the type of high school is a relevant predictor in University 1 and 2, while having a scholarship emerges as the first selected predictor only in University 4.

Table 5. LASSO Variable Importance Rankings - Dropout

Panel A: University 1					
Original	Coef (SE)	CSD	Coef (SE)	FSD	Coef (SE)
Liceo	-0.364 (0.035)	Scholarship	-1.274 (0.158)		
Total Credits first semester	-0.090 (0.003)	Liceo	-0.292 (0.034)		
Mean Score first semester	-0.059 (0.005)	Female	-0.198 (0.033)		
High school grade	-0.015 (0.002)	Mean Score first semester	-0.096 (0.005)		
		Total Credits first semester	-0.064 (0.003)		
Panel B: University 2					
Original	Coef (SE)	CSD	Coef (SE)	FSD	Coef (SE)
Liceo	-0.294 (0.018)	Liceo	-0.324 (0.019)	Scholarship	-0.234 (0.025)
Total Credits first semester	-0.078 (0.001)	Mean Score first semester	-0.099 (0.003)	Liceo	-0.183 (0.016)
Mean Score first semester	-0.062 (0.003)	Total Credits first semester	-0.046 (0.001)	Mean Score first semester	-0.085 (0.002)
High school grade	-0.014 (0.001)	High school grade	-0.019 (0.001)	Total Credits first semester	-0.044 (0.001)
				High school grade	-0.011 (0.001)
Panel C: University 4					
Original	Coef (SE)	CSD	Coef (SE)	FSD	Coef (SE)
Scholarship	-0.513 (0.054)	Scholarship	-0.521 (0.054)	Scholarship	-0.553 (0.058)
Female	-0.259 (0.024)	Female	-0.195 (0.024)	Female	-0.235 (0.026)
Mean Score first semester	-0.079 (0.004)	Mean Score first semester	-0.082 (0.004)	Mean Score first semester	-0.101 (0.004)
Total Credits first semester	-0.049 (0.002)	Total Credits first semester	-0.030 (0.002)	Total Credits first semester	-0.033 (0.002)
High school grade	-0.006 (0.001)	High school grade	-0.010 (0.001)	High school grade	-0.006 (0.001)

Notes: The table reports the ranking of the most relevant covariates selected by the lasso logit model for the prediction of *Dropout*. Variables are ordered by importance based on the magnitude of their estimated coefficients. Coefficients and standard errors (in parentheses) are reported for each selected variable. Results are shown for the original dataset (Original) and for the synthetic datasets (Centralized Synthetic Data and Federated Synthetic Data, when available). The set of potential covariates includes individual characteristics (gender, high school performance, country of birth, place of residence, scholarship status, and high school type) and first-semester academic performance (average grade and total credits earned).

[6] Estimates for broad dropout are very similar to those obtained with the standard indicator of dropout and are available upon request.

Table 6. LASSO Variable Importance Rankings - LOW ACTIVITY

Panel A: University 1					
Original	Coef (SE)	CSD	Coef (SE)	FSD	Coef (SE)
Total Credits first semester	-0.322 (0.007)	Total Credits first semester	-0.278 (0.007)		
Mean Score first semester	-0.100 (0.007)	Mean Score first semester	-0.110 (0.006)		
		High school grade	-0.022 (0.002)		

Panel B: University 2					
Original	Coef (SE)	CSD	Coef (SE)	FSD	Coef (SE)
Scholarship	-1.328 (0.056)	Scholarship	-0.487 (0.039)	Scholarship	-0.644 (0.033)
Liceo	-0.502 (0.026)	Liceo	-0.432 (0.022)	Liceo	-0.426 (0.019)
Total Credits first semester	-0.289 (0.003)	Total Credits first semester	-0.250 (0.003)	Total Credits first semester	-0.244 (0.003)
Mean Score first semester	-0.072 (0.004)	Mean Score first semester	-0.103 (0.003)	Mean Score first semester	-0.093 (0.003)
High school grade	-0.021 (0.001)	High school grade	-0.028 (0.001)	High school grade	-0.024 (0.001)

Panel C: University 4					
Original	Coef (SE)	CSD	Coef (SE)	FSD	Coef (SE)
Scholarship	-1.481 (0.056)	Liceo	-0.359 (0.033)	Total Credits first semester	-0.236 (0.005)
Total Credits first semester	-0.274 (0.002)	Total Credits first semester	-0.242 (0.005)	Mean Score first semester	-0.097 (0.005)
		Mean Score first semester	-0.044 (0.005)		
		High school grade	-0.015 (0.001)		

Notes: The table reports the ranking of the most relevant covariates selected by the lasso logit model for the prediction of *Low Activity* (enrolled with low credits). Variables are ordered by importance based on the magnitude of their estimated coefficients. Coefficients and standard errors (in parentheses) are reported for each selected variable. Results are shown for the original dataset (Original) and for the synthetic datasets (Centralized Synthetic Data and Federated Synthetic Data, when available). The set of potential covariates includes individual characteristics (gender, high school performance, country of birth, place of residence, scholarship status, and high school type) and first-semester academic performance (average grade and total credits earned).

4.3 Graduates' profiles based on future job aspirations

The present analysis complements the previous applications by extending the assessment of synthetic-data utility to a more exploratory multivariate setting. Its aim is to verify whether the synthetic data are able to preserve the same profile structure that emerges from the original data when graduates are grouped according to their work aspirations at the end of their degree programme[7].

The transition from university to the labour market is shaped not only by objective employment opportunities, but also by the system of aspirations, expectations, preferences, and constraints through which graduates imagine their future work. In this respect, the AlmaLaurea Profilo survey is especially useful because it conceptualises labour-market orientation as a multidimensional phenomenon, covering the importance attributed to different job characteristics, the willingness to accept specific contractual arrangements, the geographical areas in which one is willing to work, and a broader attitudinal orientation towards employment.

This application focuses specifically on the set of questions concerning the aspects considered relevant in job search (hereafter referred to as aspirations). The goal is twofold: first, to identify distinct profiles of graduates' work aspirations; second, to assess whether these profiles are reproduced with sufficient fidelity in the synthetic dataset. Cluster interpretation is then enriched by means of supplementary variables describing socio-demographic background, mobility preferences, and willingness to accept different contractual arrangements. The clustering procedure is estimated separately on real and simulated data, and the resulting solutions are compared in terms of centroid proximity, cluster size, and substantive interpretation.

4.3.1 Data

The clustering solution was estimated on sixteen variables measuring the aspects considered relevant in job search (job aspirations). These variables capture respondents' evaluation of different desirable characteristics of future work and constitute the core of the latent preference structure underlying graduates' orientation towards future employment. Specifically, the cluster analysis was based on the following items:

- `asp1`: possibility of earning a good income;
- `asp2`: career opportunities;
- `asp3`: job stability and security;
- `asp4`: opportunity to acquire professional skills;

[7] This application is currently available only for University 3 due to time constraints

- `asp5`: coherence with previous studies;
- `asp6`: correspondence with cultural interests;
- `asp7`: independence or autonomy;
- `asp8`: free time;
- `asp9`: social usefulness of the job;
- `asp10`: job prestige;
- `asp11`: involvement and participation in work activity and in decision-making processes;
- `asp12`: flexibility of working hours;
- `asp13`: relationships with colleagues in the workplace;
- `asp14`: workplace location and physical environment;
- `asp15`: opportunities for contacts abroad;
- `asp16`: opportunity to make full use of the competences acquired.

Responses were recorded on a four-point ordinal scale: *definitely yes* (= 5), *more yes than no* (= 4), *more no than yes* (= 2), and *definitely no* (= 1). Taken together, these indicators span several theoretically distinct dimensions, including economic returns, professional development, intrinsic motivation, work–life balance, social meaning, organisational climate, and international orientation.

The interpretation of the clusters was enriched through a set of supplementary variables that were not used to construct the clusters but were analysed ex post in order to characterise them. A first group includes socio-demographic and background characteristics: sex (`sesso`), disciplinary field of study (`gruppostat`), social class (`class_soc`), and age at graduation (`etalau_c`). More specifically, sex (`sesso`) was coded as 1 = male and 2 = female; social class (`class_soc`) as 1 = upper class, 2 = white-collar middle class, 3 = self-employed middle class, 4 = working class, and 99 = not reported; disciplinary field of study (`gruppostat`) as a sixteen-category classification including scientific, chemistry-pharmacy, geo-biological, medical, engineering, architecture, agriculture and veterinary studies, economics-statistics, political-social studies, law, literature, languages, teaching, psychology, physical education, and defence and security; and age at graduation (`etalau_c`) as 1 = under 23, 2 = 23–24, 3 = 25–26, and 4 = 27 or older.

A second group includes variables describing willingness to work in different geographical areas: the province of study (`gradoa10`), the region of study (`gradoa11`), the province of residence (`gradoa2`), Northern Italy (`gradoa5`), Central Italy (`gradoa6`), Southern Italy (`gradoa7`), another European country

(`gradoa8`), and a non-European country (`gradoa9`). These variables are measured on the same four-point response scale: *definitely yes* (= 5), *more yes than no* (= 4), *more no than yes* (= 2), and *definitely no* (= 1). Values of 0 and 99 indicate missing responses.

A third group concerns willingness to accept different contractual and organisational arrangements: apprenticeship (`dlavapp2`), self-employment or freelance work (`dlavauto2`), a fixed-term contract (`dlavdet2`), agency or temporary work (`dlavinter2`), part-time work (`dlavpartim2`), full-time work (`dlavpieno2`), an internship or traineeship (`dlavstage2`), and a contract with increasing employment protection (`dlavtutele`). These variables are again measured on the same four-point scale ranging from *definitely yes* (= 5) to *definitely no* (= 1). Values of 0 and 99 indicate missing responses.

Only students enrolled in a master's degree or a single-cycle master's degree at University 3 were included in the analysis. This restriction is substantively justified because second-level graduates are generally closer to labour-market entry and are therefore more likely to have developed clearer and more structured occupational expectations.

4.3.2 Procedure

The same analytical strategy was applied to both the original and the simulated datasets. Cluster analysis was performed using the k-means algorithm in order to identify homogeneous groups of graduates characterised by similar profiles of work aspirations. K-means is a non-hierarchical partitioning technique that assigns each observation to one and only one cluster by minimising the within-cluster sum of squares, that is, the total variability of observations around the corresponding cluster centroid. Before estimation, observations with missing values on at least one of the sixteen aspiration indicators were excluded, since k-means does not handle incomplete cases directly.

The number of clusters was selected on the basis of the original data by inspecting the within-cluster sum of squares for solutions ranging from $k = 1$ to $k = 6$. The solution identified through this criterion was then applied to both the original and the synthetic datasets in order to preserve a directly comparable framework. K-means was subsequently estimated separately on the two datasets using the same number of centres and multiple random starts.

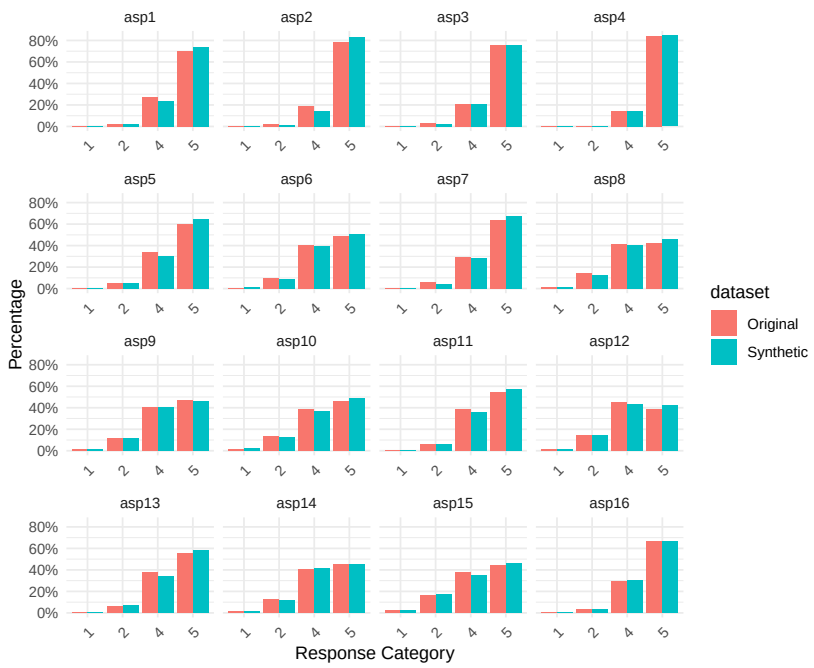
Because the two clustering solutions were estimated independently, the cluster labels in the original and the synthetic data are not directly comparable. For this reason, a matching procedure was introduced after estimation. The centroids obtained from the two datasets were compared through Euclidean distances, and the Hungarian assignment algorithm was applied through the function `solve_LSAP` in order to identify the optimal one-to-one correspondence between original and synthetic clusters. After matching, the comparison be-

tween the two solutions was based on: *(i)* the distance between centroids; *(ii)* the distribution of observations across clusters; and *(iii)* the substantive interpretation of clusters.

4.3.3 Results

For the sixteen aspiration variables used in the clustering procedure, the response distribution is consistent across the Original and Synthetic datasets: in all items, the highest shares are concentrated in the positive categories, while the negative categories remain comparatively less frequent (see Figure 15).

Figure 15. Response distribution for the sixteen aspiration variables used in the cluster analysis, comparing original and synthetic data.



The same conclusion emerges for the supplementary variables. Socio-demographic characteristics, willingness to work in different geographical areas, and willingness to accept different contractual arrangements display very similar category distributions in the Original and Synthetic data (see Figure 16).

A similar picture emerges from the correlation structure of the sixteen aspiration variables. In both the Original and Synthetic data, the correlations are predominantly positive, with several moderate associations linking conceptually

Figure 16. Distribution of response categories for the supplementary variables in the original and synthetic data.

related dimensions of future work aspirations (see Figure F.1 in Appendix F).

Regarding clustering, according to the elbow rule, a three-cluster partition was retained (see Figure 17) and then applied to both the Original and Synthetic datasets in order to ensure direct comparability.

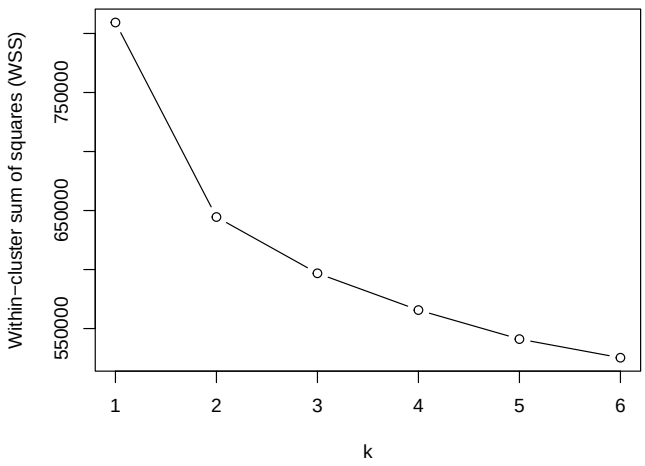
The original-data centroids reveal a highly interpretable three-profile structure (see Table 7). The first cluster is a *high-expectation* profile, with consistently high values across the sixteen items and especially strong emphasis on a broad and demanding set of job qualities, including autonomy, participation, social usefulness, workplace relationships, and the possibility of using acquired competences. The second cluster can be described as a *moderately selective and pragmatic* profile. It assigns relatively greater importance to economic security and professional consolidation than to the more expressive or relational dimensions of work. The third cluster corresponds to a *low-expectation* profile. It is characterised by systematically lower values on the aspiration indicators and appears to represent a group with weaker or less demanding job expectations overall.

Table 7. Standardized cluster centroids for the original and matched synthetic data after centroid matching

	Original data			Synthetic data (matched)		
Item	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
asp1	0.309	0.009	-1.056	0.293	0.059	-1.393
asp2	0.291	0.197	-1.414	0.292	0.191	-1.746
asp3	0.318	-0.004	-1.057	0.313	-0.033	-1.228
asp4	0.320	0.226	-1.575	0.324	0.146	-1.760
asp5	0.367	-0.179	-0.831	0.360	-0.306	-0.692
asp6	0.430	-0.381	-0.590	0.423	-0.500	-0.434
asp7	0.429	-0.290	-0.791	0.399	-0.346	-0.748
asp8	0.502	-0.540	-0.480	0.472	-0.545	-0.518
asp9	0.507	-0.504	-0.574	0.486	-0.585	-0.469
asp10	0.474	-0.299	-0.919	0.477	-0.414	-0.890
asp11	0.497	-0.299	-0.997	0.476	-0.383	-0.969
asp12	0.536	-0.564	-0.539	0.520	-0.596	-0.579
asp13	0.495	-0.404	-0.760	0.466	-0.448	-0.752
asp14	0.508	-0.526	-0.530	0.488	-0.542	-0.591
asp15	0.359	-0.263	-0.619	0.363	-0.318	-0.669
asp16	0.442	-0.220	-0.990	0.435	-0.384	-0.794

Note: synthetic clusters are reordered according to the optimal matching with the original clusters obtained through the Hungarian assignment algorithm.

Figure 17. Within-cluster sum of squares (WSS) for candidate cluster solutions from $k = 1$ to $k = 6$ in the original data.



This interpretation is also confirmed by the v-tests for the sixteen aspiration variables, which show a coherent contrast between the three groups in both Original and Synthetic data (see Figure F.2 in Appendix F). Note that although v-tests are primarily used to describe clusters with respect to a set of supplementary variables, in this case they are also reported for the active variables (aspirations) in order to provide a direct comparison between the Original and Synthetic datasets.

As also shown in Table 7 and Figure F.2 (Appendix F), the synthetic data reproduce the same general profile structure. After centroid matching, the distance between the original and the synthetic centroids is small for all three matched clusters and especially small for the first one (see Table 8).

Table 8. Euclidean distance between matched Original and Synthetic data centroids

Matched cluster	Distance
Cluster 1	0.088
Cluster 2	0.241
Cluster 3	0.427

These values indicate a close correspondence between the original and the synthetic solutions. Figure F.3 in Appendix F reports the centroid differences

for each item. The largest discrepancies are observed for `asp1`, `asp5`, `asp6`, and `asp16`; however, they remain limited in magnitude, confirming that the synthetic data preserve the geometry of the cluster solution obtained from the original data.

A second useful diagnostic concerns the distribution of observations across matched clusters. Table 9 shows that the proportion of observations assigned to each cluster is very similar in the Original and Synthetic datasets.

Table 9. Percentage of observations in each matched cluster in the original and the synthetic data

Cluster	Original data (%)	Synthetic data (%)
Cluster 1	50.9	53.2
Cluster 2	33.9	34.1
Cluster 3	15.2	12.6

Supplementary variables for cluster characterisation. The clusters are strongly associated with several supplementary variables (see Figures F.4, F.5, and F.6 in Appendix F).

The high-expectation cluster is the one most clearly associated with an expansive orientation towards both geographical mobility and employment options, confirming that it represents the most open and broadly oriented group. It is positively characterised by favourable categories of willingness to work in several geographical areas and by greater openness towards a range of contractual arrangements, especially more stable and demanding options such as full-time work and contracts with increasing employment protection. This cluster is also characterised by a higher presence of female than male.

By contrast, the low-expectation cluster is associated with a more constrained configuration. It is negatively characterised by many of the most expansive categories, including willingness to work in different geographical areas and willingness to accept several contractual arrangements, especially full-time work and contracts with increasing employment protection. This indicates that the cluster is more limited not only in the importance attributed to job characteristics, but also in its broader orientation towards employment opportunities and mobility. This cluster is characterised by a higher presence of male than female.

The intermediate pragmatic cluster occupies a position between these two extremes. It is more selective than the high-expectation profile, but less uniformly constrained than the low-expectation one. Its interpretation is therefore consistent with a profile in which work is valued primarily in terms of security, employability, and feasible opportunities, rather than in terms of a broad and highly demanding ideal of future work. This cluster is especially associated with

graduates from disciplinary fields such as engineering and economics-statistics, and it is also characterised by a higher proportion of male.

Note that social class (`class_soc`) and age at graduation (`etalau_c`) proved to be only weakly informative for cluster characterisation, as they showed limited discriminatory power across the three profiles.

The synthetic data reproduce the same broad hierarchy of explanatory variables, as is evident in the graphical comparison of v-tests for the supplementary categorical variables by cluster (see Figures F.4, F.5, and F.6), where the original and synthetic results display highly similar patterns. Taken together, these findings indicate that the synthetic data preserve not only the core geometry of the cluster solution, but also much of the substantive interpretative structure underlying the profiles. Overall, the analysis confirms the ability of the synthetic data to recover meaningful relationships among variables even within an exploratory multivariate framework such as cluster analysis.

5 Concluding remarks

This report has assessed whether synthetic data can be used to get reliable results in empirical research where access to confidential micro-data is restricted by legal, organizational, or privacy constraints. Using relational university administrative records and graduate survey data from four universities in the GRINS network, we compared original and synthetic datasets along two complementary dimensions: first, through global metrics of privacy protection and statistical fidelity; second, through three different empirical applications based on different sample selection criteria and estimators.

The comparison between original and synthetic data in Section 3 provides encouraging results: both statistical fidelity and privacy protection are generally high across all universities and all data tables considered. Marginal distributions, missing value patterns and many bivariate statistics are closely aligned between original and synthetic data in tables (relating to demographic characteristics, enrollment, exams, graduates' profiles, and employment results). However, the two datasets seem to perform differently in the case of sparse categorical variables or where the share of missing values is relatively high. More generally, while the synthetic data preserve the overall multivariate structure well, some specific pairwise correlations are harder to replicate, especially when they involve rare cases/events or thin cells. This confirms that synthetic data are particularly reliable for analyzes that exploit the broad geometry of the data used to bind them, but they should be used with more caution for analyzes that hinge on fine-grained local dependence patterns.

The applications in Section 4 support this interpretation. In the peer-effects application, summary statistics on the selected sub-sample and the derived variables are very similar between the original and the synthetic data in all the three Universities for which the analysis was carried out. In two Universities out of three (i.e., Universities 2 and 4) also the econometric estimates are qualitatively similar: higher peer ability is associated with better educational attainment and lower dropout. The comparison between CSD and FSD highlights that these results hold, especially when the synthetic data are generated from the merged (complete) fully integrated original data set, and hence exploiting all the potential relational features across variables available in different datasets (i.e., administrative and survey data). By contrast, synthetic data perform less well when they are obtained by applying independent generating processes to different data sources (that is, without exploiting the full relational features of the original data), which are subsequently merged through unique identifiers. This is the case of FSD: although they preserve part of the residual identifying variation, they reproduce the joint distribution between peer quality and extensive-margin outcomes less accurately, leading to attenuated peer effects. The results of University 1 are less consistent between the original and synthetic

data, suggesting that these differences may also reflect specific features of the institutional setting (such as degree definition) and the empirical design itself, especially when the target parameter is sensitive to outliers or to definitions of peer-groups that the synthetic data do not fully identify.

The application concerning the prediction of the drop-out risk confirms that synthetic data preserve the main descriptive structure of the selected sample rather well in all the Universities considered. However, predictive performance is not fully equivalent to that obtained with the original data. The classification accuracy at a fixed threshold remains broadly similar, but the AUC is systematically higher in the original data. This suggests that the synthetic data set retains much of the ranking information needed for prediction, but with a smaller dispersion in individual risk. Importantly, the ranking of the most relevant predictors is largely stable across original and synthetic data: first-semester academic performance remains central, and the LASSO model generally selects the same variables with similar coefficients across datasets.

The third application, based on cluster analysis of graduates according to job aspirations (and currently available only for one University), provides the strongest evidence in favor of synthetic data as a reliable tool for empirical research. In this case, the synthetic data reproduce the summary statistics (including the correlation structure) and the main outputs of the cluster analysis with high fidelity. The distances between the matched centroids are small, especially for the first cluster, and the substantive interpretation of the clusters remains stable across the original and synthetic data. The structure of the clustering output is preserved, as well as the distribution of supplementary variables used to characterize the profiles. This suggests that synthetic data are especially promising for exploratory multivariate analyzes aimed at recovering latent structures, typologies, or broad association patterns.

In general, evidence indicates that synthetic data are likely to be a valid substitute for real data under some conditions, such as when the empirical objective does not depend critically on rare events or "thin" cells, or when the target of inference is a broad pattern, such as a predictor ranking or a cluster structure. Comparison between the two types of synthetic data also highlights that they are more reliable when generated by exploiting the full relational structure of the underlying data sources. Under these conditions, the synthetic data used in this study preserve enough of the original statistical structure to support a reliable and meaningful empirical analysis while substantially relaxing access constraints.

This study is subject to some limitations. The empirical evidence is based on a limited number of universities and on a restricted, though heterogeneous, set of applications. In addition, only two universities were able to implement both generation processes for the synthetic data within the project horizon.

Furthermore, while the synthetic data generation methods employed are state-of-the-art, they may not fully replicate the complexity of real-world data in all dimensions. Finally, although representative, the scope of the applications considered does not exhaust the range of possible analytical scenarios in which synthetic data could be deployed. Despite these limitations, the study provides a solid and promising empirical foundation for future research in this area.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abdelhameed, S. A., Moussa, S. M., and Khalifa, M. E. (2018). Privacy-preserving tabular data publishing: A comprehensive evaluation from web to cloud. In *Computers & Security*, volume 72, pages 74–95.

Abowd, J. M., Aref, F., et al. (2020). The 2020 Census disclosure avoidance system. Technical report, U.S. Census Bureau. Published in *Harvard Data Science Review*, 2022.

Abowd, J. M. and Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202.

Abowd, J. M. and Stinson, M. H. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, 95(5):1451–1467.

Abowd, J. M., Stinson, M. H., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau.

Abowd, J. M. and Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in administrative data. *Journal of Business & Economic Statistics*, 23(2):133–152.

Abowd, J. M. and Woodcock, S. (2001). Disclosure limitation in longitudinal linked data. In Doyle, P., Lane, J. I., Theeuwes, J. J., and Zayatz, L. V., editors, *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 215–277. North-Holland.

AINDO (2025). Neural model for relational data: Technical documentation. `https://docs.rdml.aindo.com/latest/guide/neural/intro/`. Accessed: April 2026.

Appenzeller, A., Leitner, M., Filz, M., Houghton, A., and Sedlmeir, J. (2022). Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23):12320.

Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., and Veloso, M. (2020). *Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls*. ACM.

Baak, M., Koopman, R., Snoek, H., and Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Bellomarini, L., Laurenza, E., and Sallinger, E. (2022). AI, synthetic data and regulation. *Proceedings of the International Conference on Information Systems (ICIS)*.

Benedetto, G., Stinson, M., and Abowd, J. (2013). The creation and use of the SIPP synthetic beta. *Journal of Applied Economics*, 28(3):373–397.

Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Lachowicz, M., Rose, B. A., Shaw, T., Stapleton, L. M., Woolley, M. E., and Zheng, Y. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*, 12(4):702–727.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022a). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Borisov, V., Sessler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. (2022b). Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Boudewijn, A., Ferraris, A. F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., and Chauvenet, C. R. (2023). Privacy measurement in tabular synthetic data: State of the art and future research directions. *arXiv preprint arXiv:2311.17453*.

Bowen, C. M. and Liu, F. (2020). Comparative study of differentially private synthetic data algorithms from the NIST 2018 challenge. *Journal of Privacy and Confidentiality*, 10(1):1–21.

Bowen, C. M. and Snoke, J. (2021). Differentially private data synthesis methods. *Annual Review of Statistics and Its Application*, 8:473–495.

Cheng, X., Su, S., Xu, S., and Li, Z. (2017). Towards de-anonymization of public datasets. *IEEE Transactions on Big Data*, 3(2):155–167.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer, New York.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357.

Drechsler, J. and Vilhuber, L. (2014). Synthetic longitudinal business databases for international comparisons. *Statistical Journal of the IAOS*, 30(2):137–146.

El Emam, K. (2020). *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly Media.

El Emam, K., Mosquera, L., and Hoptroff, R. (2020). *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., and Roventini, A. (2019). Validation of agent-based models in economics and finance. *LEM Working Paper Series*.

Figueira, A. and Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, 10(15):2733.

Ganev, G., Oprisanu, B., and De Montjoye, Y.-A. (2024). On the inadequacy of similarity-based privacy metrics: Reconstruction attacks against "truly anonymous synthetic data". *arXiv preprint arXiv:2312.05114*.

Giomi, M., Boenisch, F., Wehmeyer, C., and Tascón, B. (2022). A unified framework for quantifying privacy risk in synthetic data. *Proceedings on Privacy Enhancing Technologies*, 2023(2):312–328.

Giuffrè, M. and Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digital Medicine*, 6(1):186.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(1):108.

Gonzales, A., Guruswamy, G., and Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082.

Gueye, M., Attabi, Y., and Dumas, M. (2023). Row conditional-TGAN for generating synthetic relational databases. In *IEEE ICASSP 2023*.

Gupta, R., Gupta, R., and Gupta, A. (2016). Synthetic data generation. *International Journal of Computer Applications*, 148(7).

Hernandez, M., Epelde, G., Pantaleo, A., Beristain, A., Larrea, M., and Molano, A. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45.

Hittmeir, M., Mayer, R., and Ekelhart, A. (2020). *A Baseline for Attribute Disclosure Risk in Synthetic Data*. ACM.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Hudovernik, V., Jurkovič, M., and Štrumbelj, E. (2024). Benchmarking the fidelity and utility of synthetic relational data. *arXiv preprint arXiv:2410.03411*.

Ji, Z., Lipton, Z. C., and Elkan, C. (2014). Differential privacy and machine learning: A survey and review. *arXiv preprint arXiv:1412.7584*.

Jordon, J., Szpruch, L., Houssiau, F., Sherborne, M., and van der Schaar, M. (2022). Synthetic data – what, why and how? *arXiv preprint arXiv:2205.03257*.

Jordon, J., Yoon, J., and van der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR)*.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, 79(3):362–384.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. In *International Conference on Machine Learning (ICML)*. arXiv preprint arXiv:2209.15421.

Li, J. and Tay, W. P. (2023). Incremental relational generator for tabular data. *arXiv preprint*.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9(2):407–426.

Mami, C. A., Bacciu, D., and Numerosi, L. (2022). Generating relational data with variational graph autoencoders. *arXiv preprint*.

McKenna, R., Sheldon, D., and Miklau, G. (2019). Graphical-model based estimation and inference for differential privacy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4435–4444.

Meehan, C., Chaudhuri, K., and Dasgupta, S. (2020). A non-parametric test to detect data-copying in generative models. *Proceedings of Machine Learning Research (AISTATS)*, 108:3546–3556.

Nowok, B., Raab, G. M., and Dibben, C. (2016). `synthpop`: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74(11):1–26.

O'Keefe, C. M. and Rubin, D. B. (2015). Individual privacy versus public good: Protecting confidentiality in health research. *Statistics in Medicine*, 34(23):3081–3103.

Pang, W., Zhao, Z., Hu, Y., and Luo, S. (2024). ClavaDDPM: Multi-relational data synthesis with cluster-guided diffusion models. *arXiv preprint*.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The Synthetic Data Vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

Plasencia Palacios, M. N., Boudewijn, A., Saccani, S., Ferraris, A. F., Sofronieva, D., D'Acquisto, G., Brozzetti, F., Panfilo, D., and Bortolussi, L. (2025). Empirical evaluation of structured synthetic data privacy metrics: Novel experimental framework. *arXiv preprint arXiv:2512.16284*.

Raab, G. M., Nowok, B., and Dibben, C. (2020). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 10(1):1–27.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188.

Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131(2):365–377.

Rosenblatt, L. and Howe, B. (2022). Epistemic parity: Reproducibility as an evaluation metric for differential privacy. *ACM SIGMOD Record*, 51(3):34–41.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Scassola, D., Saccani, S., and Bortolussi, L. (2025). Graph-conditional flow matching for relational data generation. *arXiv preprint arXiv:2505.15668*.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Snoke, J., Raab, G., Nowok, B., Dibben, C., and Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):663–688.

Solatorio, A. V. and Dupriez, O. (2023). REaLTabFormer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*.

Stadler, T., Oprisanu, B., and Troncoso, C. (2020). Synthetic data – anonymisation groundhog day. In *Proceedings of the 29th USENIX Security Symposium*, pages 347–363.

Stadler, T., Oprisanu, B., and Troncoso, C. (2022). Synthetic data — anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 217–234. USENIX Association.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 7333–7343.

Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R., and Susskind, J. (2021). An attention free transformer.

Zhang, G., He, Y., Oganian, A., and Cai, B. (2025). Creating synthetic data for complex surveys using the Research and Development Survey: A comparison study. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, 2(212):1–10.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017). PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems*, 42(4):1–41.

Appendix for the paper
Synthetic data for academic research: (more) opportunities & (than) challenges by

V. Atella[1] P. Bello [4]C. Davino[2] R. Fabbricatore[2] M. Fort[3] G. Lima [3] C. Marconi [4] F. Origo [4] E. Pisanelli [4] V. Rattini [3] D.Vuri[1]

Appendix

Table of Contents

A Method and procedural challenges **3**

B Description of the reference population **3**

B.1 University 1 3

B.2 University 2 4

B.3 University 3 5

B.4 University 4 5

C Additional Tables and Figures for Section 3 **7**

D Additional Tables and Figures for Section 4.1 **9**

E Additional Tables and Figures for Section 4.2 **14**

F Additional Tables and Figures for Section 4.3 **17**

A Method and procedural challenges

To generate synthetic data with AINDO—under the preferred approach of SDK-based synthesis executed within each university's secure environment—the process relies on a lightweight, code-driven technology stack rather than a full platform deployment. In practice, each university provisions a dedicated compute server (on-premise or cloud), hosts the source data locally or via a connected database, and provides AINDO with secure VPN access. Once AINDO personnel have been formally appointed as data processors, they run the synthesis through the AINDO SDK, which consists of Python libraries and runtime components that execute data preparation, model training, and synthetic data generation directly on the university infrastructure. This setup is operationally advantageous because it enables rapid troubleshooting and iterative refinement.

From a technical standpoint, the server should offer adequate compute and memory resources for model training, including an Intel CPU with 16+ cores and support for AVX-512 and AVX2 instruction sets (with at least 8 cores available), 64 GB of RAM, and roughly 200 GB of storage (subject to variation depending on data volume and intermediate artefacts). The recommended operating system is Ubuntu Server 22.04 x86-64, although Windows can also be used. The software environment requires Python 3.10, with support for virtual environments to ensure isolated and reproducible dependencies. The machine must have outbound connectivity to PyPI to install required Python packages, and it must allow installation of the AINDO SDK via a distributable .whl package. In addition, selected network ports should be reachable to enable web-based interfaces used for monitoring and interactive workflows. Under this configuration, AINDO can install and operate the SDK without root privileges, reducing operational burden on university IT teams while maintaining standard security controls. Finally, given the scale and heterogeneity of the raw data sources, a preliminary extraction and curation phase is required to construct a synthesis-ready dataset that complies with project guidelines, including the selection of available variables, the treatment of structural missingness, and the harmonisation of linkage keys across administrative and AlmaLaurea domains.

B Description of the reference population

B.1 University 1

The reference population for University 1 includes 70,634 students enrolled in first-cycle (Bachelor's), second-cycle (Master's), or single-cycle degree programs under DM 270 or DM 509 from academic year 2012/2013 to 2024/2025. The data combine administrative records with survey data. The first include: (i) the student registry dataset, which contains information on socio-demographic

characteristics upon enrollment, including prior education and qualifications; (ii) the academic career registry, including, for each academic year, the specific degree of enrollment, income used to compute tuition fees, scholarships, and academic progression (that is, academic status of the student); (iii) the exams registry, which provides detailed information on all the exams taken by each student (number of credits, number of trials and grade) and the final degree grade (including the date of degree and socio-demographic characteristics (gender and tenure) of the supervisor). Administrative data consist of 218,656 student-degree-year observations and 974,227 exam-level observations. Starting from a.y. 2018-2019, information on the score of the entry test (TOLC) is also available for 20,878 students. Survey data come from Almalaurea and include: (i) the survey on graduates (30,808 observations); (ii) the survey on labour market outcomes 1, 3 and 5 years since graduation (25,924 , 5,786 and 3,269 observations, respectively). The different sources can be merged using the anonymised student ID, the degree and year of enrollment (or graduation year to merge Almalaurea with administrative data).

B.2 University 2

The reference population for University 2 comprises 275,605 graduates who enrolled in first-cycle (Bachelor's), second-cycle (Master's), or single-cycle degree programmes under DM 270 or DM 509 between academic years 2012/2013 and 2024/2025. Individuals who obtained more than one degree within the reference window are represented by multiple records, since each degree is treated as a distinct statistical observation. The original data integrate two complementary sources: a relational database of university administrative records and a set of AlmaLaurea survey datasets – graduate profile and employment conditions measured 1, 3, and 5 years after graduation. These sources are designed to be linkable through anonymised individual identifiers, namely an anonymised tax code and an anonymised career ID. The resulting information spans three main domains. First, administrative data reconstruct students' educational pathways – such as enrolments, background characteristics and socio-economic status – organised at the student-by-academic-year level and indexed by the anonymised tax code, anonymised career ID, and enrolment year. Second, exam records provide course-level information at the exam-by-student level. Third, AlmaLaurea modules capture graduates' profiles at the time of degree completion and subsequent labour-market outcomes, which are observed only for graduates and follow different coverage rules at the 1-year versus the 3- and 5-year follow-ups. In terms of scale, the available datasets include 275,605 students in the "anagrafica" archive, which contains the core student registry and basic administrative identifiers; 332,356 records in "carriera", which document enrollment histories, degree-program information, and academic progression over time; 4,999,617

records in the "esami" dataset, which contains exam-level administrative records for exams taken by students; the same course exam may appear multiple times for a student when it is taken more than once, and the dataset includes information on whether each attempt was passed or failed; 151,890 observations in the "AlmaLaurea profile" survey, which reports graduates' characteristics and self-declared profiles at graduation; 131,833 observations in the "AlmaLaurea employment-status survey at 1 year after graduation" survey; 50,967 observations at 3 years; and 33,388 observations at 5 years, providing medium- and longer-term measures of labour-market outcomes.[1]

B.3 University 3

The reference population for University 3 comprises 111,839 graduates (2013-2025) who first enrolled in first-cycle (Bachelor's), second-cycle (Master's), and single-cycle degree programs under DM 270 between 2012 and 2022. Students who earned more than one degree within the reference period are represented by multiple records, as each degree constitutes a separate observation. The data cover three main domains: (i) students' educational pathways based on university administrative records, (ii) graduates' profiles from AlmaLaurea, and (iii) labor market outcomes at 1, 3, and 5 years after graduation from AlmaLaurea.

Regarding the administrative datasets on student careers, several variables are not available in the source data used to generate the synthetic data. Specifically, the data set for the entrance tests is not available. Moreover, missing information concerns detailed enrollment history variables (e.g., recognition of prior credits and part-time enrollment status), scholarship and economic indicators (e.g., ISEE and tuition fee exemptions), information on prior education and inter-university mobility (details on previous qualifications and student mobility from other universities), and (for exams) attempt data. Note that data on entrance tests, scholarships, and exam attempt records are generally not available in the university administrative archives.

B.4 University 4

The reference population for University 4 comprises students first enrolled in first-cycle (Bachelor's), second-cycle (Master's), and single-cycle degree programs under DM 270 or DM 509 from academic year 2012/2013 to 2024/2025. The data are drawn from multiple sources: (i) the student registry dataset, which contains information on enrollment, degree type, prior education, previous qualifications, and mobility; (ii) the scholarship dataset, which includes information on financial aid, fee exemptions, and economic indicators; (iii) the

[1] The data do not have information on students who started their career at University 2 but then transfer to another university.

degree dataset for graduates, which provides information on degree grades and date of degree as well as thesis supervisor characteristics (age and gender). These sources are designed to be linkable by year of enrollment through an anonymised student ID. By assembling these three dataset we obtain a sample of 331,011 year of enrollment-student records. In addition, we get information from (iv) the exam dataset, which records details of attempted and passed exams, including exact dates, credit values, and scores. It includes 1,681,124 student-exam records. Because entrance exam data are not centrally administered by the University, entrance test scores are available only for 35,449 students enrolled in Economics and Science programs (and for the latter only for selected years). By assembling these datasets, we reconstruct each student's complete academic trajectory from first enrollment through graduation (or final year of enrollment), including information on entrance test results when available. The resulting dataset is structured as a panel of 1,681,124 records, with each observation representing a student-exam record. Student administrative records are enriched with data from AlmaLaurea, which captures graduate characteristics and self-reported profiles at the time of degree completion for 55,760 individuals. AlmaLaurea also provides labor market outcomes at 1, 3, and 5 years following graduation. Employment surveys at the 3- and 5-year marks are administered exclusively to graduates of second-cycle (Master's) and single-cycle degree programs. Labor market data are available for 42,006, 14,232, and 9,835 graduates at 1, 3, and 5 years post-graduation, respectively. Almalaurea data are linked to University data through an anonymised student ID and a "degree course" identifier.

C Additional Tables and Figures for Section 3

Table C.1. PhiK difference values by data table and University: Number of considered variables (#var), Minimum difference value (Min), Maximum difference value (Max), and Most Critical Variables

Data table	University	#var	Min	Max	Most Critical Variables
			Registry data from University archives		
Personal details	University 1	9	0.00	0.24	cittadinanza × gender
	University 2	9	0.00	0.12	comune_nascita × gender
					anno_nascita x cittadinanza
	University 3	8	0.00	0.53	regione_nascita × anno_nascita
	University 4	9	0.00	0.23	provincia_nascita × anno_nascita
College career	University 1	15	0.00	0.38	comune_domicilio × regione_residenza
	University 2	19	0.00	0.32	titolo_straniero_flg × tipo_corso
	University 3	14	0.00	0.71	comune_residenza × aa_nasc_docente_laurea1
	University 4	20	0.00	0.98	comune_residenza × aa_nasc_docente_laurea1
Enrollment	University 1	17	0.00	0.25	curriculum × anni_fuori_corso
	University 2	19	0.00	0.23	trasferito_a × curriculum_id
	University 3	7	0.00	0.17	trasferito_a × idcorsodistudio
	University 4	20	0.00	0.94	AbbrevCarriera × cds_normativa
					AbbrevCarriera × part_time_flg
					cds_normativa × part_time_flg
					trasferito_da_altro_ateneo × cds_normativa
					trasferito_da_altro_ateneo × part_time_flg
					passaggio_da_altro_corso × cds_normativa
					passaggio_da_altro_corso × part_time_flg
					passaggio_da_altro_corso × trasferito_da_altro_ateneo
Exams	University 1	16	0.00	0.24	ric_id × num_trial
	University 2	18	0.00	0.19	flag_respinto_ritirato × num_trial
	University 3	15	0.00	0.28	facolta × id_career
	University 4	20	0.00	1	data_sup × cfu
Entry test	University 1	11	0.00	0.43	TolcPunti × comprensione_last
	University 2	15	0.00	0.12	aa_last × cdldiassegnazione
	University 3	NA	NA	NA	NA
	University 4	15	0.00	0.55	postiaconcorso × inglese_last
			Graduates' Profile, at graduation		
Graduates' Profile	University 1	17	0.00	0.30	facoa × durata
	University 2	18	0.00	0.25	etam × diplnome2
	University 3	16	0.00	0.19	class × codicione
	University 4	20	0.00	0.29	diplnome3 × corsa
					diplnome3 × univ
			Graduates' Employment Status, post graduation		
in 1 Year	University 1	20	0.35	0.30	U9AL_univ_1a × tipo_indagine_1a
	University 2	20	0.00	0.32	U9AL_univ_1a × tipo_indagine_1a
	University 3	20	0.00	0.35	U9AL_univ_1a × tipo_indagine_1a
	University 4	20	0.00	0.31	U9AL_univ_1a × tipo_indagine_1a
in 3 Years	University 1	20	0.35	0.60	Q2f_3a × Q1F_3a
	University 2	20	0.00	0.60	Q1H_3a × Q1A_3a
	University 3	20	0.00	0.41	kq1hb_3a × interv_3a
					interv_3a × almenuna_noq1ha_3a

Table C.1 continued.

Data table	University	#var	Min	Max	Most Critical Variables
					interv_3a × incorimp_noq1ha_3a
	University 4	20	0.00	0.43	kq1hb_3a × interv_3a
					interv_3a × almenuna_noq1ha_3a
					interv_3a × incorimp_noq1ha_3a
in 5 Years	University 1	20	0.35	0.84	impegn_noq1ha_5a × interv_5a
	University 2	20	0.00	0.34	klav_cerc_c2_5a × interv_5a
	University 3	20	0.00	0.25	Q1HB_5a × impegn_noq1ha_5a
					Q1HB_5a × Q1A_5a
					Q1HB_5a × Q1C_5a
					impegn_noq1ha_5a × Q1B_5a
	University 4	17	0.00	0.43	interv_5a × noncerca_maformFL_5a
					noncerca_maformFL_5a × incorimp_noq1ha_5a

Note: `cittadinanza` = citizenship; `anno_nascita` = year of birth; `comune_nascita` = municipality of birth; `regione_nascita` = region of birth; `provincia_nascita` = province of birth; `gender` = gender; `comune_domicilio` = municipality of temporary residence; `regione_residenza` = region of permanent residence; `comune_residenza` = municipality of permanent residence; `titolo_straniero_flg` = foreign qualification flag; `tipo_corso` = type of degree programme; `aa_nasc_docente_laurea1` = year of birth of the thesis advisor; `curriculum` = study track; `anni_fuori_corso` = years enrolled beyond the standard programme duration; `trasferito_a` = transferred to; `idcorsodistudio` = degree programme identifier; `AbbrevCarriera` = abbreviated academic career; `cds_normativa` = degree programme regulations; `part_time_flg` = part-time flag; `trasferito_da_altro_ateneo` = transferred from another university; `passaggio_da_altro_corso` = transferred from another degree programme; `ric_id` = exam recognition identifier; `num_trial` = number of exam attempts; `flag_respinto_ritirato` = rejected or withdrawn flag; `facolta` = faculty/department; `id_career` = career identifier; `data_sup` = exam pass date; `cfu` = university credits; `TolcPunti` = score in the standardized entrance test; `comprensione_last` = score in the comprehension section of the standardized entrance test; `aa_last` = entrance test academic year; `cdldiassegnazione` = assigned degree programme; `postiaconcorso` = number of places available in the programme; `inglese_last` = score in the english section of the standardized entrance test; `facoa` = faculty; `durata` = duration of studies (in years); `etam` = age at graduation (years); `diplnome2` = Type of diploma (aggregated); `class` = degree programme class (field of study); `codicione` = degree programme code (post-reform); `diplnome3` = type of diploma (detail); `corsa` = pre-reform degree programme (AL code); `univ` = university; `U9AL_univ_1a` = Master's vs bachelor's university; `tipo_indagine_1a` = type of survey (CAWI, CATI); `Q2f_3a` = job offer after the internship; `Q1F_3a` = post-graduation internship in a company; `Q1H_3a` = other activities under a research fellowship; `Q1A_3a` = unpaid voluntary work; `kq1hb_3a` = aggregation of q1hb (other activities with scholarship/work grant); `interv_3a` = indicator for being interviewed in the survey; `almenuna_noq1ha_3a` = participation in training activities (including unpaid voluntary work, excluding activities under a research fellowship); `incorimp_noq1ha_3a` = currently undertaking a significant post-graduation training activity (including unpaid voluntary work, excluding activities under a research fellowship); `impegn_noq1ha_5a` = seeking employment while engaged in training (including unpaid voluntary work, excluding activities under a research fellowship); `interv_5a` = indicator for being interviewed in the survey; `klav_cerc_c2_5a` = aggregation of lav_cercFL and c2 (job search/employment status); `Q1HB_5a` = other activities with scholarship/work grant; `Q1A_5a` = unpaid voluntary work; `Q1C_5a` = PhD participation; `Q1B_5a` = post-graduation internship/traineeship; `noncerca_maformFL_5a` = not seeking employment and engaged in training; `incorimp_noq1ha_5a` = currently undertaking a significant post-graduation training activity (including unpaid voluntary work, excluding activities under a research fellowship).

D Additional Tables and Figures for Section 4.1

Table D.2. Summary Statistics Across Different Datasets - University 1

	Original Data			CSD		
	Obs.	Mean	S.D.	Obs.	Mean	S.D.
Outcomes:						
Graduate (%)	38,397	54.4	49.80	38,742	54.3	49.82
Graduate on Time (%)	38,397	46.6	49.89	38,742	46.3	49.87
Drop Out (%)	38,397	26.8	44.29	38,742	26.9	44.33
Exam Grades (Credit Weighted)	29,754	24.8	2.40	30,283	24.5	2.53
Ability:						
Own Ability (HS GPA, 60-100)	38,397	75.4	10.77	38,742	74.9	10.53
Peer Ability (60-100)	38,397	75.4	2.64	38,742	74.9	2.25
Covariates:						
Female (%)	38,397	63.0	48.27	38,742	61.1	48.75
Get Scholarship (%)	38,397	2.7	16.25	38,742	2.3	15.12
Migrant (%)	38,397	6.3	24.25	38,742	6.5	24.68
Previous College Degree (%)	38,397	0.0	0.00	38,742	0.0	1.34
Cohort Size	38,397	548.6	351.51	38,742	526.8	316.81
Females in Cohort (%)	38,397	62.8	24.12	38,742	61.0	23.48
Get Scholarship in Cohort (%)	38,397	2.8	1.82	38,742	2.4	1.42
Migrants in Cohort (%)	38,397	7.8	3.24	38,742	7.7	2.90
Previous College Degree in Cohort (%)	38,397	0.0	0.04	38,742	0.0	0.15

Notes. The table reports summary statistics for the main variables used in the analysis, across the different dataset versions.

Table D.3. Summary Statistics Across Different Datasets - University 2

	Original Data			CSD			FSD		
	Obs.	Mean	S.D.	Obs.	Mean	S.D.	Obs.	Mean	S.D.
Outcomes:									
Graduate (%)	130,279	66.4	47.23	128,033	66.3	47.26	130,559	55.2	49.73
Graduate on Time (%)	130,279	50.1	50.00	128,033	50.3	50.00	130,559	42.2	49.40
Drop Out (%)	130,279	17.8	38.23	128,033	17.9	38.35	130,559	24.8	43.18
Exam Grades (Credit Weighted)	114,604	25.9	2.43	113,259	26.2	2.53	106,658	25.5	2.67
Ability:									
Own Ability (HS GPA, 60-100)	130,279	81.0	11.97	128,033	81.9	12.16	130,559	79.7	12.20
Peer Ability (60-100)	130,279	81.0	4.37	128,033	81.9	4.15	130,559	79.7	4.19
Covariates:									
Female (%)	130,279	55.1	49.74	128,033	54.9	49.76	130,559	52.9	49.92
Get Scholarship (%)	130,279	13.1	33.75	128,033	11.0	31.32	130,559	12.5	33.10
Migrant (%)	130,279	5.0	21.71	128,033	4.9	21.67	130,559	6.3	24.29
Previous College Degree (%)	130,279	1.1	10.27	128,033	1.4	11.91	130,559	0.7	8.44
Cohort Size	130,279	306.0	261.48	128,033	292.5	233.49	130,559	307.1	257.70
Females in Cohort (%)	130,279	55.1	21.85	128,033	54.7	20.52	130,559	52.9	21.92
Get Scholarship in Cohort (%)	130,279	13.5	4.40	128,033	11.5	3.55	130,559	12.7	3.83
Migrants in Cohort (%)	130,279	8.0	4.52	128,033	8.0	4.18	130,559	9.7	5.19
Previous College Degree in Cohort (%)	130,279	1.4	2.21	128,033	1.9	2.30	130,559	1.1	1.91

Notes. The table reports summary statistics for the main variables used in the analysis, across the different dataset versions.

Table D.4. Summary Statistics Across Different Datasets - University 4

	Original Data			CSD			FSD		
	Obs.	Mean	S.D.	Obs.	Mean	S.D.	Obs.	Mean	S.D.
Outcomes:									
Graduate (%)	65,324	50.6	50.00	63,951	50.2	50.00	62,091	50.0	50.00
Graduate on Time (%)	65,324	43.5	49.57	63,951	43.4	49.56	62,091	43.0	49.50
Drop Out (%)	65,324	29.5	45.62	63,951	30.0	45.81	62,091	29.7	45.70
Exam Grades (Credit Weighted)	51,520	25.3	2.68	51,012	25.2	2.82	49,063	25.3	2.83
Ability:									
Own Ability (HS GPA, 60-100)	65,324	78.3	12.19	63,951	78.8	12.25	62,091	78.2	11.78
Peer Ability (60-100)	65,324	78.3	4.08	63,951	78.8	4.08	62,091	78.2	3.60
Covariates:									
Female (%)	65,324	50.6	50.00	63,951	50.2	50.00	62,091	52.9	49.92
Get Scholarship (%)	65,324	4.9	21.60	63,951	5.1	22.04	62,091	5.3	22.35
Migrant (%)	65,324	7.4	26.16	63,951	6.4	24.41	62,091	7.1	25.64
Previous College Degree (%)	65,324	2.3	15.00	63,951	2.8	16.40	62,091	3.2	17.50
Cohort Size	65,324	338.0	310.26	63,951	345.7	321.80	62,091	337.5	303.15
Females in Cohort (%)	65,324	50.7	22.54	63,951	50.2	20.37	62,091	52.8	20.90
Get Scholarship in Cohort (%)	65,324	4.9	4.31	63,951	5.1	3.62	62,091	5.3	3.92
Migrants in Cohort (%)	65,324	7.4	9.90	63,951	6.4	7.03	62,091	7.1	8.03
Previous College Degree in Cohort (%)	65,324	2.3	3.95	63,951	2.8	4.21	62,091	3.2	4.77

Notes. The table reports summary statistics for the main variables used in the analysis, across the different dataset versions.

Table D.5. Peer Effects Analysis: Results Across Different Datasets - University 1

	(1) Graduate (%)	(2) Graduate on Time (%)	(3) Drop-Out (%)	(4) Exam Grades (Std.)
Panel A: Original Data				
Peer Ability (Std.)	-2.23	-4.96	0.39	-0.02
	(0.893)	(1.175)	(0.595)	(0.025)
	[-4.00,-0.46]	[-7.28,-2.63]	[-0.79,1.57]	[-0.07,0.03]
Observations	38,397	38,397	38,397	29,754
Mean Dependent	54.45	46.60	26.79	0.00
Mean Independent	75.4	75.4	75.4	75.4
Panel B: CSD				
Peer Ability (Std.)	2.41	1.21	-2.90	0.04
	(1.255)	(1.290)	(0.992)	(0.018)
	[-0.07,4.89]	[-1.35,3.76]	[-4.86,-0.94]	[0.01,0.08]
Observations	38,742	38,742	38,742	30,283
Mean Dependent	54.30	46.33	26.87	0.00
Mean Independent	74.9	74.9	74.9	74.9
Individual Covariates	Yes	Yes	Yes	Yes
Cohort Covariates	No	No	No	No
Year FE	Yes	Yes	Yes	Yes
Degree FE	Yes	Yes	Yes	Yes

Notes. The table reports the β coefficient in Equation 7, for different educational attainment outcomes across columns. The independent variable of interest is the standardized leave-out mean of peer ability in the cohort, measured as described in Equation 6. Panel A reports the coefficients for the model estimated in the original administrative data. Panel B reports the same coefficients in the Centralized synthetic data (CSD). Standard errors, clustered at the cohort-level, are presented in parentheses, with the corresponding 95% confidence intervals reported between brackets below.

Table D.6. Peer Effects Analysis: Results Across Different Datasets - University 2

	(1) Graduate (%)	(2) Graduate on Time (%)	(3) Drop-Out (%)	(4) Exam Grades (Std.)
Panel A: Original Data				
Peer Ability (Std.)	3.11	2.99	-2.19	0.04
	(0.666)	(0.689)	(0.516)	(0.017)
	[1.81,4.42]	[1.64,4.35]	[-3.20,-1.17]	[0.01,0.08]
Observations	130,279	130,279	130,279	114,604
Mean Dependent	66.42	50.11	17.78	0.00
Mean Independent	81.0	81.0	81.0	81.0
Panel B: CSD				
Peer Ability (Std.)	3.31	3.86	-1.98	0.06
	(0.561)	(0.584)	(0.380)	(0.011)
	[2.21,4.41]	[2.71,5.01]	[-2.73,-1.23]	[0.04,0.08]
Observations	128,033	128,033	128,033	113,259
Mean Dependent	66.32	50.26	17.91	-0.00
Mean Independent	81.9	81.9	81.9	81.9
Panel C: FSD				
Peer Ability (Std.)	2.38	2.21	-1.01	0.05
	(0.560)	(0.504)	(0.459)	(0.013)
	[1.28,3.48]	[1.22,3.20]	[-1.91,-0.11]	[0.02,0.07]
Observations	130,559	130,559	130,559	106,658
Mean Dependent	55.21	42.24	24.79	-0.00
Mean Independent	79.7	79.7	79.7	79.7
Individual Covariates	Yes	Yes	Yes	Yes
Cohort Covariates	No	No	No	No
Year FE	Yes	Yes	Yes	Yes
Degree FE	Yes	Yes	Yes	Yes

Notes. The table reports the β coefficient in Equation 7, for different educational attainment outcomes, presented across columns. The independent variable of interest is the standardized leave-out mean of peer ability in the cohort, measured as described in Equation 6. Panel A reports the coefficients for the model estimated in the original administrative data. Panel B reports the same coefficients in the first version of the synthetically generated data (CSD). Panel C reports the same coefficients in the second version of the synthetically generated data (FSD). Standard errors, clustered at the cohort-level, are presented in parentheses, with the corresponding 95% confidence intervals reported between brackets below.

Table D.7. Peer Effects Analysis: Results Across Different Datasets - University 4

	(1) Graduate (%)	(2) Graduate on Time (%)	(3) Drop-Out (%)	(4) Exam Grades (Std.)
Panel A: Original Data				
Peer Ability (Std.)	2.22	2.45	-1.89	0.03
	(0.966)	(1.002)	(0.841)	(0.019)
	[0.32,4.12]	[0.48,4.42]	[-3.54,-0.24]	[-0.00,0.07]
Observations	65,324	65,324	65,324	51,520
Mean Dependent	50.56	43.47	29.54	-0.00
Mean Independent	78.3	78.3	78.3	78.3
Panel B: CSD				
Peer Ability (Std.)	0.54	1.59	-0.17	0.02
	(1.008)	(1.012)	(0.752)	(0.016)
	[-1.44,2.52]	[-0.39,3.58]	[-1.64,1.31]	[-0.01,0.05]
Observations	63,951	63,951	63,951	51,012
Mean Dependent	50.19	43.41	29.96	-0.00
Mean Independent	78.8	78.8	78.8	78.8
Panel C: FSD				
Peer Ability (Std.)	2.09	3.17	-2.13	0.05
	(1.000)	(1.034)	(0.781)	(0.017)
	[0.13,4.06]	[1.14,5.20]	[-3.66,-0.59]	[0.02,0.09]
Observations	62,091	62,091	62,091	49,063
Mean Dependent	50.05	42.97	29.71	-0.00
Mean Independent	78.2	78.2	78.2	78.2
Individual Covariates	Yes	Yes	Yes	Yes
Cohort Covariates	No	No	No	No
Year FE	Yes	Yes	Yes	Yes
Degree FE	Yes	Yes	Yes	Yes

Notes. The table reports the β coefficient in Equation 7, for different educational attainment outcomes, presented across columns. The independent variable of interest is the standardized leave-out mean of peer ability in the cohort, measured as described in Equation 6. Panel A reports the coefficients for the model estimated in the original administrative data. Panel B reports the same coefficients in the first version of the synthetically generated data (CSD). Panel C reports the same coefficients in the second version of the synthetically generated data (FSD). Standard errors, clustered at the cohort-level, are presented in parentheses, with the corresponding 95% confidence intervals reported between brackets below.

E Additional Tables and Figures for Section 4.2

Table E.8. Descriptive Statistics: Original vs Synthetic Data - University 1

	Original Data			CSD		
	Obs.	Mean	S.D.	Obs.	Mean	S.D.
Outcomes:						
Broad Dropout (%)	52,972	34.3313	47.4820	53,917	34.3602	47.4915
Dropout (%)	52,972	23.1537	42.1819	53,917	23.7402	42.5495
Enrolled w Low Credits (%)	52,972	30.8975	46.2075	53,917	34.4882	47.5334
Covariates:						
Female (%)	52,972	62.2650	48.4728	53,917	60.7211	48.8375
Scholarship (%)	52,972	2.9563	16.9379	53,917	2.5966	15.9035
Migrant (%)	52,972	7.1944	25.8397	53,917	7.1276	25.7288
International (%)	52,972	0.5078	7.1081	53,917	0.5546	7.4262
Non-resident (%)	52,972	42.7094	49.4661	53,917	44.3496	49.6802
High school grade	52,072	76.4912	11.0862	53,013	76.0791	11.2868
Liceo (%)	52,972	32.6210	46.8830	53,917	36.1908	48.0557
Total Credits first semester	52,972	10.8828	9.8261	53,917	8.5816	8.6882
Mean Score first semester	36,031	24.8503	3.0050	34,323	24.9908	3.1918

Table E.9. Descriptive Statistics: Original vs Synthetic Data - University 2

	Original Data			CSD			FSD		
	Obs.	Mean	S.D.	Obs.	Mean	S.D.	Obs.	Mean	S.D.
Outcomes:									
Broad Dropout (%)	213,804	22.88	42.009	213,817	21.31	40.950	248,980	30.37	45.987
Dropout (%)	213,804	14.27	34.973	213,817	14.37	35.077	248,980	18.47	38.803
Enrolled with Low Credits (%)	213,804	22.39	41.689	213,817	23.60	42.462	248,980	34.46	47.525
Covariates:									
Female (%)	213,804	56.41	49.587	213,817	55.71	49.673	248,980	54.61	49.787
Scholarship (%)	213,804	12.07	32.577	213,817	10.62	30.808	248,980	10.79	31.029
Migrant (%)	213,804	8.28	27.555	213,817	8.14	27.342	248,980	9.50	29.326
International (%)	213,804	3.20	17.594	213,817	3.41	18.160	248,980	3.69	18.862
Non-resident (%)	213,804	73.60	44.080	213,817	76.11	42.641	248,980	76.01	42.704
High school grade	203,571	82.38	12.222	203,460	83.17	14.343	231,302	81.40	14.254
Liceo (%)	213,804	58.94	49.195	213,817	62.51	48.411	248,980	56.76	49.542
Total Credits first semester	213,804	14.28	10.527	213,817	11.63	9.394	248,980	9.44	8.783
Mean Score first semester	166,658	25.86	3.047	161,392	26.01	3.199	165,574	25.65	3.330

Table E.10. Descriptive Statistics: Original vs Synthetic Data - University 4

	Original Data			CSD			FSD		
	Obs.	Mean	S.D.	Obs.	Mean	S.D.	Obs.	Mean	S.D.
Outcomes:									
Broad Dropout (%)	96,773	34.3143	47.4761	95,429	34.0557	47.3899	92,196	33.7151	47.2740
Dropout (%)	96,773	24.4572	42.9836	95,429	24.9484	43.2717	92,196	24.0271	42.7250
Enrolled w Low Credits (%)	96,773	29.4318	45.5738	95,429	32.0469	46.6659	92,196	32.8474	46.9661
Covariates:									
Female (%)	96,773	52.6707	49.9289	95,429	52.4170	49.9418	92,196	54.2507	49.8193
Scholarship (%)	96,773	5.9438	23.6444	95,429	6.4372	24.5417	92,196	6.1760	24.0720
Migrant (%)	96,773	7.6106	26.5169	95,429	6.7233	25.0427	92,196	6.7357	25.0640
International (%)	96,773	1.7753	13.2053	95,429	1.9543	13.8425	92,196	2.1053	14.3562
Non-resident (%)	96,773	30.7162	46.1320	95,429	32.8653	46.9726	92,196	32.5632	46.8613
High school grade (%)	96,744	79.6494	12.4367	95,233	80.0767	12.4229	92,029	79.4053	11.9716
Liceo (%)	96,773	60.7039	48.8411	95,429	58.6164	49.2522	92,196	63.0548	48.2659
Total Credits first semester	96,773	10.7902	10.3401	95,429	9.2059	9.3100	92,196	8.2290	8.8160
Mean Score first semester	61,341	25.0066	3.2200	57,519	24.8800	3.3925	51,951	25.0567	3.4148

F Additional Tables and Figures for Section 4.3

Figure F.1. Correlation matrices of the sixteen aspiration variables in the original data (left panel) and the synthetic data (right panel).

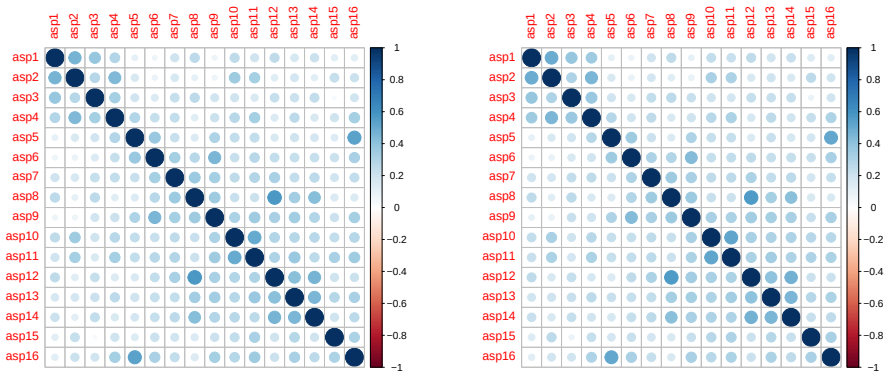


Figure F.2. v-tests for the sixteen aspiration variables in Cluster 1 (top panel), Cluster 2 (middle panel), and Cluster 3 (bottom panel): original versus synthetic data.

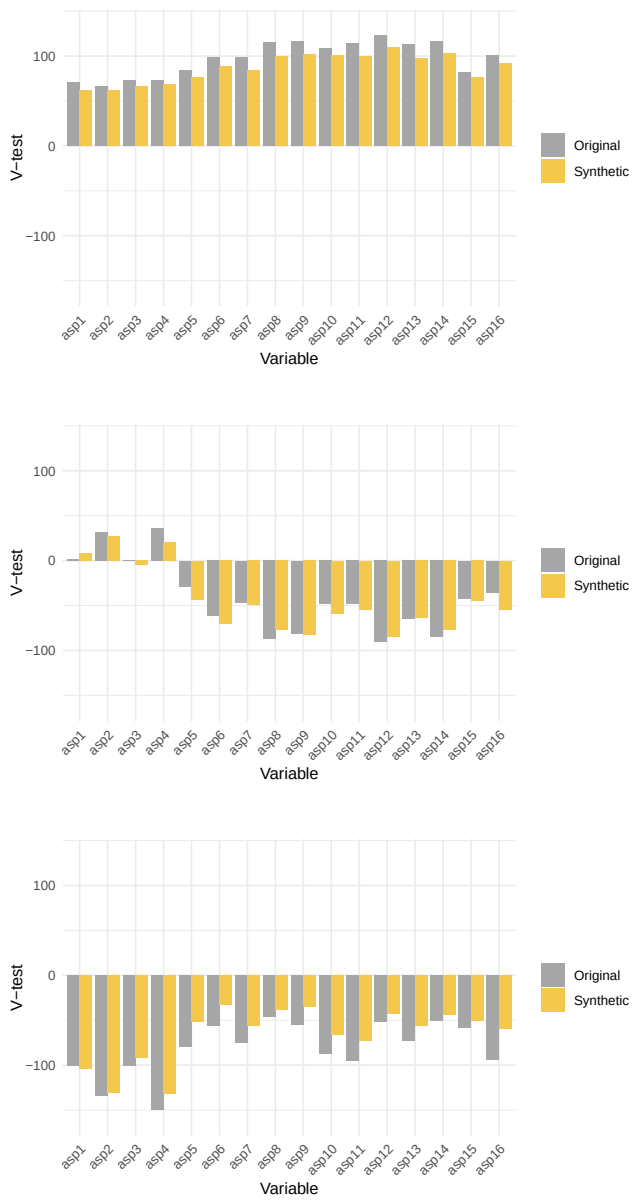


Figure F.3. Differences between matched original and synthetic cluster centroids for each aspiration item.

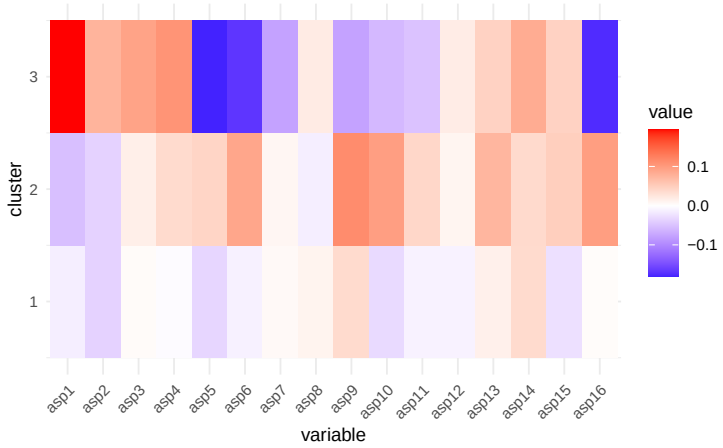


Figure F.4. v-TESTS FOR SOCIO-DEMOGRAPHIC AND BACKGROUND SUPPLEMENTARY VARIABLES IN CLUSTER 1 (TOP PANEL), CLUSTER 2 (MIDDLE PANEL), AND CLUSTER 3 (BOTTOM PANEL): ORIGINAL VERSUS SYNTHETIC DATA.

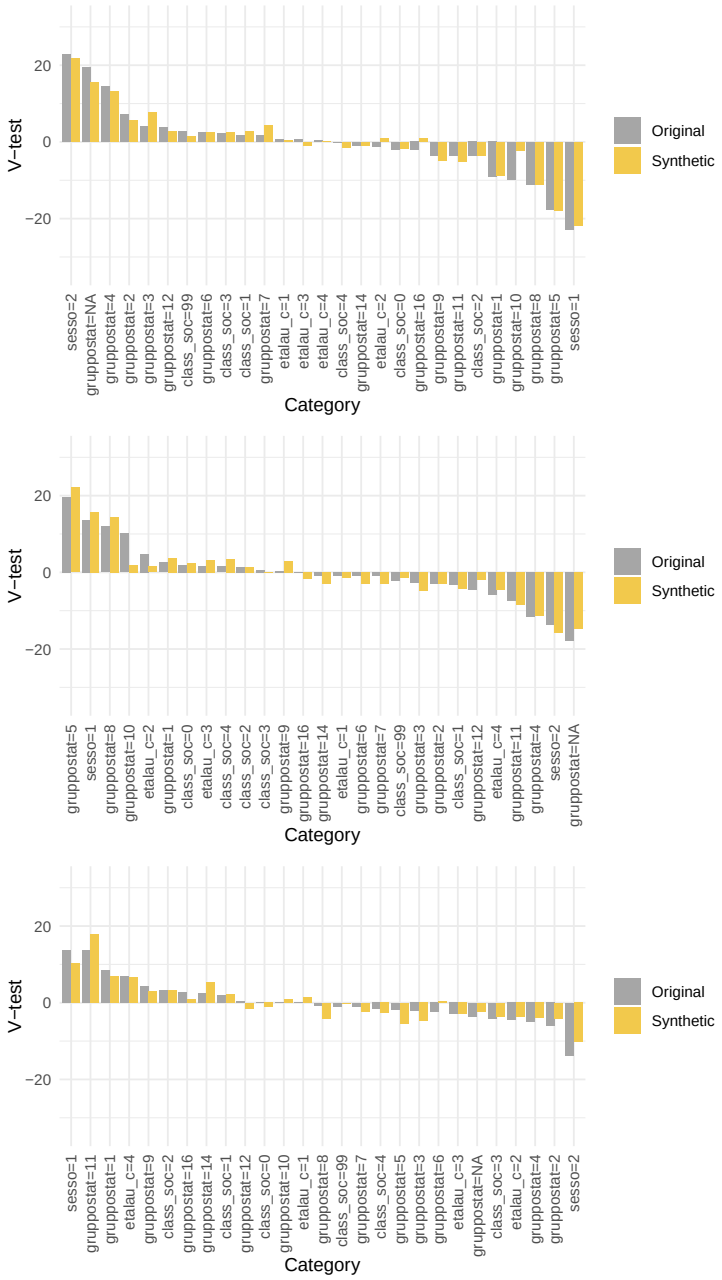


Figure F.5. v-TESTS FOR GEOGRAPHICAL MOBILITY SUPPLEMENTARY VARIABLES IN CLUSTER 1 (TOP PANEL), CLUSTER 2 (MIDDLE PANEL), AND CLUSTER 3 (BOTTOM PANEL): ORIGINAL VERSUS SYNTHETIC DATA.

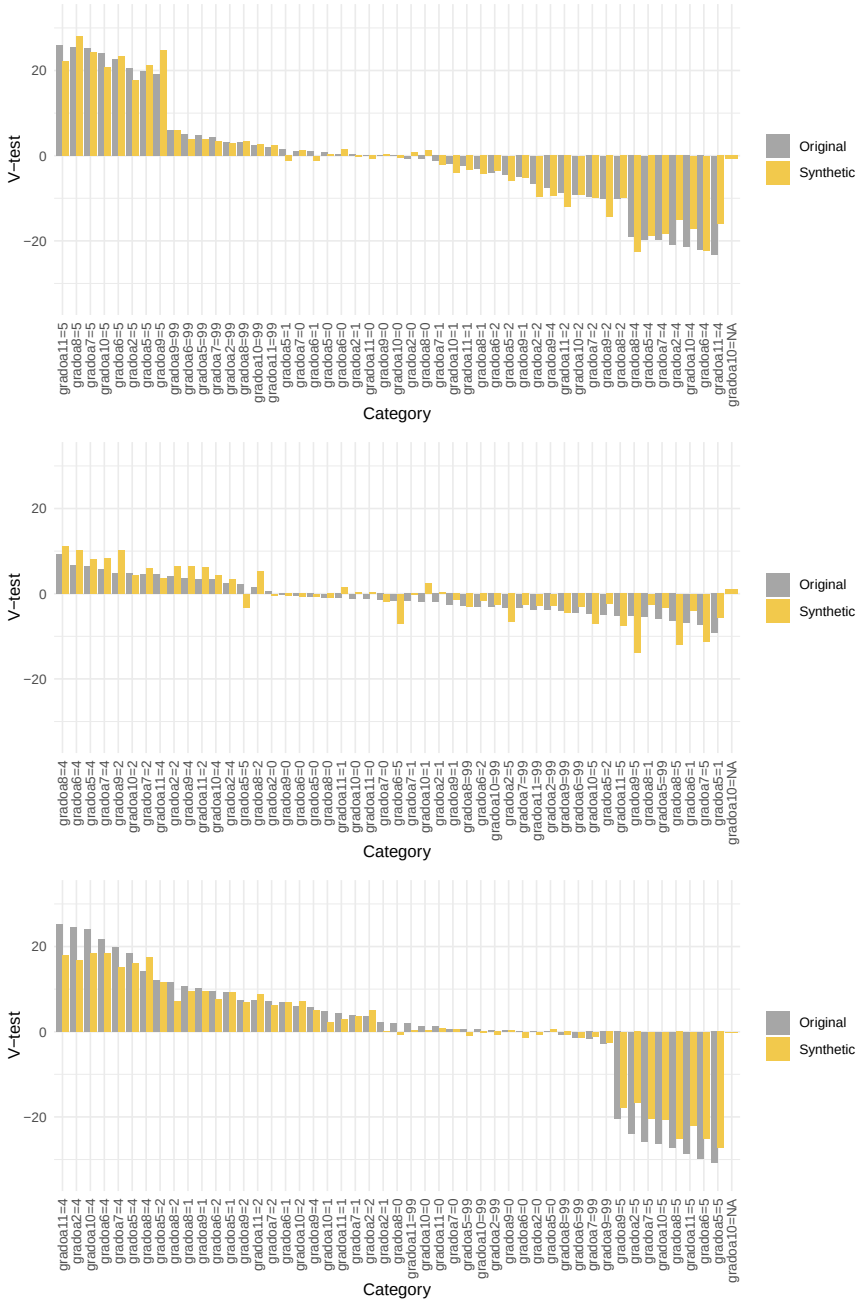
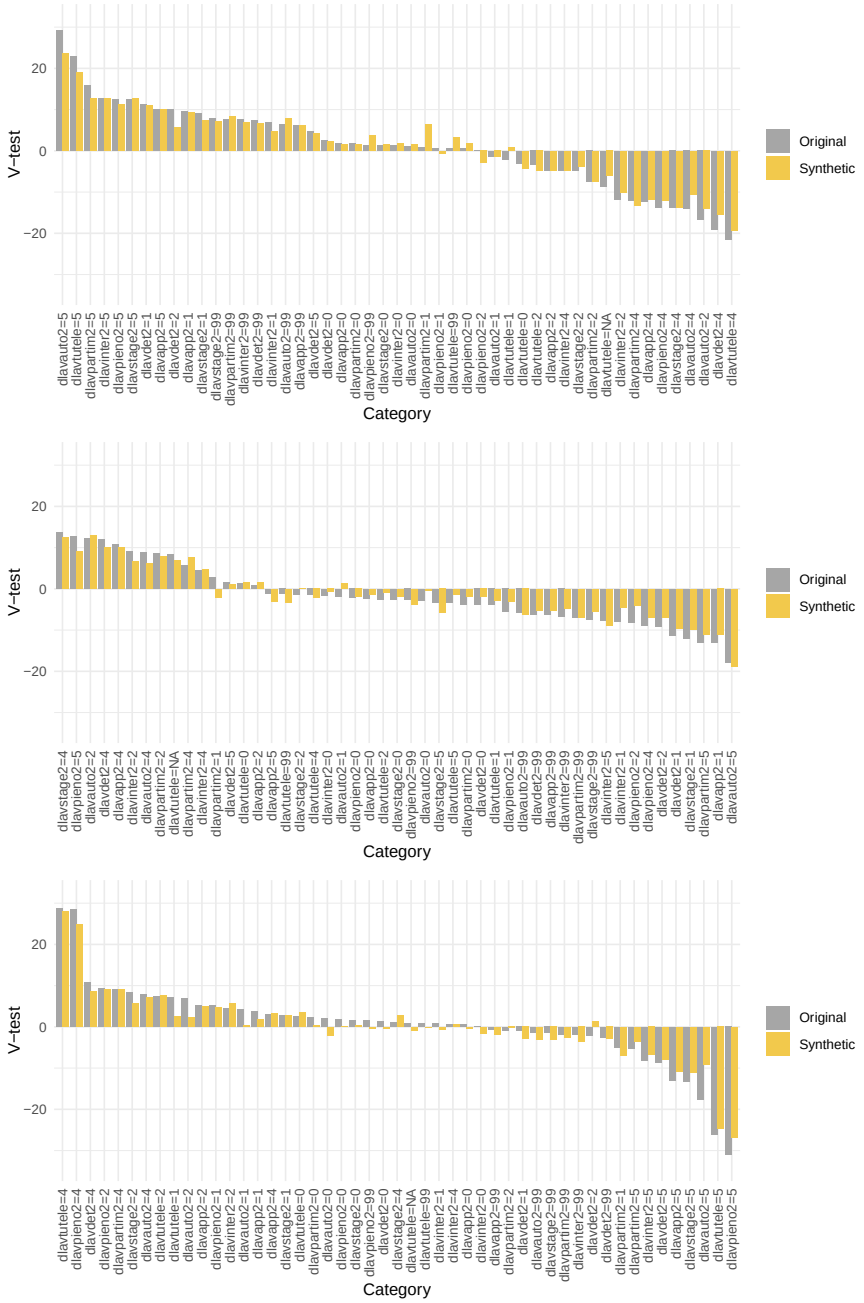



Figure F.6. v-TESTS FOR CONTRACTUAL ARRANGEMENTS SUPPLEMENTARY VARIABLES IN CLUSTER 1 (TOP PANEL), CLUSTER 2 (MIDDLE PANEL), AND CLUSTER 3 (BOTTOM PANEL): ORIGINAL VERSUS SYNTHETIC DATA.