TC Document

I. Basic Information for TC

▪ Country/Region:	REGIONAL
▪ TC Name:	AI Generated Synthetic Data (AIGSD) for better Productive Development Programs (PDP)
▪ TC Number:	RG-T4768
▪ Team Leader/Members:	Goni Pacchioni, Edwin Antonio (PTI/CTI) Team Leader; Pelaez Sierra Sergio (PTI/CTI); Becker Seco Rosario Paz (LEG/SGO); Medina Vasquez Exequiel Enrique (PTI/CTI)
▪ Taxonomy:	Research and Dissemination
▪ Operation Supported by the TC:	
▪ Date of TC Abstract authorization:	
▪ Beneficiary:	Chile and Perú
▪ Executing Agency and contact name:	Inter-American Development Bank
▪ Donors providing funding:	OC SDP Window 2 - Institutions(W2C)
▪ IDB Funding Requested:	US$280,000.00
▪ Local counterpart funding, if any:	US$0
▪ Disbursement period (which includes Execution period):	36 months
▪ Required start date:	August 2025
▪ Types of consultants:	Individuals
▪ Prepared by Unit:	PTI/CTI-Competitiveness, Technology, and Innovation Division
▪ Unit of Disbursement Responsibility:	PTI/CTI-Competitiveness, Technology, and Innovation Division
▪ TC included in Country Strategy (y/n):	Yes
▪ TC included in CPD (y/n):	No
▪ Alignment to the Institutional Strategy 2024-2030:	Productive development and innovation through the private sector

II. Objectives and Justification of the TC

2.1 **Objectives**. The main objective of this Technical Cooperation (TC) is to strengthen the institutional capacities of national counterparts by enhancing the design, selection, and evaluation of Productive Development Programs (PDP) (especially, programs to promote technical-extension and creation/adoption of innovation in firms) through the application of Artificial Intelligence Generated Synthetic Data (AIGSD) and Machine Learning (ML) methodologies. This will be achieved by: (i) creating privacy-compliant synthetic datasets that replicate the statistical properties of confidential administrative records, enabling robust evaluation of PDPs without compromising data confidentiality; (ii) improving the selection mechanisms of PDPs to identify candidates with the highest potential for positive outcomes; and (iii) building institutional capacity within implementing agencies to apply these methodologies independently.

2.2 **Justification**. PDPs face two challenges that limit their effectiveness and evaluation. First, while administrative data (such as employer-employee payrolls, tax and trade records, and social security registries) can significantly enhance program evaluation, confidentiality restrictions often prevent access to these data sources (Koenecke and Varian, 2020; Nagaraj and Tranchero, 2023; Barrientos et al., 2024). Second, the

mechanisms used to select suitable beneficiaries typically rely on heuristics or theory-based (deductive) approaches, rather than on inductive methods that learn from complex patterns in relevant data (Fontagné et al., 2013; Buffart et al., 2020).

2.3 AIGSD addresses the first challenge by generating synthetic datasets that statistically mimic real administrative records while preserving privacy. Causal inference, combined with ML methods, addresses the second challenge by enabling the development of more effective selection algorithms using historical program data with known outcomes. The application of these methodologies is expected to significantly enhance the design, implementation, and evaluation of PDPs throughout the region.

2.4 For generating synthetic data from confidential administrative records, we will implement several approaches. A well-established method is based on Generative Adversarial Networks (GANs). GANs consist of two neural networks competing against each other - a generator that creates fake data and a discriminator that tries to detect fake from real data. Through this adversarial training process, the generator learns to create increasingly realistic outputs. Conditional Tabular GAN (CTGAN) (Xu et al., 2019) excels at handling heterogeneous tabular data (mixing numerical and categorical variables) and enables conditional generation (e.g., synthesizing firm-level data for specific industries or regions). CTGAN has shown competitive privacy performance despite lacking differential privacy mechanisms that would add noise or randomness (Liu et al., 2025). More recent approaches include diffusion-based models and transformer-based models. Diffusion models learn to generate data by modeling the reverse process of gradually adding noise to data. They are trained to remove noise step by step, learning to transform pure noise into realistic outputs. TabDDPM (Tabular Denoising Diffusion Probabilistic Model) is one such model specialized in tabular data (Kotelnikov et al., 2023). Transformer-based models, such as Gretel Tabular Fine-Tuning, take a pre-trained language model and subject it to additional pre-training on diverse tabular datasets.

2.5 Challenges in synthetic data generation include preserving temporal dynamics and causal relationships. Regarding temporal dynamics, the synthesizers typically treat each row in a dataset as an independent observation missing the sequential logic and autocorrelation structures that govern time series. As for causal relationships, while the synthesizers reproduce correlation patterns, they typically do not preserve the selection mechanisms, quasi-random variation, and unobserved heterogeneity structures that make causal identification possible in the original data. On both sides, methods have advanced. Modern methods implement training processes that attempt to preserve step-by-step temporal transitions and mimic real data-generating processes that maintain causal relationships. However, time series and causal synthesis are still developing without the robustness of cross-sectional analysis (Brophy et al., 2023; Amad et al., 2024).

2.6 The quality of synthetic data generated by each method will be evaluated using both statistical similarity metrics (comparing distributions of key variables and their relationships) and utility metrics (the ability to derive similar conclusions from analyses performed on both real and synthetic data). This includes comparisons of marginal distributions of individual variables to test attribute fidelity (e.g., Kullback-Leibler (KL) divergence and chi-squared tests), joint distributions and correlations between variables to test bivariate fidelity (e.g., Jensen-Shannon distance), performance of downstream regression and ML tasks (e.g., Train-Synthetic-Test-Real (TSTR) methodology) and validation of causal inference estimation to test application fidelity.

2.7 Privacy evaluation metrics will also be examined. These include delta-presence analysis, which measures the maximum probability that a specific firm's data can be inferred from the synthetic dataset, and k-anonymization scores, which assess whether each synthetic record is indistinguishable from at least k–1 other based on key identifying attributes. Identifiability scores will be used to quantify re-identification risk by estimating how many original records can be uniquely matched to synthetic ones. Additional assessments will include membership inference attack simulations to test whether adversaries can detect the presence of a specific firm in the training data, and attribute inference evaluations to determine whether sensitive firm characteristics can be inferred from synthetic data patterns.

2.8 Prior work across a range of fields demonstrates how synthetic data can enhance analysis while safeguarding privacy and maintaining fidelity. In education, GANs combined with differential privacy have been used to replicate student performance patterns without exposing individual records (Liu et al., 2025). In agricultural economics, synthetic farm-level datasets have reproduced real-world elasticities and efficiency scores (Wimmer & Finger, 2023). In autonomous driving, diffusion models have generated rare-condition road scenes to improve system robustness (Goel & Narasimhan, 2024). In healthcare, privacy-preserving synthetic patient records have been used to augment clinical cohorts for reliable predictive modeling (Giuffrè & Shung, 2023). In finance, CTGANs have simulated transaction time series for stress-testing fraud detection and risk models (Assefa et al., 2020). To the best of our knowledge, however, these methods have not yet been applied to the research or evaluation of innovation policy.

2.9 To improve beneficiary selection, we will implement the following approach that combines causal inference, multi-program data integration, and synthetic data-augmented ML: (i) Treatment effect calculation: We will conduct causal impact evaluations using PSM (Propensity Score Matching) to estimate treatment effects for the primary program under evaluation (e.g, Startup Perú or Crece), measuring impacts on key outcomes such as innovation performance, productivity, and sales. Simultaneously, we will incorporate treatment effects from external programs implemented by the same or similar agencies, ensuring that each firm observation has both comprehensive characteristic data and a corresponding individual treatment effect measure; (ii) Data assembly: The internal and external datasets will be merged into a unified database containing firm characteristics (size, age, sector, geographic location, education levels, etc.) as features and standardized treatment effects as the target variable; and (iii) Synthesis: To address sample size limitations and class imbalance issues, this combined dataset will be processed using the AI-based synthetic data generation methodologies discussed in the previous section. The synthetic data generation will expand the training dataset while preserving the statistical relationships between firm characteristics and treatment effects observed across multiple programs, creating a robust and balanced training set.

2.10 Model training and validation: Multiple ML algorithms will be trained on the synthesized dataset to learn patterns between firm characteristics and program success across different PDP interventions. Models will be validated using holdout samples of real data to ensure that insights derived from synthetic training translate effectively to actual selection decisions.

2.11 This TC is consistent with the Inter-American Development Bank (IDB) Group Institutional Strategy: Transforming for Scale and Impact (CA-631) and is aligned with the objective of: (i) Bolster sustainable growth, as the TC result will contribute with the

designing and implementing of better public policies to increase productivity in the LAC region. The TC is also aligned with the operational focus areas of: (i) Institutional Capacity and Rule of Law and Citizen Security; and (ii) Productive Development and Innovation through the Private sector. The TC is also aligned with the CTI Sector Framework Document (SFD) (GN-2791-13), and with the Ordinary Capital Strategic Development Program (OC SDP) Window 2, Priority Area 3: Effective, Efficient, and Transparent Institutions (W2C) (GN-2819-14), which has as expected results strengthen the quality of institutions and policies as well as the provision of services and implementation of policies, to improve public management and promote the development of the private sector; and leverage digital transformation to promote more effective, efficient and transparent governments, better and more equitable opportunities for citizens, and more productive and innovative companies. The TC is also aligned with the IDB Group Country Strategy of Chile 2022-2026 (GN-3140-3) and Peru 2022-2026 (GN-3110-1) with the strategic area of strengthening the efficiency and quality of the public agencies and the strategic objective of increasing productivity by improving evaluability.

III. Description of activities/components and budget

3.1 **Component 1. Development and implementation of AI-based synthetic data generation methodologies (US$140,000).** This component will advance the following activities: (i) Data collection and preparation of administrative records: Identifying and obtaining access to relevant confidential administrative datasets (such as employer-employee payrolls, tax records, and social security registries) from participating institutions, establishing secure data handling protocols, and preparing the data for synthetic generation; (ii) Implementation and comparison of synthetic data generation methodologies; and (iii) Evaluation and validation of synthetic data quality.

3.2 **Component 2. Development of ML-based beneficiary selection algorithms (US$80,000).** This component will carry out the following activities: (i) Causal impact evaluation and data integration: Implementing PSM to estimate individual treatment effects for firms from the primary program under evaluation, while incorporating treatment effect data from external programs implemented by similar agencies, creating a unified dataset with firm characteristics as features and standardized treatment effects as target variables; (ii) Synthetic data generation and model development: Applying AI-based synthetic data generation methodologies to the integrated dataset to address sample size limitations and class imbalance, then creating and training multiple ML algorithms on the synthesized data to learn patterns between firm characteristics and program success across different PDP interventions; and (iii) Model validation and optimization: Evaluating model performance using holdout samples of real data through cross-validation, fine-tuning parameters to ensure synthetic training translates to effective real-world selection decisions, and optimizing selection algorithms to identify candidates with the highest potential for positive outcomes across diverse program contexts.

3.3 **Component 3. Institutional capacity building for AI methodology adoption (US$40,000).** This component will develop the following activities: (i) Development of comprehensive training materials and documentation: Creating user manuals, technical guides, and tutorial materials for both synthetic data generation and ML-based selection methodologies; and (ii) Implementation of hands-on training workshops: Conducting in-person and virtual training sessions with technical staff from participating institutions, focusing on practical application of the developed methodologies.

3.4 **Component 4. Data sharing and institutional cooperation (US$20,000).** This component will establish and strengthen the institutional frameworks necessary for sustainable data access and collaboration between the IDB and partner institutions in Chile and Peru. Activities will include: (i) developing and formalizing interagency cooperation agreements with data custodians such as Chile's SII and SERCOTEC, and Peru's SUNAT and ProInnovate, establishing clear protocols for secure data access, usage rights, and confidentiality provisions; and (ii) creating standardized data governance frameworks that ensure compliance with national privacy regulations while enabling synthetic data generation activities.

3.5 Budget. The total cost of this TC is US$280,000. The execution and disbursement period will be 36 months, and the unit responsible for disbursements will be CTI. The funds will be provided by OC SDP Window 2 - Institutions(W2C). The following table provides a breakdown of the budget by components and activities.

Indicative Budget (US$)

Activity/Component	Description	IDB/W2C	Total Funding
Component 1. Development and implementation of AI-based synthetic data generation methodologies	1.1. Specialized consulting in generation of synthetic data for causal inference	100,000	140,000
	1.2. Specialized consulting in evaluation and validation of synthetic data for causal inference	40,000	
Component 2. Development of ML-based beneficiary selection algorithms	2.1. Consulting for development of Machine Learning-based beneficiary selection algorithms for productive development program	80,000	80,000
Component 3. Institutional capacity building for AI methodology adoption	3.1. Consulting for development of training materials and technical documentation for AI methodologies	20,000	40,000
	3.2. Workshop organization in Chile and Peru	20,000	
Component 4. Data sharing and institutional cooperation	4.1. Consulting to prepare collaboration and confidentiality agreement documents between agencies for permanent access to administrative records of Chile and Peru	20,000	20,000
TOTAL		280,000	280,000

IV. Executing agency and execution structure

4.1 In accordance with the Operational Guidelines for Technical Cooperation Products (OP-619-4 Annex II), this technical cooperation will be executed by the IDB through the Competitiveness, Technology and Innovation Division (PTI/CTI), which will be responsible for contracting. Bank execution is justified because: (i) the TC requires

specialized expertise in AI and synthetic data generation that is not available within beneficiary countries' institutional frameworks; (ii) PTI/CTI has demonstrated technical capacity in innovation policy evaluation and beneficiary selection and can coordinate the complex, multi-disciplinary research required; and (iii) the IDB's regional mandate enables knowledge transfer and methodological standardization across Chile and Peru, facilitating future replication throughout the region.

4.2 For selection and contracting, the Bank's team will consider: (i) hiring individual consultants, in accordance with the provisions of the Complementary Workforce policy (AM-650); and (ii) contracting logistical services and other non-consulting services, in accordance with the Institutional Procurement Policy GN-2303-33 and its guidelines.

V. Major issues

5.1 Data access: Gaining access to confidential administrative records presents both legal and bureaucratic hurdles. However, it is expected that the TC "Improving evaluability during execution through continuous evaluation with administrative data" (ATN/OC-21360-RG) will facilitate this process. This TC aims to establish interagency cooperation agreements for administrative data sharing in both Chile and Peru—the same countries targeted by this TC—and to build capacity for exploiting these records. If successful, this TC will have already established the necessary legal frameworks and relationships with data custodians.

5.2 Privacy compliance: Despite the expected progress from TC ATN/OC-21360-RG, additional measures will still be implemented: (i) establishing formal data sharing agreements with clear confidentiality provisions specifically for synthetic data generation; (ii) implementing secure data handling protocols that comply with relevant privacy laws; and (iii) engaging legal experts from participating institutions early in the process to address compliance requirements.

5.3 Institutional adoption and cultural resistance: New AI-based methodologies may face resistance due to organizational inertia, skepticism about "black box" algorithms, or concerns about replacing human judgment in selection processes. Mitigation strategies include: (i) demonstrating tangible benefits through pilot implementations; (ii) ensuring transparency in how AI models make recommendations; and (iii) framing the new methodologies as decision-support tools that enhance rather than replace expert judgment.

5.4 Synthetic data effectiveness for causal inference applications: While synthetic data generation methods excel at preserving distributions, correlations and enabling downstream prediction tasks such as ML, their effectiveness for causal inference applications remains less established in the literature. To ensure that synthetic data maintains the causal structures necessary for reliable treatment effect estimation, we will implement specialized methodologies that model the actual data generation process (Amad et al., 2024). However, these adaptations may reduce the data's utility for other analytical applications. Still, regardless of causal inference performance, the generated synthetic data will provide value for broader data analysis tasks and ML applications and can be released for use by other researchers and practitioners.

VI. Exceptions to Bank policy

6.1 Exceptions to Bank policies are not expected.

VII. Environmental and Social Aspects

7.1 This Technical Cooperation is not intended to finance pre-feasibility or feasibility studies of specific investment projects or environmental and social studies associated with them; therefore, this TC does not have applicable requirements of the Bank's Environmental and Social Policy Framework (ESPF).

Required Annexes:

Results Matrix - RG-T4768

Terms of Reference - RG-T4768

Procurement Plan - RG-T4768