Innovative Approaches to Synthetic Data Generation: Insights from HealthData4EU Cluster Projects

On October 2, 2024, a session titled "Showcasing Innovative Research for Synthetic Data Generation" highlighted key achievements from three European projects under the **HealthData4EU Cluster: AISYM4MED, SYNTHEMA, and SECURED**. Each project addresses distinct challenges in healthcare, focusing on synthetic data, privacy-preserving technologies, and the use of artificial intelligence to improve medical outcomes.

The session was moderated by **Serena Battaglia, Project Officer of these sister projects at the European Commission (HaDEA – European Health and Digital Executive Agency)** and a key figure in the European open cloud movement, who provided a strategic overview of the discussions. Ms. Battaglia emphasised the growing importance of synthetic data in addressing healthcare challenges, particularly regarding data scarcity and privacy. She also underscored the need for collaboration across projects to achieve common solutions for data sharing, innovation, and privacy preservation within the healthcare sector.

Together with the three previously stated sibling projects—PHEMS, PHASE IV AI, and FLUTE—this cluster was established as part of a collection of initiatives funded under the same call, **Horizon-HLTH-2022-IND-13-02**. This call is a flagship effort designed to provide funding for creative initiatives that use technology to improve healthcare services and tackle important issues in the field. To address these problems, the call outlines several aims, including generating synthetic data, data anonymisation approaches, and scaling up multi-party computation. Through the promotion of safe, compatible, and open international health data

centres, these programs aid in the creation and uptake of creative data-driven solutions.

AISYM4MED: Evaluating the Quality of Synthetic Medical Data

Led by **Dr. Luis Rosado (Senior Researcher at Fraunhofer AICOS Portugal)**, the AISYM4MED project tackles a fundamental question in synthetic data generation: *How do we evaluate the realism, representativeness, and usefulness of synthetic medical data?* As synthetic data becomes more prevalent in addressing data scarcity and privacy concerns, it is critical to ensure that this data is reliable for healthcare applications.

AISYM4MED's innovative platform introduces an open-source library for data auditing, designed to automatically evaluate both synthetic and real data across multiple modalities, including time series, tabular, and imaging data. This library features a comprehensive set of metrics to assess the quality of synthetic data, helping medical professionals ensure that machine learning models trained on this data remain effective and trustworthy. Preliminary findings from the project demonstrate how the AISYM4MED platform can assist healthcare professionals in evaluating the synthetic data they use. The project focuses on bridging the gap between technical solutions and the practical needs of clinicians, offering tools to evaluate the realism and representativeness of synthetic data in medical settings.

SYNTHEMA: Tackling Data Scarcity in Rare Diseases

Dr. Sofia Tsekeridou (Senior Research & Innovation Manager at Netcompany – Intrasoft), a member of the SYNTHEMA project, addressed the challenge of accessing scarce health data for rare diseases and generating synthetic data in a secure, privacy-preserving manner. Rare diseases present a unique challenge for AI-driven healthcare, as the low number of patients and data silos across unconnected clinical sites hinder effective diagnosis and treatment.

SYNTHEMA proposes a novel solution: federating scarce data across dispersed clinical sites and health registries using federated learning models. This approach allows the secure and privacy-preserving training of synthetic data generation engines while maintaining data protection compliance, especially in rare hematological diseases. By federating data across borders and clinical sites, SYNTHEMA reduces bias and improves the quality of synthetic data, helping to generate "virtual patients" who can be used to enhance diagnostic capacities and support treatment decisions.

HEALTHDATA4EU

The project's focus on federated learning is particularly important in an era where privacy concerns and data protection regulations, such as GDPR, are paramount. By ensuring that data remains decentralised, SYNTHEMA provides a secure framework for accessing and utilising sensitive health data, addressing both ethical and legal challenges in synthetic data generation.

SECURED: Privacy-Preserving Technologies in Healthcare

Privacy is a major concern in healthcare, and the SECURED project, led by **Dr. Francesco Regazzoni (EU project coordinator at University of Amsterdam & Universitá della Svizzera Italiana)**, explores the use of privacy-preserving technologies to safeguard sensitive medical data. These technologies include various cryptographic methods such as homomorphic encryption (FHE), secure multiparty computation (SMPC), and differential privacy.

Dr. Regazzoni highlighted the advantages and limitations of these techniques, noting that while they offer significant protection for patient data, they can also present challenges in terms of scalability and processing speed. For instance, homomorphic encryption allows data to be analysed while still encrypted, ensuring privacy throughout the process. However, the computational demands of FHE are significant, making it less practical for certain healthcare applications where real-time analysis is required.

SECURED aims to balance these challenges by exploring how privacy-preserving technologies can be adapted to meet the specific needs of healthcare, ensuring that medical professionals, researchers, and developers can access the data they need without compromising patient privacy.

Finding Common Ground: The HealthData4EU Cluster's Collaborative Approach

The HealthData4EU Cluster aims to find common solutions to the complex challenges of health data sharing, privacy, and innovation. Across projects like AISYM4MED, SYNTHEMA, and SECURED, there is a shared goal: to harness the power of synthetic data and privacy-preserving technologies while maintaining the highest ethical standards.

As the session concluded, it became clear that collaboration between projects is key to overcoming these challenges. By developing frameworks that are flexible yet robust, the HealthData4EU Cluster is paving the way for future breakthroughs in synthetic data generation, rare disease research, and secure

data collaboration in healthcare, all supported by initiatives like the Horizon-HLTH-2022-IND-13-02 CALL.

This conference is a component of the European Big Data Value Forum (EBDVF), the main event organised by BDVA that unites the whole European data-driven AI research and innovation community in a shared learning environment, collaborative spirit, and celebration of accomplishments. To promote policy initiatives and improve industrial and research operations in the fields of data and artificial intelligence, EBDVF brings together professionals from the industry, business developers, researchers, and policymakers from throughout Europe and beyond.

Visit our project websites to learn more about potential future collaborations with the HealthData4EU Cluster and individual projects!

	● Website: https://synthema.eu/ ● LinkedIn: https://www.linkedin.com/company/synthema/ ● Contact: arantxa@australo.org
	● Website: https://aisym4med.eu/ ● LinkedIn: https://www.linkedin.com/company/aisym4med ● Contact: alorenzini@zabala.eu, inunez@zabala.eu
	● Website: https://secured-project.eu/ ● LinkedIn: https://www.linkedin.com/company/secured-project/ ● Contact: f.regazzoni@uva.nl