SECURED for Health: Scaling up privacy to enable the integration of the European health data space

Francesco Regazzoni[*xviii], Gergely Acs[xii], Albert Zoltan Aszalos[x], Christos Avgerinos[xv], Nikolaos Bakalos[xvi], Josep Ll. Berral[††], Joppe W. Bos[§], Marco Brohet[*], Andrés G. Castillo Sanz[‡], Gareth T. Davies[§], Stefanos Florescu[‖], Pierre-Elisée Flory[¶], Alberto Gutierrez-Torre[††], Evangelos Haleplidis [xi], Alice Héliou[¶], Sotirios Ioannidis [†], Alexander Islam El-Kady[xi, 14], Katarzyna Kapusta[¶], Konstantina Karagianni [xi], Pieter Kruizinga[‖], Kyrian Maat[*], Zoltán Ádám Mann[*], Kalliopi Mastoraki[†], SeoJeong Moon[§], Maja Nisevic[**], Balázs Pejó[xii], Kostas Papagiannopoulos[*], Vassilis Paliuras[xi xix], Paolo Palmieri[xiii], Francesca Palumbo[‡‡], Juan Carlos Perez Baun[xvii], Peter Pollner [x], Eduard Porta-Pardo[xiv], Luca Pulina[‡‡], Muhammad Ali Siddiqi[‖], Daniela Spajic[**], Christos Strydis[‖], Georgios Tasopoulos[*], Vincent Thouvenot[¶], Christos Tselios[xi], Apostolos P. Fournaris[xi]

[*]*University of Amsterdam (NL)*, [†]*Circular Economy Foundation (BE)*, [‡]*FHUNJ (ES)*, [§]*NXP (BE)*, [¶]*Thales (FR)*, [‖]*Erasmus Medical Center (NL)*, [**]*KU Leuven (BE)*, [††]*Barcelona Supercomputing Center (ES)*, [‡‡]*University of Sassari (IT)*, [x]*Semmelweis University (HU)*, [xi]*Industrial Systems Institute/R.C. ATHENA (GR)*, [xii]*BME (HU)*, [xiii]*University College Cork (IE)*, [xv]*Catalink (CY)*, [xvi]*ICCS (GR)*, [xvii]*EVIDEN (ES)*, [xiv]*Josep Carreras Leukaemia Research Institute (ES)*, [xviii]*Università della Svizzera italiana (CH)* [xix]*University of Patras (GR)*,

***Abstract*—In this paper, we present the SECURED project[1], aimed at improving privacy-preserving processing of data in the health domain. The technologies developed in the project will be demonstrated in four health-related use cases and with the involvement of SME's selected through an open funding call.**

I. Introduction

The 'Scaling up secure processing, anonymization and generation of health data for EU cross border collaborative research and innovation" SECURED project (https://secured-project.eu/), started in January 2023, aims to enhance the scalability and efficiency of multiparty computation, data anonymization, and synthetic data generation, focusing on private and unbiased artificial intelligence and data analytics. Specifically addressing challenges in secure multiparty computation protocols, data anonymization methods for health data, dynamic on-demand services for synthetic data generation, federation protocols for machine learning, and support for health technology providers, SECURED employs algorithmic improvements and implementation efficiency to scale up privacy technologies. The project targets well-being, prevention, diagnosis, treatment, and follow-up care in health-related data, addressing ethical and legal challenges. SECURED developed technologies are showcased in four real-world use cases, including real-time tumor classification, telemonitoring for children, synthetic data generation for education, and access to genomic data. SECURED technologies can be further evaluated by selected SMEs thought a public open call that will be opened at the end of the second year of the project.

[1]Funded in part by the European Union (EU), Grant Agreement no. 10109571. Views and opinions expressed are those of the authors and do not necessarily reflect those of the EU or the Health and Digital Executive Agency. Neither the EU nor the granting authority are responsible for them.

II. SECURED Concept and Architecture

The SECURED architecture provides a secure and trusted environment for decentralized, cooperative processing of health data, employing secure computation, anonymization (with preemptive de-anonymization assessment), and the creation of high-quality synthetic data. The vision is to enhance the sharing of health datasets in Europe by securely connecting EU health data hubs, the health data analytics research community, e-health SMEs, and end users. The SECURED approach involves two parallel yet interacting innovation flows: the data flow and the processing flow, detailed in Fig. 1.

A. Data flow

The SECURED data flow focuses on enhancing data privacy through anonymization, de-anonymization validation, and synthetic data generation. The first goal is to enable health data producers to properly anonymize their datasets using SECURED's tools, validated through de-anonymization attacks. The second goal involves augmenting datasets through privacy-preserving synthetic data generation, ensuring sufficient volume for AI model training and data analysis.

For anonymization, SECURED provides a suite of tools and an assessment mechanism generating an anonymization "score" that can intuitively convey the level of protection offered, allowing data producers to meet specific protection requirements. If the score falls below a set threshold, the anonymization process can be adjusted. To evaluate the usefulness of datasets for AI model training, the volume of anonymized datasets is assessed. SECURED's synthetic data generation techniques help enhance datasets to adequate volumes, combining with unbiasing processes to prevent bias in the final anonymized datasets. The output is

979-8-3503-4859-0/DATE24/

unbiased, anonymized, actionable datasets stored in data producer premises (e.g., EU data hub's data lakes), registered in SECURED Innohub Knowledge base and dataset inventory, accessible to other stakeholders upon request.

B. Processing flow

The SECURED processing flow is focused on scaling up existing private processing technologies, and designing novel ones to enable collaborative, privacy-preserving analysis and processing of health data, without requiring data holders to share private datasets with other parties. SECURED develops a secure multiparty computation (SMPC) software library that supports SMPC-enabled operations for ML/DL (including AI model training, AI model updating, and AI inference support). Using this library, stakeholders can adapt their AI-based data analysis tools using an SMPC transformation process, enabling the formation of clusters of data producers and processors that collaboratively compute private datasets without actually sharing those data. The library supports the training of AI models using a cluster dataset following a federated learning paradigm. The federation infrastructure is supported by the SECURED Innohub. The processing flow is aimed at ultimately allowing Innohub members to contribute to the SECURED federation with their clusters of AI models (local cluster client models collaboratively trained through the SMPC-enhanced Innohub member tools). The produced AI models (aggregated from various clusters in the federation) are always anonymized (following a variant of the SECURED data flow) and stored in the SECURED Knowledge base. These AI models can be shared with the SECURED Innohub members, thus forming a "privacy-preserving SECURED marketplace" and can also be used for the SECURED synthetic data generation mechanism.

C. Innohub

SECURED aims to create and manage a privacy-enhancing hub, the *SECURED InnoHub*, that provides tools, services, and support for the privacy-preserving processing of health data to stkeholders in the healthcare domain, including researchers, innovators, health data users as well as EU data Hubs across Europe. The goal of the hub is to enable stakeholders to leverage available datasets to perform accurate, distributed data analytics, while preserving the privacy of the data. The SECURED Innohub promotes collaboration among parties by acting as a one-stop collaboration point, for sharing results and collaboratively building expertise. As the data analytics technology most widely adopted for health data is machine/deep learning, the hub focuses the offered tools and services on enhancing the privacy of ML/DL solutions, including an SMPC-capable toolbox that can operate in various modes under a SECURED federation infrastructure. The SECURED Innohub will bring together providers and consumers of health data and offer them a trusted, secure and privacy-preserving environment to research, test their solutions, and collaborate.

III. SECURED Privacy-Preserving Techniques

A. Homomorphic Encryption

Homomorphic Encryption (HE) enables functions to be evaluated on encrypted data. For SECURED, this allows the inference or training of AI models to be performed while the confidentiality of the medical data is still guaranteed [1], [2]. We mainly focus on Fully HE (FHE), which allows arbitrary polynomial functions to be evaluated, and schemes including BGV [3], BFV [4], and CKKS [5], as well as TFHE [6].

Following the growth of research on HE, a number of open-source libraries/frameworks have been developed that offer HE functionality. In SECURED we start from these existing works, including HElib [7] developed by IBM, SEAL [8] developed by Microsoft, TFHE [9] an open-source project that uses Fast Fully Homomorphic Encryption over the Torus, HEAAN [5] developed by HEAAN CryptoLab [10]. These libraries support different HE schemes and offer various trade-offs between speed, memory, data transfer, data representation, and supported operations. Because of the intrinsic complexity of HE and the diverse nature of every scheme, HE libraries/frameworks are mainly focused on solving specific problems, thus no library can be considered the best overall. SECURED will provide support for choosing the most suitable solution for a target problem.

B. Secure Multi-Party Computation

Secure Multi Party Computation (SMPC) techniques relevant to SECURED can be divided into two classes: Garbled Circuits and Secret Sharing. Garbled circuits [11] allow parties to compute together with private inputs while minimizing the risk of private inputs becoming known to other parties. Secret sharing is based on the idea that a secret can be spread over multiple shares (thus multiple parties), where all or a majority of shares need to be combined to retrieve the secret.

A number of frameworks exist for SMPC. The ABY framework [12] allows quick conversion between different data representations, which is a challenge in standard SMPC settings. MP-SPDZ [13] encompasses multiple SMPC protocols and security models and acts as a unifying tool to benchmark SMPC protocols against each other. Other implementations have been designed specifically for privacy-preserving deep learning with SMPC components, such as Chameleon [14] in a two-party setting and SecureML [15] with three parties where SMPC is combined with HE. These works serve as guidelines of what can be integrated into the SECURED pipeline.

C. Data Anonymization, de-Anonymization and Private Synthetic Data Generation

There is consensus on the benefits of sharing health data for medical research [16], but it is a complex task that requires recollection, permissions, and security measures as this kind of data is sensitive. For this reason, data anonymization, de-anonymization and synthetic data generation have recently grown along with Deep Learning techniques, as it provides a way to remove sensitive personal information and generate new data that can be used for analysis (as base dataset or for

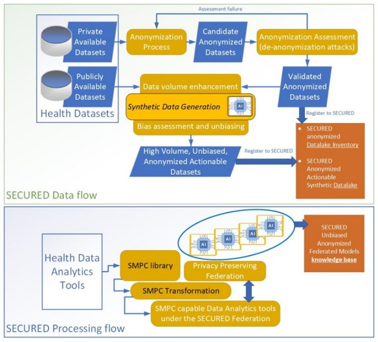

Fig. 1. The data flow and the processing flow in the SECURED architecture

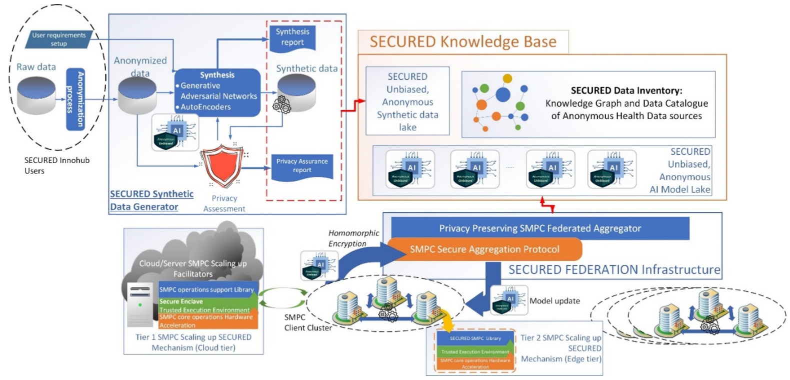

Fig. 2. The architecture of SECURED

data augmentation) as well as for education, as shown in the use cases.

However, the anonymization and generative techniques are not perfect in accuracy, usability, and privacy, and are subject to advanced de-anonymization and re-identification attacks. Generative models can be improved to create better synthetic data, and the usability of these methods can be improved, for example, by providing the user with the ability to condition the generation with input parameters [17]. Finally, the generated data should be different enough from the original data so that it cannot be re-identified with engineered attacks [18]. SECURED will try to address these challenges during the project to provide good, usable, and secure methods to anonymize and generate health data.

IV. Legal Implications

Techniques for comprehensive data collection, processing, and sharing activities are underpinned by a legal framework that addresses data and privacy protection, AI, cybersecurity, and medical devices. The European Commission has set up a strategic approach for data. It is recognized that data is pivotal for innovation in a data-agile health economy, but it is crucial that the handling and processing of patient data is maintained under the relevant rules and principles at different jurisdictional levels. In healthcare, the right to privacy and data protection is shaped by a multitude of hard-law and soft-law frameworks, including the European Convention on Human Rights and the Council of Europe's Convention for the Protection of Individuals concerning the automatic processing of personal data. Some aspects needs to be further clarified.

For instance, the status of whether SMPC anonymizes or pseudonymizes data remains unclear. ENISA, for instance, categorizes SMPC as a pseudonymization technique. Despite securing input data, there is a potential risk that the generated output could reveal personal information if the input data itself contains personal data. In addition, with synthetic data, it is debated where synthetic data transitions from being personal to non-personal. The European Data Protection Supervisor (EDPS) recommends privacy assurance assessments to determine the extent to which data subjects can be identified when dealing with synthetic data.

To ensure the safeguarding of patient health and safety, the EU has introduced the In Vitro Diagnostic Regulation (IVDR) and the Medical Devices Regulation (MDR). These frameworks establish safety requirements and information technology measures applicable to all medical devices. Software, in particular, may qualify as a medical device if intended for use, alone or in combination, for a purpose specified in the medical devices regulation. Alongside IVDR and MDR, the Regulation on Health Technology Assessment (HTA) aims to harmonize health technology assessments across EU Member States. The HTA is defined as a multidisciplinary, systematic, and transparent process to evaluate the effectiveness of health technologies.

V. SECURED Pilots

A. Real-Time, ultrasound-guided neurosurgery

Functional Ultrasound Imaging (fUSI) is a groundbreaking neuroimaging technique that visualizes cerebral blood flow, akin to Functional Magnetic Resonance Imaging (fMRI). In healthcare, fUSI can revolutionize early diagnosis and treatment of brain diseases, especially during surgery. As a valuable adjunct to Magnetic Resonance Imaging (MRI) in neurosurgery, fUSI compensates for intraoperative brain changes, enhancing tumor classification and aligning with preoperative MRI data. However, this co-registration method encounters two primary limitations: Control-Point Identification and Computational Demands. Concerning the first, control points are crucial for accurately recalculating the grid upon which the MRI data is interpolated, however, identifying corresponding control points in both MRI and fUSI images is challenging. Concerning the second, updating MRI scans, especially those with high voxel counts and intricate deformations, is computationally intensive. To tackle the above problems, we make two condierations. First, we will use intraoperative, 3D, ultrafast-Doppler ultrasound imaging to guide the MRI realignment process.Second, we will use an HPC cluster that can update the MRI scans in real time during the operation. We will use state-of-the-art HE technology to ensure that MRI data is securely stored on the cluster and that the interpolation/regridding, and the multimodal data (MRI & fUSI) fusion, is only performed using encrypted MRI data without compromising its privacy.

B. Telemonitoring for children

Cancer centres have increased their use of telehealth as part of the cancer care delivery continuum . Patient-centered cancer care includes high level of decentralization and broadens precision medicine from "the right treatment, for the right patient, at the right time" to include "in the right place". The ability to undergo chemotherapy treatment at home without jeopardising patient safety is a main line of innovation in oncology departments. The development of models based on clinical data sets from patient telemonitoring demands novel techniques that ensure data security. SECURED tools for federated learning and scientific data synthesis, as well as reduced computing costs, are critical to meeting clinicians' expectations. In addition to these promising lines of research, the incorporation of wearable and tracking devices as part of the telehealth experience is already emerging as a future cancer care model. Ensuring that telehealth platforms can track these novel technologies will be critical for coalescing data into the most effective telehealth visit possible. The tools in SECURED's anonymization techniques are also critical to accomplishing this.

C. Synthetic-data generation for education

Using the SECURED architecture, this pilot will facilitate the education of doctors by using synthetic data generated based on patient data. For instance, educators can integrate ML tools into their daily practice when generating exams, guaranteeing that questions never repeat. Online education tools have transitioned to robust learning management systems (LMS) that offer a wide range of features. As more educational institutions and learners rely on digital platforms for remote learning, the access and storage of data have raised significant challenges. The primary concern lies in how these platforms handle sensitive information. Balancing the need for data-driven insights to improve educational outcomes with safeguarding individuals' privacy rights is an ongoing challenge. The SECURED tools directly address these challenges by enabling the development of privacy-preserving education environments and consultation platforms tailored for highly sensitive healthcare data.

D. Access to Genomic Data

The availability of human genetic data has grown significantly. Examples include the Database of Genotypes and Phenotypes (dbGaP) and the National Cancer Institute's (NCI) Genomic Data Commons. Initiatives have also started national biobanks, i.e., longitudinal cohorts with data from hundreds of thousands of volunteers. A famous example is the UK BioBank, which has led to key discoveries in the genetic architecture of several diseases. While this trove of information has extraordinary potential for biomedical research, particularly when analyzed with AI, several ethical, legal, and technical barriers currently limit the impact that these data can have.

SECURED tools will be used to overcome several of these challenges. For example, we will use SECURED's federated learning to train models on independent genetic datasets from several sources. Once this model is successful, researchers will be able to analyze and train models on genetic data from different projects without the need to download local copies of these datasets. Similarly, this could also potentially allow biobanks to provide access to researchers to the data to ML models while preserving patient privacy. As part of this pilot, the limits of patient anonymization using genetic, environmental, and clinical data will be assessed.

VI. Conclusions

SECURED aims to increase the efficiency of privacy-preserving data processing by scaling up multi-party computation, data anonymisation and synthetic data generation. Focusing on private and unbiased AI and data analytics, it will demonstrate technologies developed in health-related use cases like real-time tumor classification, telemonitoring for children, education, and access to genomic data. SECURED will also analyse the current ethical and legal challenges associated with data sharing and privacy-preserving technologies.

References

[1] F. Regazzoni, P. Palmieri, and A. P. Fournaris, "Treating a different kind of patient: curing security weaknesses in digital health systems of the future," in *9th Intl. Workshop on Advances in Sensors and Interfaces, IWASI 2023*. IEEE, 2023, pp. 99–102.

[2] Z. Á. Mann, C. Weinert, D. Chabal, and J. W. Bos, "Towards practical secure neural network inference: The journey so far and the road ahead," *ACM Computing Surveys*, vol. 56, no. 5, 2023.

[3] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "Fully homomorphic encryption without bootstrapping," *Electron. Colloquium Comput. Complex.*, vol. TR11-111, 2011.

[4] Z. Brakerski, "Fully homomorphic encryption without modulus switching from classical gapsvp," in *CRYPTO 2012*, 2012.

[5] J. H. Cheon, A. Kim, M. Kim, and Y. S. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *ASIACRYPT 2017*, 2017, pp. 409–437.

[6] I. Chillotti, N. Gama, M. Georgieva, and M. Izabachène, "Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds," in *ASIACRYPT 2016*, ser. Lecture Notes in Computer Science, 2016.

[7] S. Halevi and V. Shoup, "Algorithms in HElib," in *CRYPTO 2014*, 2014.

[8] "Microsoft SEAL," https://github.com/Microsoft/SEAL, Jan. 2023, Microsoft Research, Redmond, WA.

[9] I. Chillotti, N. Gama, M. Georgieva, and M. Izabachène, "TFHE: fast fully homomorphic encryption over the torus," *J. Cryptol.*, 2020.

[10] "Heaan," https://github.com/snucrypto/HEAAN, Dec. 2023, CryptoLab Inc, Seoul, Korea.

[11] A. C.-C. Yao, "How to generate and exchange secrets," in *27th annual symposium on foundations of computer science (Sfcs 1986)*. IEEE, 1986, pp. 162–167.

[12] D. Demmler, T. Schneider, and M. Zohner, "ABY - A framework for efficient mixed-protocol secure two-party computation," in *NDSS*, 2015.

[13] M. Keller, "MP-SPDZ: A versatile framework for multi-party computation," in *CCS 2020*, 2020.

[14] M. S. Riazi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar, "Chameleon: A hybrid secure computation framework for machine learning applications," in *AsiaCCS 2018*, 2018.

[15] P. Mohassel and Y. Zhang, "Secureml: A system for scalable privacy-preserving machine learning," in *2017 IEEE S&P*, 2017.

[16] K. P. Seastedt, P. Schwab, Z. O'Brien, E. Wakida, K. Herrera, P. G. F. Marcelo, L. Agha-Mir-Salim, X. B. Frigola, E. B. Ndulue, A. Marcelo *et al.*, "Global healthcare fairness: We should be sharing more, not less, data," *PLOS Digital Health*, vol. 1, no. 10, p. e0000102, 2022.

[17] T. Weber, M. Ingrisch, B. Bischl, and D. Rügamer, "Implicit embeddings via gan inversion for high resolution chest radiographs," in *MICCAI Workshop on Medical Applications with Disentanglements*, 2022.

[18] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," *CoRR*, vol. abs/2301.13188, 2023.