

# The SERUMS tool-chain: Ensuring Security and Privacy of Medical Data in Smart Patient-Centric Healthcare Systems

V. Janjic, J.K.F. Bowles, A.F. Vermeulen, A. Silvina  
*School of Computer Science, University of St Andrews*  
*St Andrews, United Kingdom*  
{vj32,jkfb,afv,as362}@st-andrews.ac.uk

M. Belk, C. Fidas, A. Pitsillides  
*Department of Computer Science, University of Cyprus*  
*Nicosia, Cyprus*  
{belk,fidas,andreas.pitsillides}@cs.ucy.ac.cy

M. Kumar, M. Rossbory  
*Data Analysis Systems*  
*Software Competence Center Hagenberg*  
*Hagenberg, Austria*  
{mohit.kumar,michael.rossbory}@scch.at

M. Vinov  
*IBM Research Laboratory*  
*Haiifa, Israel*  
vinov@il.ibm.com

T. Given-Wilson, A. Legay  
*Universite Catholique de Louvain*  
*Louvain-la-Neuve, Belgium*  
{thomas.given-wilson,axel.legay}@uclouvain.be

E. Blackledge  
*Sopra Steria*  
*Edinburgh, United Kingdom*  
euan.blackledge@soprasteria.com

R. Arredouani, G. Stylianou and W. Huang  
*Accenture B. V.*  
*Amsterdam, Netherlands*  
{r.arredouani,georgios.stylianou,wanting.huang}@accenture.com

**Abstract**—Future-generation healthcare systems will be highly distributed, combining centralised hospital systems with decentralised home-, work- and environment-based monitoring and diagnostics systems. These will reduce costs and injury-related risks whilst both improving quality of service, and reducing the response time for diagnostics and treatments made available to patients. To make this vision possible, medical data must be accessed and shared over a variety of mediums including untrusted networks. In this paper, we present the design and initial implementation of the SERUMS tool-chain for accessing, storing, communicating and analysing highly confidential medical data in a safe, secure and privacy-preserving way. In addition, we describe a data fabrication framework for generating large volumes of synthetic but realistic data, that is used in the design and evaluation of the tool-chain. We demonstrate the present version of our technique on a use case derived from the Edinburgh Cancer Centre, NHS Lothian, where information about the effects of chemotherapy treatments on cancer patients is collected from different distributed databases, analysed and adapted to improve ongoing treatments.

**Keywords**—Medical data, Smart Healthcare, Data Sharing, Privacy, Security, Personalised Medicine

## I. INTRODUCTION

The healthcare systems of the future will be highly decentralised, integrating home-, work- and environment-based monitoring systems with existing hospital diagnostic systems. The benefits of integrating such a variety of systems and information on patients include a reduction of costs and travel-associated risks while allowing patients to get faster diagnostics and better medical treatments that more accurately suit their needs. As a consequence, medical data will

need to be collected from a variety of sources and exchanged in a variety of ways, including over public networks that cannot be implicitly trusted. At the same time, however, we have stricter regulations on ownership and handling of personal data. Transnational standards for data protection, such as the EU General Data Protection Regulation<sup>1</sup>, will need to be combined with local regulations, giving very strict rules about who is allowed to access (parts of) patient data. Complying with data protection regulations whilst facilitating data exchange and analytics in a decentralised way is a key challenge for future healthcare systems.

In this paper, we describe a methodology and complete tool-chain that will be developed over the course of the ongoing EU H2020 project SERUMS<sup>2</sup> to address safe, secure and privacy-preserving storage, access, communication and analysis of the medical data in future-generation smart health centres. Our main goal is to put patients at the centre of the future healthcare provision in Europe, enhancing their personal care and maximising the quality of treatment that they will receive, whilst ensuring trust in the security and privacy of their confidential medical data.

To reduce the scope of the paper, we restrict our attention to a subset of the SERUMS technologies. We propose a universal format for patient records, to allow a uniform representation of patient data across different use cases and describe its implementation. We describe FlexiPass, an

<sup>1</sup>Information on GDPR can be found at <https://gdpr-info.eu/>

<sup>2</sup>Securing Medical Data in Smart Patient-Centric Healthcare Systems (SERUMS): <https://serums-h2020.weebly.com>

authentication mechanisms to access these records, together with the application of blockchain technology to control permissions, ensuring that only allowed staff have access to required parts of patient records, and to save the access history of all records. We describe a novel, privacy-preserving data analytics mechanism which ensures that the analytics model itself does not accidentally leak sensitive information. Finally, we present a data fabrication approach that allows the generation of synthetic but realistic data, given a strict format of patient records and dependency rules between its elements. In the context of the SERUMS project, we only use generated synthetic data for the development and verification of our technologies, but we will furthermore prove formally the closeness of the synthetic and real data. For illustration, we present one of SERUMS real-world use cases on predicting toxicity levels of cancer treatments.

## II. BACKGROUND

The emergence of Internet-of-Things (IoT) technology is having a profound effect on the development of modern healthcare systems. Traditionally healthcare systems were highly centralised with data relevant to a patient, as well as the devices used to obtain this data (e.g., blood pressure monitors, CT scanners), residing in a central location, for example within a hospital. From a security and privacy point of view, collecting, storing and processing such data was relatively simple, since the data only needed to be communicated over trusted networks. However, as personal medical devices become cheaper and more prevalent, and there is an increased realisation of the benefits of integrating a variety of health data sources for improved healthcare provision, new significant challenges emerge with sharing private and confidential data across public networks. In particular, we need to be able to ensure:

- Trust: Patients must be able to trust that systems operate as intended and that their data is fully protected.
- Security, Privacy and Anonymity: Systems must operate efficiently and guarantee the best possible quality of healthcare, whilst simultaneously providing high levels of security and expectations on data privacy and anonymity.
- Data Control: Patients must have full control of their data according to expectation and law, whilst allowing medical staff data access as required.
- Regulation Compliance: The smart healthcare system must comply with regulations at various levels, including GDPR, local legislation and policy that may at times conflict with the above goals or other legislation.

SERUMS tackles the above problems by (1) addressing security and protection of shared medical data across untrusted networks; (2) integrating personal medical data, coming from various sources, into coherent and structured smart patient records; (3) enabling data analytics techniques over distributed data; and (4) developing authentication and trust

mechanisms that will ensure that only properly authorised staff have access to (parts of) personal and medical data. At the same time, we consider world-leading levels of compliance to existing and emerging legal and ethical standards.

## III. SERUMS TOOL-CHAIN

Figure 1 gives an overview of the SERUMS tool chain and the overall process of accessing data across a distributed healthcare system. The core of it is a centralised data lake that holds the smart patient records (see Section IV-A). Note that, while the patient records are centralised, the data in them may refer to databases distributed inside and outside of the hospital environment. These records contain all information about the patients, from static information such as date of birth, gender and contact information, to vital information such as weight, body mass index, allergies, to dynamic information about treatments and examinations. Some of the data for the records will be collected from within the healthcare system over trusted networks, while other may be collected from personal health monitoring devices, etc. Data sent over untrusted networks must be secured using data encryption mechanisms.

When staff needs to access patient data, they first log in to the central healthcare system using secure authentication mechanisms. In the SERUMS project, our aim is to develop personalised and adaptive multi-factor user authentication schemes (see Section IV-D). Once the user logs in, their access rights are checked using the blockchain backend which is linked to a distributed blockchain database. Different classes of users (e.g. patients, GPs, specialists, insurers) will have different levels of permissions, according to GDPR and other legal and ethical regulations. For example, the patient has full access to their record, while a specialist can only access parts of the record that are relevant to them. The blockchain ensures that only authorised agents can access the data, and depending on permissions, possibly only be part of the data. The blockchain contains all access rules and transitions, and keeps a record the data access history. Note, however, that no actual data is stored in the blockchain.

Once the user is authenticated and the access rights are checked, the requested data from the smart patient records data lake is sent back to the user. If the user does not have full access rights to the record, the data transfer may involve masking parts of the data, i.e., hiding parts of the record that the user has no access right to see. The access transaction itself is stored in the blockchain database.

Finally, different kinds of analysis will need to be performed on patient data. In the SERUMS project, we focus on deep learning analytics to drive diagnostics and prediction of treatment outcomes (see the use case in Section V-A). Since the data referred to from the smart patient records is distributed, and we assume that some of this data cannot leave the place where it is stored, the analytics will also need to be performed in a distributed way. We need to

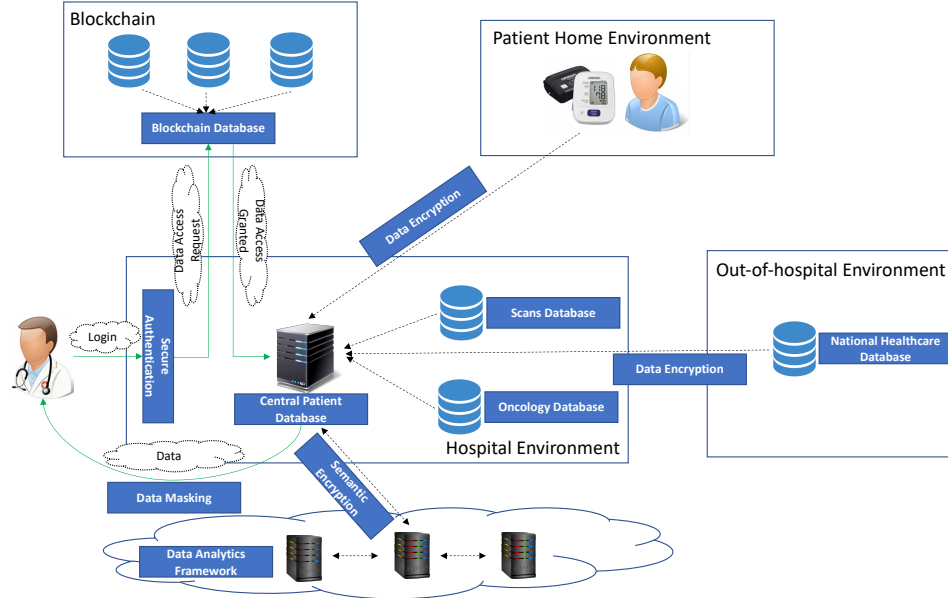


Figure 1. The overview of SERUMS tool-chain

make sure that no unsafe information is revealed by the learning models, as well as to ensure security of the data communicated between the central patient record database and the analytics model (which may reside in the cloud). In this context, our aim is to develop privacy-preserving distributed deep-learning analytics models for data analysis (see Section IV-C).

For the purposes of developing, verifying, and testing the complete SERUMS infrastructure, the SERUMS project will use synthetic instead of real patient data, to avoid any privacy and security concerns. Data fabrication (See Section IV-B) technology allows us to rapidly generate large volumes of data that is the same in terms of structure as real data, but which was synthetically generated. The strict format of the smart patient records, the formally defined rules on possible values for each field, the relationships between different fields and the well-formulated data interaction rules, makes it possible to automate this process.

#### IV. SERUMS TECHNOLOGIES

##### A. Smart Patient Records

Good organisation of patient data is essential to the smooth and correct operation of any health system. Furthermore, new legislation for privacy and ownership of the data (such as GDPR) together with a highly-decentralised organisation of modern health providers impose additional requirements for health data. Ideally, patient data should be owned by the patient, and only they have full access to their data. Other system users, such as specialists, general practitioners and insurers, are expected to have access to parts of the data relevant to the services they provide (e.g. diagnostics, treatment, insurance etc.). In addition to access restrictions,

the distributed nature of health systems means that we cannot assume that data is stored in one central location. Secure communication of data across untrusted networks might be required at any point the patient record is accessed. To develop a generic infrastructure for safe and secure communication of distributed medical data, it is highly desirable for the patient data (including pointers to any data that resides on remote systems) to be stored in a precise and machine-readable format.

In the SERUMS technology tool-chain, the Smart Patient Record represents a central information source for information about patients registered in Europe. These records aggregate a complete patient medical history across approved healthcare providers. The information in a single record includes both relatively static information (such as name, age, address, type of insurance, allergies) and highly dynamic information (such as undergoing treatments, results of scans and hospital admissions). For each healthcare institution, smart patient records will reside in a Smart Healthcare Data Lake. The Smart Patient Record Format represents metadata that describes the data in the records. We propose a universal format for patient records that can be used for describing different use cases within SERUMS and is applicable to future healthcare systems.

Our universal format is based on the concept of data vault [16], which consists of hubs (unique business keys), links (that represent associations between hubs) and satellites (where attributes of the hubs and links are stored). The general data vault has unlimited types of hubs, links and satellites to model real-world data. The SERUMS project has introduced a more limited type of hub, link and satellite classification [19] to force a more generalised view of all data

sources. This will support scaling [20] of the data vault. We propose a Time-Person-Object-Location-Event (T-P-O-L-E) data vault as a universal smart patient record format, such that:

- Time: the dates and times of events are stored in Coordinated Universal Time.
- Person: information about patients is stored using the concept of "Golden Nominal". This type of record is a single person record with a unique reference to that person.
- Object: other referable entities that are stored, including organisations (hospital, bank, medicine suppliers etc.), physical objects (medicine, bank cards, vehicles, hospital beds), buildings etc.
- Location: described by latitude, longitude and altitude [18].
- Event: an abstraction of any event or action in the real-world, including scans, home visits by a doctor and treatments.

The T-P-O-L-E data vault supports a future-proof design of the healthcare solution by enabling adding data at any point with full history capabilities. This model is the basis for the single-truth records data sharing and processing engine of the SERUMS project. The solution then uses advanced security to protect the information in a cross-country configuration respecting patient consent. This health record system will be able to support evolving coordinated services [25]

### B. Data Fabrication

IBMs Data Fabrication Platform (DFP) is a web based central platform that provides a consistent and organisational-wide methodology (rule-guided fabrication) for generating high-quality data for testing, development, and training. Fabrication of synthetic data consists of two stages - data modelling and data generation. Furthermore, data modelling comprises resources and structure definitions, constraint rules definitions and fabrication configuration definitions. Input and output resources are standard relational databases (e.g., DB2, Oracle, PostgreSQL, SQLite), standard file formats (e.g., Flat file, XLS, CSV, XML, JSON) and streaming via MQTT protocol.

In rule guided fabrication, the database logic is extracted automatically and is augmented by application logic and testing logic modelled by the user. The application logic and the testing logic can be modelled using rules that the platform provides, but the users can also add new rules. Once the user requests the generation of a certain amount of data into a set of test databases, the platform internally ensures that the generated data satisfies the modelled rules as well as the internal databases consistency requirements. The platform can generate data from scratch, inflate existing databases, move existing data, and transform data from previously existing resources, such as old test databases or even

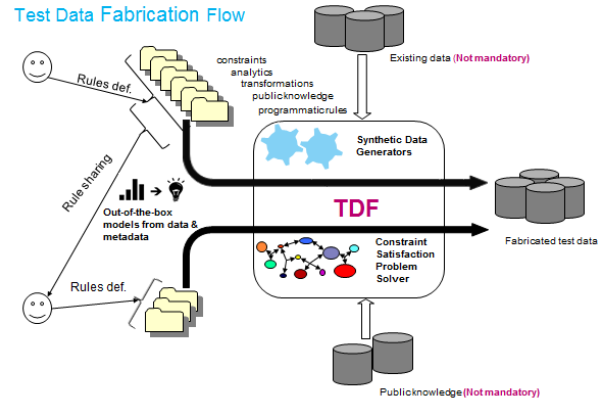


Figure 2. Flow of Generating Fabricated Data

production data. The platform provides a comprehensive and hybrid solution that can create a mixture of synthetic and real data according to user requirements.

To overcome the shortcomings of existing data generation techniques, DFP generates data using a proprietary Constraint Satisfaction Problem (CPS) solver (See Figure 2). This methodology is generic and does not require access to real data, making it very safe to use in our setting. Data fabrication consists of the following steps.

- 1) The user defines a data project which contains the structure of the data, the constraint rules and the fabrication configuration. In order to construct a constraint satisfaction problem for the solver, the platform analyses the table metadata to get the desired properties (columns data types, referential integrity constraints etc.).
- 2) The platform then selects a subset of the relevant rules and tables using the fabrication configuration, with possible addition of relevant parent tables and some default rules (e.g. PK and Unique Column). This information is used for the construction of a database table dependency graph. For each table in that graph, starting at root nodes, structural record dependencies are built recursively.
- 3) Based on the dependency graph, the fabrication pattern is computed where each target table record is assigned to one of the following fabrication modes: New, Reuse or Other. Given the patterns, the graph and the rules, a CSP problem can be created. The problem consists of variables and rules, and a solution is an assignment of values to variables that satisfies the rules.
- 4) Finally, the CSP problem is submitted to the solver, which produces a desired number of solutions to the problem and stores them in the appropriate places (e.g. database, file or stream).

### C. Distributed Privacy-Preserving Data Analytics

Machine learning methods such as deep neural networks have delivered remarkable results in data-analytics for a wide range of application domains, including healthcare. However, their training requires large data-sets which might be containing sensitive information that need to be protected from model inversion attack [13] and adversaries with access to model parameters and knowledge of the training procedure. This problem is addressed within the framework of differential privacy [2], [24]. Machine learning algorithms typically operate on data in the form of a matrix where e.g. rows correspond to features and columns correspond to samples. The particular problem in the context of matrix-valued data is to protect a machine learning algorithm, under differential privacy framework, from an adversary who seeks to gain an information about the data from an algorithm's output by perturbing the value in an element of the training data matrix. Despite the fact that random noise adding mechanism has been widely studied in privacy-preserving machine learning, there remains the challenge of studying privacy-utility trade-off for matrix-valued query functions. Our recent work [17] has suggested a novel entropy based approach for resolving the privacy-utility trade-off for real-valued data matrices. The study in [17] mathematically derives the probability density function of noise that minimizes the expected noise magnitude together with satisfying sufficient conditions for  $(\epsilon, \delta)$ -differential privacy.

1) *An Optimal  $(\epsilon, \delta)$ -Differentially Private Noise for Real-Valued Matrices:* Consider a data-set consisting of  $N$  number of samples with each sample having  $p$  number of attributes represented by a matrix  $Y \in \mathbb{R}^{p \times N}$ . A given machine learning algorithm, training a model using data matrix  $Y$ , can be represented by a mapping,  $\mathcal{A} : \mathbb{R}^{p \times N} \rightarrow \mathbf{M}$ , where  $\mathbf{M}$  is the model space.

*Definition 1 (A Private Algorithm):* Let  $\mathcal{A}^+ : \mathbb{R}^{p \times N} \rightarrow \text{Range}(\mathcal{A}^+)$  be a mapping defined as

$$\mathcal{A}^+(Y) = \mathcal{A}(Y + V), \quad V \in \mathbb{R}^{p \times N} \quad (1)$$

where  $V$  is a random noise matrix with  $f_{v_j^i}(v)$  being the probability density function of its  $(j, i)$ -th element  $v_j^i$ ;  $v_j^i$  and  $v_j^{i'}$  are independent from each other for  $i \neq i'$ ; and  $\mathcal{A}(\cdot)$  is a given mapping representing a machine learning algorithm.

*Definition 2 ( $d$ -Adjacency for Data Matrices):* Two matrices  $Y, Y' \in \mathbb{R}^{p \times N}$  are  $d$ -adjacent if for a given  $d \in \mathbb{R}_+$ , there exist  $i_0 \in \{1, 2, \dots, N\}$  and  $j_0 \in \{1, 2, \dots, p\}$  such that  $\forall i \in \{1, 2, \dots, N\}, j \in \{1, 2, \dots, p\}$ ,

$$|y_j^i - y_j^{i'}| \leq \begin{cases} d, & \text{if } i = i_0, j = j_0 \\ 0, & \text{otherwise} \end{cases}$$

where  $y_j^i$  and  $y_j^{i'}$  denote the  $(j, i)$ -th element of  $Y$  and  $Y'$  respectively. Thus,  $Y$  and  $Y'$  differ by only one element and the magnitude of the difference is upper bounded by  $d$ .

*Definition 3 ( $(\epsilon, \delta)$ -Differential Privacy for  $\mathcal{A}^+$ ):* The algorithm  $\mathcal{A}^+(Y)$  is  $(\epsilon, \delta)$ -differentially private if

$$\Pr\{\mathcal{A}^+(Y) \in \mathcal{O}\} \leq \exp(\epsilon) \Pr\{\mathcal{A}^+(Y') \in \mathcal{O}\} + \delta \quad (2)$$

for any measurable set  $\mathcal{O} \subseteq \text{Range}(\mathcal{A}^+)$  and for  $d$ -adjacent matrices pair  $(Y, Y')$ . Here,  $\Pr\{\cdot\}$  is the probability taken over the randomness used by algorithm.

*Result 1 (An Optimal  $(\epsilon, \delta)$ -Differentially Private Noise):* The probability density function of noise that minimise the expected noise magnitude together with satisfying the sufficient conditions for  $(\epsilon, \delta)$ -differential privacy of  $\mathcal{A}^+$  is given as

$$f_{v_j^i}^*(v) = \begin{cases} \delta \text{Dirac}\delta(v), & v = 0 \\ (1 - \delta) \frac{\epsilon}{2d} \exp(-\frac{\epsilon}{d}|v|), & v \in \mathbb{R} \setminus \{0\} \end{cases} \quad (3)$$

where  $\text{Dirac}\delta(v)$  is Dirac delta function satisfying  $\int_{-\infty}^{\infty} \text{Dirac}\delta(v) dv = 1$ . The optimal value of expected noise magnitude is given as

$$E_{f_{v_j^i}^*} [|v|] = (1 - \delta) \frac{d}{\epsilon}. \quad (4)$$

*Proof:* The proof follows from [17]. ■

2) *Differentially Private Distributed Deep Learning:* The post-processing invariant property [10] of differential privacy allows one to compose a global private deep model from local private deep models.

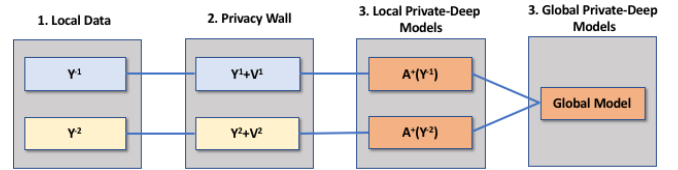


Figure 3. A structural representation of the differentially private distributed learning for deep models.

The distributed form of differentially private deep learning is represented in Fig. 3 where a privacy wall is inserted between training data and the globally shared data. The privacy wall uses noise adding mechanisms to attain differential privacy for each participant's private training data. Therefore, the adversaries have no direct access to the training data.

### D. Flexible User Authentication

The SERUMS user authentication scheme will go beyond traditional "one-size-fits-all" practices towards adopting a personalised and adaptable multi-factor user authentication scheme which will be based on a flexible authentication paradigm, coined FlexPass [5], [8], [14]. A first conceptual design of the proposed flexible user authentication paradigm is depicted in Figure 4. Our approach attempts to provide a new user authentication paradigm that leverages upon theories in Cognitive Psychology (dual coding, episodic and

semantic memory), which suggest that humans’ episodic and semantic memories, represented as verbal and visual information, can be transformed into memorable and personal authentication secrets. Such secrets can be semantically similarly reflected on both textual and graphical password keys, and accordingly used complimentary based on user preference (Figure 4) [5]. The paradigm relies on a single, open-ended, user-selected secret that can be reflected as a textual key and a graphical key.

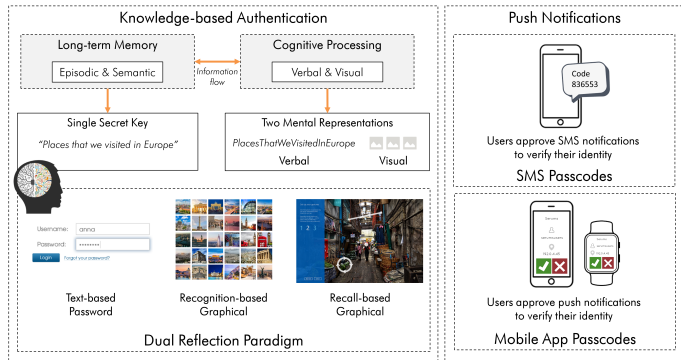


Figure 4. Conceptual design of the Flexible User Authentication Paradigm

The FlexPass paradigm extends existing works in knowledge-based user authentication based on theories of human cognition with the aim: a) to enhance memorability through ownership, and prior experience and knowledge of each single user; and b) to support user authentication adaptability since users can choose their preferred way to login based on their needs and context of use. For example, users that are on the move and interact on their smartphone might prefer to login with a graphical password, instead of entering text on a virtual keyboard which is considered a demanding and time-consuming task [26]. The same user however, in a different context, e.g., while at home working on the desktop computer, can choose to login through his textual password key. Note that in both cases, the user is only required to recall the same single secret, which can be reflected differently based on the users preference. Similarly, older adults might prefer to always login with a graphical password since they find it easier than textual passwords, as opposed to younger adults that instead, prefer traditional textual passwords [22].

Nevertheless, the dual nature of FlexPass embraces new security vulnerabilities that need to be addressed, i.e., it introduces a new observational attack since adversaries can see the set of pictures (user-selected and decoy images) during login. A brute-force algorithm could use such information from the graphical representation to guess the secret. Aiming to add an additional layer of security, we use a second factor for authentication through push notifications as a first step before proceeding to login. In particular, at a first stage users will be required to approve a push notification

that is realised as an SMS notification including an OTP, and a mobile application notification. After verifying their identity, users will login through their preferred user authentication type based on the FlexPass paradigm. Furthermore, the open-ended nature of the paradigm might affect users towards misuse strategies. To assure that users will not create semantically insecure (predictable) grids of images, automated image tagging technologies and policies need to be investigated to prevent users unsafe coping strategies.

### E. Blockchain

Blockchain is a programmable, distributed ledger with an immutable history of transactions. For every transaction consensus has to be reached among the participating organisations (or commonly denoted as nodes) before it can be written on the ledger. Blockchain is programmable via the notion of a smart contract that is simply a piece of code, that is installed and executed within the Blockchain network; the execution of a smart contract’s function creates a transaction. Note that the transaction is written on the ledger of each node concurrently. Consequently, the ledgers are always synchronised. If a node has some downtime, when it restarts, it automatically synchronises its ledger to the ledgers of the rest of the nodes. In addition, a single ledger (of a single node) cannot be tampered unless the attacker can manage to concurrently infiltrate at least the majority (if not all) of the nodes, depending on the consensus protocol used.

In the proposed architecture a blockchain network is created where every relevant organisation (e.g a hospital) participates. The user’s permissions that control access for the SPHR are programmed using smart contracts. This allows versatility as the rules used to form the permissions can be updated whenever required. However, due to the Blockchain’s nature, a single organisation cannot force an update of these rules as transactions will not be able to reach consensus and inevitably will not be written on the ledger. The process flow for setting up access control is shown in Figure 5. The medical organisation (e.g. hospital) creates generic smart contracts (access rules) and stores them in the blockchain (step 1). The Patient also creates custom smart contracts about their data, which are also stored in the blockchain (step 2). The patient’s ID is shared with the doctor (step 3), who then authenticate themselves to the system (step 4). The Doctor requests access to data about a patient (step 5). The IDs of the doctor and the patient are checked against the access rules in the blockchain (check/audit trail). This results in a request for an access token from the data vault (step 6). The data vault provides the access token (step 7) and the response with this token is sent to the doctor (step 8). The doctor requests data about the patient from the data vault using the token (step 9) and the data is afterwards fetched from the vault (step 10).

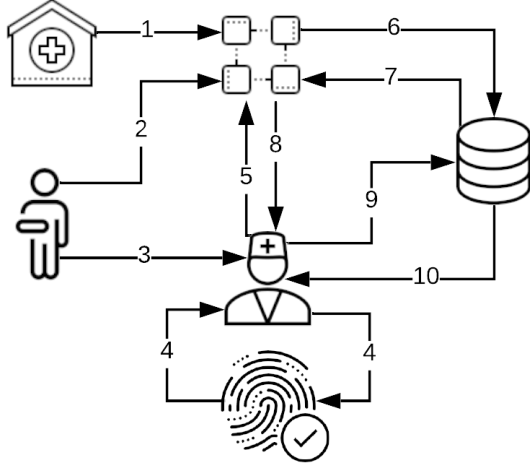


Figure 5. Process flow for access request

## V. EVALUATION

We present an initial evaluation of the SERUMS technologies presented in Section IV on a use case based on the Edinburgh Cancer Data Gateway (ECDG).

### A. Use Case - Edinburgh Cancer Data Gateway

We are developing a dashboard to help oncologists observe, monitor, and analyse the condition of their patients over time. It can also be used to analyse the effect of different chemotherapy treatments when given to patients with similar characteristics, and consequently influence future decisions to improve the well-being and survival rate of patients. Our ultimate aim is to build a toxicity predictor (Figure 6) to predict the toxicity of chemotherapy treatments based on history and feedback from patients. Figure 7 shows the data structure we use for training the toxicity predictor. We extracted data for training the machine learning models from three main databases (i.e., Chemocare, Trak, and Oncology DB) within the Edinburgh Cancer Centre (ECC). The data contains the information on treatment cycles, recorded side effects (here, toxicity level), comorbidities, and various observations concerning breast cancer patients for three years (from 2014 to 2016). The extraction has data for 51,661 treatments, of which 13,030 are breast cancer treatments. There are 933 unique patients, and some patients may have two or three different treatments/regimes. Each regime has several cycles ranging from one to more than 50 cycles.

### B. Smart Patient Records

The data from the Edinburgh Cancer Gateway (see Table I), is abstracted out into their data vault structure. For example, the original form of NDC\_SMR01 is given in Figure 8. Each of the columns is examined and classified under one of the hubs of the TPOLE data vault:

- Time: admission\_date, discharge\_date, length\_of\_stay;
- Person: sex, age\_in\_years, ethnic\_group, marital\_status, postcode; or

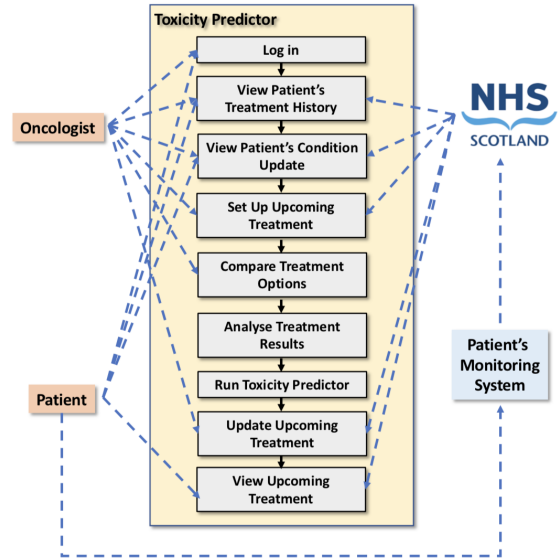


Figure 6. Toxicity Predictor for Breast Cancer Treatment

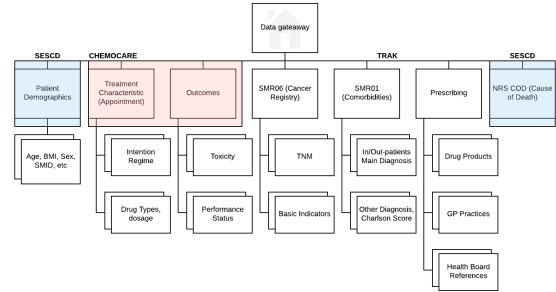


Figure 7. Database Structure for Training the Toxicity Predictor Model

Table name	# vars	# num	# categorical	# bool
NDC_SMR01	17	3	13	1
NDC_SMR06	9	2	7	0
NDC_Charlson	20	9	5	6
Chemocare_Toxicity	17	14	3	0
Chemocare_Treatment	19	8	11	0

Table I  
DATABASE TABLES STRUCTURE FROM THE EDINBURGH CANCER GATEWAY USE CASE.

- Object: main\_operation\_a, main\_operation\_b, main\_condition, other\_condition\_1, other\_condition\_2, other\_condition\_3, other\_condition\_4.

These are then broken up into smaller subcategories which will form the satellites of the data vault. In this example, the Object category can be seen to be made up of two sets: one containing details about the operations, and the other containing details about the conditions.


















public.ndc_smr01	
 chi	int4
 admission_date	date
 discharge_date	smallint
 length_of_stay	smallint
 sex	smallint
 age_in_years	smallint
 ethnic_group	char(2)
 marital_status	char(1)
 postcode	char(8)
 main_condition	char(10)
 other_condition_1	char(10)
 other_condition_2	char(10)
 other_condition_3	char(10)
 other_condition_4	char(10)
 main_operation_a	char(10)
 main_operation_b	char(10)
 ndc_smr01_pk	constraint « pk »

Figure 8. Example of source table

### C. Data Fabrication

In order to synthesise data, we must pass database table definitions and metadata to the DFP. The metadata itself contains high level information about the data, describing details about its nature, without revealing any of the actual values that make up the source data. In addition, we need to define the rules that the data conforms to, in order to keep the synthetic data as accurate as possible. This might include, for example, the range of values that the data takes and the distribution of these values, as well as any relationships between different data elements. For instance, we might have a column with the appointment date. The metadata would contain the information that it is a date type, the format the date should be stored in, whether it can be null, etc. The rules for it might include that it must be greater than the date of birth for the patient.

For the Edinburgh Cancer Gateway use case, we have collected many aspects of the metadata including common, maximum, minimum, and extreme values for each reading. In addition, we derived the distribution of the data value measurements and the correlations between the different values. This profiling can be seen to work to derive the required rules to fabricate new data. These rules were then used to synthesise data to be used in the development and evaluation of the SERUMS tool chain.

### D. Blockchain

The blockchain smart contracts will use the hyper-ledger format and will enable the storage of the preference contract of the patient, the vault of the current active data transport contracts and the valid user contracts of the SERUMS data exchange process.

### E. Authentication

The dual nature of the proposed user authentication scheme allows us to move from "one-size-fits-all" authentication schemes to flexible authentication schemes since users

can choose their preferred way to authenticate; either by entering the textual password or the graphical password that represents their single secret. Consider a password creation scenario in which a user chooses a secret derived from his episodic memory, e.g., Places that we visited in Europe. In this scenario, the textual password key is based on the articulation of the secret, e.g., the system will generate a textual password key PlacesThatWeVisitedInEurope. For the creation of the graphical password key, the user chooses pictures illustrating relevant images through search in Web engines. Other related images from the image search default to decoy images (in the case of recognition-based graphical authentication). Both user-selected and decoy images are finally assigned to the users profile to be used for login. Users will also be able to choose a single background image and then draw secret gestures on the image that will be based on the chosen single secret.

A preliminary evaluation study with 32 volunteers (age ranging 20-49 (m=33.84; sd=9.43) has been conducted to investigate likeability aspects and user acceptance of the proposed paradigm. More details on the prototype designs of FlexPass and evaluation results are reported in [5]. Participants interacted with initial prototypes of FlexPass and rated their experience using a 5-point Likert scale (1: Not at all 5: Absolutely). Example statements included: I would adopt FlexPass as my main authentication method, FlexPass login is fast to use, "Long registration time is bad", etc. Initial evaluation results are promising for further development of the proposed paradigm since most of the participants are positive to adopt FlexPass as their main authentication method and they particularly like the flexibility of switching between textual and pictorial passwords (81.25%). Furthermore, participants rated FlexPass login process as memorable (87.5%), easy to use (84.37%), and efficient to use (68.75%). Nevertheless, given that the new paradigm adds an additional amount of time in the secret creation process compared to the current state-of-the-art approach, participants had mixed opinions with regards to the higher password creation times (during registration). In particular, 53.13% participants stated that the higher registration times might negatively affect their opinion about FlexPass, and 21.87% rated that long registration times might prevent them from using FlexPass.

### F. Noise Adding Mechanism for Differential Privacy

The optimal noise adding mechanism to attain differential privacy is compared with the classical Gaussian mechanism via quantify the gain (over Gaussian mechanism) achieved by optimal  $(\epsilon, \delta)$ -differentially private noise in term of reduction in expected noise magnitude. The ratio of expected noise magnitude of classical Gaussian mechanism to that of optimal mechanism is calculated as

$$R(\delta) = \frac{2}{(1-\delta)\sqrt{\pi}} \sqrt{\log(1.25/\delta)}. \quad (5)$$



It is observed in Fig. 9 that noise magnitude reduction factor is increasingly more pronounced in the high privacy regime (i.e. low  $\delta$ ), however, also shoots up in the low privacy regime as  $\delta \rightarrow 1$ . The optimal mechanism reduces the noise magnitude by more than 4 times in the high privacy regime over the Gaussian mechanism.

## VI. RELATED WORK

*Smart Patient Records:* Unified Medical Language System (UMLS [7]) proposes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. OpenEHR Specification Program [1] provides specifications and their computable expressions to enable development and deployment of open, interoperable and computable patient-centric health information systems.

*Generating Synthetic Data:* Several studies address generating data for given queries. Most of these approaches (e.g. QAGen [6], De La Riva et al. [9] and Emmi et al. [11]) address only subsets of the SQL language as well as a simple subset of the possible data types of databases. Many of these works have performance and scalability issues as well. Adorf and Varendorff [3] propose a scalable solution that generates data for form-centric applications using an SMT solver. However, constraint solvers cannot deal with the variety of data types, such as decimal numbers, calendar types, and strings, this is also not an ideal solution and requires workarounds that increase complexity of the overall system and affect performance and quality of results.

*Authentication:* Recent works have investigated the influence of specific human, technology and design factors affecting user authentication preference and task performance, aiming to apply that knowledge in designing usable and personalised authentication schemes. Nicholson et al. [22] suggested personalizing the user authentication type based on age differences; Belk et al. [4] proposed an extensible authentication framework for personalizing authentication tasks based on the users' cognitive processing styles and abilities; Ma et al. [21] suggested personalizing user authentication types by considering users' cognitive disabilities; and Forget et al. [12] proposed an authentication scheme for enabling users to choose the preferred user authentication mechanism instead of providing a single authentication type.

*Privacy of Medical Analytics:* Shade [15], a framework for Apache Spark that provides strong privacy guarantees for users, includes two mechanisms - SparkLAP, which provides Laplacian perturbation based on a user's query and SparkSAM, which uses the contents of the database itself in order to calculate the perturbation. Palanisami et al. [23] present a privacy-aware data disclosure scheme that considers group privacy requirements of individuals in bipartite association graph datasets where even aggregate information about groups of individuals may be sensitive.

## VII. CONCLUSION

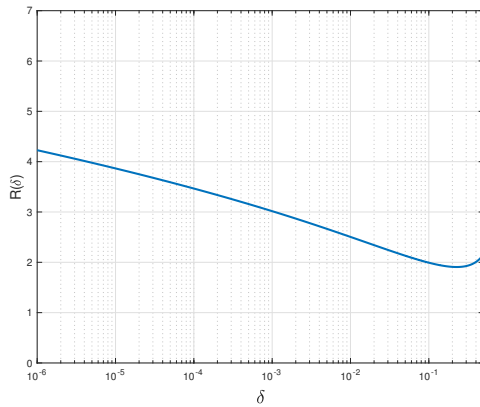
In this paper, we have outlined the problems that the distributed health systems of the future will face in terms of safe storing and sharing of confidential patient data. We have also proposed the SERUMS methodology for managing confidential, distributed medical data, covering all the phases in its lifetime, from retrieval and storing to end-point data analytics. Furthermore, we have described the initial versions of the tools from the SERUMS tool-chain, including new universal smart patient record format, blockchain for controlling access to the health records and recording lineage of the data, authentication mechanisms for logging in to healthcare systems and privacy-preserving data analytics techniques. We have also described Data Fabrication Platform (DFP), a platform for generating large volumes of synthetic but realistic medical data that will be used for development and evaluation of the SERUMS tool-chain. Finally, we have described its proposed use in the Edinburgh Cancer Gateway use case that collects and analyses information about effects of chemotherapy treatments on breast cancer patients, to predict the outcome of the treatment and improve treatment.

## ACKNOWLEDGEMENT

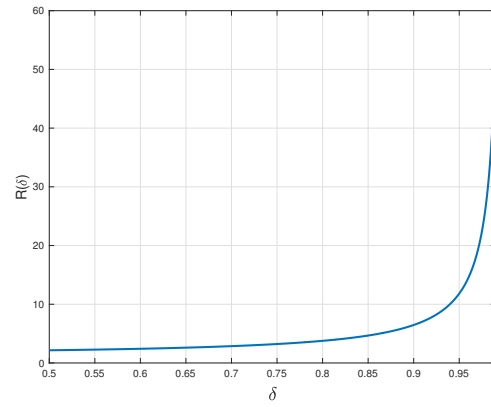
This research was funded by the EU H2020 project Serums: Securing Medical Data in Smart Patient-Centric Healthcare Systems (grant code: 826278).

## REFERENCES

- [1] OpenEHR Specification Program. <https://www.openehr.org/programs/specification/>.
- [2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proc. CCS '16*, pages 308–318. ACM, 2016.
- [3] H.-M. Adorf and M. Varendorff. Constraint-based automated generation of test data. In D. Winkler, S. Biffl, and J. Bergs-mann, editors, *Software Quality. Model-Based Approaches for Advanced Software and Systems Engineering*, pages 199–213, Cham, 2014. Springer International Publishing.
- [4] M. Belk, C. Fidas, P. Germanakos, and G. Samaras. The interplay between humans, technology and user authentication. *Comput. Hum. Behav.*, 76(C):184–200, Nov. 2017.
- [5] M. Belk, C. Fidas, and A. Pitsillides. Flexpass: Symbiosis of seamless user authentication schemes in iot. In *Extended Abstracts of the CHI 2019*.
- [6] C. Binnig, D. Kossman, E. Lo, and M. T. Özsu. Qagen: Generating query-aware test databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07*, pages 341–352, New York, NY, USA, 2007. ACM.
- [7] O. Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res.* 2004 Jan 1;32(Database issue), 2004.



(a) High privacy regime.



(b) Low privacy regime.

Figure 9. Ratio of expected noise magnitude of the classical Gaussian mechanism to that of optimal mechanism for  $(\epsilon, \delta)$ -differential privacy.

- [8] A. Constantinides, M. Belk, C. Fidas, and A. Pitsillides. On the accuracy of eye gaze-driven classifiers for predicting image content familiarity in graphical passwords. In *Proc. UMAP 2019*, pages 201–205. ACM, 2019.
- [9] C. de la Riva, M. J. Suárez-Cabal, and J. Tuya. Constraint-based test database generation for sql queries. In *Proceedings of the 5th Workshop on Automation of Software Test, AST '10*, pages 67–74, New York, NY, USA, 2010. ACM.
- [10] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
- [11] M. Emmi, R. Majumdar, and K. Sen. Dynamic test input generation for database applications. In *Proc. ISSTA '07*, pages 151–162. ACM, 2007.
- [12] A. Forget, S. Chiasson, and R. Biddle. [paper] choose your own authentication. In *New Security Paradigm Workshop (NSPW)*. ACM, 2015. Conference Papers.
- [13] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. CCS '15*, pages 1322–1333. ACM, 2015.
- [14] G. Hadjidemetriou, M. Belk, C. Fidas, and A. Pitsillides. Picture passwords in mixed reality: Implementation and evaluation. In *Extended Abstracts of the CHI 2019*, 2019.
- [15] A. Heifetz, V. Mugunthan, and L. Kagal. Shade: A differentially-private wrapper for enterprise big data. In *Proc. IEEE Big Data 2017*, pages 1033–1042, 2017.
- [16] W. Inmon and D. Linstedt. Introduction to Data Vault. In *Data Architecture: a Primer for the Data Scientist*. 2015.
- [17] M. Kumar, M. Rossbory, B. A. Moser, and B. Freudenthaler. Deriving an optimal noise adding mechanism for privacy-preserving machine learning. In G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil, editors, *Database and Expert Systems Applications*, pages 108–118. Springer International Publishing, 2019.
- [18] S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, and T. Cheng. Geospatial big data handling theory and methods: A review and research challenges, 2016.
- [19] D. Linstedt. Data Vault 2.0 Being Announced, 2012.
- [20] D. Linstedt and M. Olschimke. Scalable Data Warehouse Architecture. In *Data Vault 2.0*. 2016.
- [21] Y. Ma, J. Feng, L. Kumin, and J. Lazar. Investigating user behavior for authentication methods: A comparison between individuals with down syndrome and neurotypical users. *ACM Trans. Access. Comput.*, 4(4):15:1–15:27, July 2013.
- [22] J. Nicholson, L. Coventry, and P. Briggs. Age-related performance issues for pin and face-based authentication systems. In *Proc. CHI '13*, pages 323–332. ACM, 2013.
- [23] B. Palanisamy, C. Li, and P. Krishnamurthy. In *Proc. IEEE Big Data 2017*, pages 1043–1052, 2017.
- [24] N. Phan, Y. Wang, X. Wu, and D. Dou. Differential privacy preservation for deep auto-encoders: An application of human behavior prediction. In *Proc. AAAI '16*, pages 1309–1316. AAAI Press, 2016.
- [25] S. Silow-Carroll, J. N. Edwards, and D. Rodin. Using electronic health records to improve quality and efficiency: the experiences of leading hospitals. *Issue brief (Commonwealth Fund)*, 2012.
- [26] E. von Zezschwitz, A. De Luca, and H. Hussmann. Honey, i shrunk the keys: Influences of mobile devices on password composition and authentication performance. In *Proc. NordiCHI '14*, pages 461–470. ACM, 2014.