Privacy Preserving Interoperability for eHealth Systems

Alevtina Dubovitskaya
LSIR, I&C, EPFL
AISLab, HES-SO Valais

Abstract—Development and implementation of convenient, miniaturized and integrated instruments that enable therapeutic drug monitoring and dosage individualization for the patients will significantly improve modern therapeutics. To make it possible we aim at tackling the following questions: how to achieve (semantic) interoperability and medical data integration in a distributed environment; and how to ensure privacy of the patients in case of reusing medical data for providing personalized recommendations that enhance patient care.

Index Terms—Access Control, eHealth, Electronic Medical Record, Interoperability, Point-of-Care System, Privacy

I. INTRODUCTION

The treatment of certain diseases such as cancer, HIV, or other serious medical conditions relies on a regular administration of critical drugs that are absolutely required to keep those life-threatening diseases under control. Constant monitoring allows precise personalization of the treatment. This can improve patient care in particular for drugs with a narrow effective therapeutic range, such as Efavirenz, used for the treatment of HIV. Another example is the drugs that are used in achieving anemia correction. For those drugs the relationship between the dose prescribed to the patient and the concentration of the drug in the blood is poorly predictable and highly variable among individuals.

At present, concentration measurements are performed in large medical laboratories. The results are sent to medical doctors in charge of the patients. Unfortunately, clinical pharmacologists are not largely available, and for general practitioners it is a true challenge to interpret and translate mere drug concentration values to an appropriate decision of dosage adjustment for the patient.

Technological advances during recent years make it possible to envisage a new approach to perform drug concentration measurements and translate these values into personalized treatment advices. A personal monitoring instrument, a portable Point-Of-Care (POC) stand-alone system, will fulfill the following main objectives:

1) Perform measurements of the drug concentration in blood samples.
2) Provide the medical doctor with all necessary information about the patient.
3) Collect drug usage and measurement data into a remote database, enabling further refinements in dosage adjustment procedures.

A system like this is the goal of the ISyPeM2 project—a continuation of the Nano-Tera project: Intelligent Integrated Systems for Personalized Medicine (ISyPeM).

A prototype of a POC system aims at assisting medical doctors with the interpretation of drug concentration values and conducting clinical studies. It will make a critical step towards personalized medicine, and demonstrate the feasibility of performing Therapeutic Drug Monitoring (TDM) using portable devices while decreasing costs and improving treatment efficacy.

Building such a system and integrating it in the eHealth environment requires the combination of know-how in a variety of different fields, from molecular biology and biomedical microtechnology to mathematical modeling and information technologies. From the perspective of data management the following challenges need to be overcome to achieve the goals listed above:

- Distributed data processing There is a need to ensure interoperability to allow different entities (such as medical doctors, patients, clinical institutions) to access the required data about particular patient. The necessity to access the data from different locations and using different
A new approach to therapeutic drug concentration monitoring for personalized medicine.

In the next three sections we discuss the aforementioned works in order to gain valuable insights from. First, in Section II we survey the work of Buyya et al. [1], in which the authors present a way to achieve interoperability by proposing a cloud federation framework—InterCloud. In Section III we pass through the different aspects of use of patient’s data when conducting medical research presented by Elger et al. in [2]. Then, in Section IV we move to the first approach to anonymization of RT-datasets proposed by Poulis et al. in [3]. Finally, we conclude with our research proposal in Section V.

II. INTERCLOUD: UTILITY-ORIENTED FEDERATION OF CLOUD COMPUTING ENVIRONMENTS FOR SCALING OF APPLICATION SERVICES

Buyya et al. suggest a way to achieve interoperability by creating a federated cloud computing environment. They present the vision and challenges of this approach, and the architectural elements of InterCloud—the utility-oriented framework proposed in the paper.

A. Why do we need a cloud federation?

Cloud computing becomes more and more popular. It enables the delivery of infrastructure, platform and software as a service to the users. This has the following advantages: consumers need to pay providers only while accessing the services; there is no need to invest a lot of money and efforts in building and maintaining a complex IT infrastructure, and access to the services is provided without regards to where the services are hosted.

However in their work [1] Buyya et al. also pointed out several shortcomings of this approach, including the following: difficulties in determining upfront the best location for hosting services (which is expected from the customers) and meeting the quality of service (QoS) expectations of consumers originating from multiple geographic locations (from the side of cloud providers.) In addition, no single cloud provider will be able to establish its data centers at all possible locations through the world in order to meet the QoS expectations of all the consumers. Hence, there is a need to build mechanisms for a federation of cloud infrastructure service providers.

B. Research Issues

To exploit the potential of a federated cloud infrastructure it is crucial to overcome several challenges.

- Application service behavior prediction: The system has to be able to dynamically scale services up and down over federated cloud infrastructures. The challenge is to fit statistical models to the observed service behaviors.
- Flexible mapping of services to resources: Operating costs and energy requirements of composite systems, makes appropriate mapping of services to resources critical. It is important to maximize system efficiency and utilization without affecting QoS targets. Mapping is aggravated by the stochastic behavior of resources and services.

---

1 e.g., EC Data Protection Directive 95/46/EC in Europe. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) in US
2 e.g., HL7 CDA, a standard supporting message-based information exchange of medical data. http://www.hl7.org
• **Optimization techniques for market-driven decision making problems:** There is a need to develop optimization models that ultimately aim to optimize both resource-centric (utilization, availability, reliability) and user-centric (response time, budget spent, fairness) QoS targets.

• **Integration and interoperability:** Privacy and security issues create barriers for migration of sensitive data into the Cloud. There also exist many SMEs that use on-premise IT assets. Therefore there are issues related to interoperability between on-premise software and cloud services, in particular, identity and data management and business process orchestration.

• **Scalable monitoring of system components:** Existing centralized approaches to overall system monitoring is not an appropriate solution in the case of the distributed components of the federated system due to concerns of scalability, performance, and reliability. Therefore, Buyya et al. advocate for architecting service monitoring and management services based on decentralized messaging and indexing models.

### C. Architecture of InterCloud

To address the challenges described above the authors propose an articulated model for constructing a federated Cloud infrastructure. Figure 1 shows the high level components of the framework consisting of clients brokering and coordinator services supporting the federation of clouds. They perform tasks such as application scheduling, resource allocation and migration of workloads.

Every consumer instantiates a Cloud Broker to express his/her needs. Cloud Coordinators publish their services to the federation. The Cloud Exchange acts as a mediator bringing together service providers and customers. It aggregates the infrastructure demands from the application brokers and evaluates them by supplying available resources published by the Cloud Coordinators.

- **Cloud Coordinator (CC):** The Cloud Coordinator service is responsible for the management of enterprise clouds inside the domain and their membership to the overall federation. It provides a programming, management, and deployment environment for applications in the federation. A CC exports the services of a cloud to the federation by implementing basic functionalities for resource management such as scheduling, allocation, workload and performance models, dynamic sensing/monitoring, and discovery.

- **Cloud Broker (CB):** the Cloud Broker identifies suitable cloud service providers through the Cloud Exchange and negotiates with Cloud Coordinators for an allocation of resources that meets the QoS needs of the consumers. The architecture of Cloud Broker consists of the following components: User Interface, Core Services, Execution Interface, and Persistence. For example: **Core Services** enable the main functionality of the broker by determining the most appropriate cloud services for the user application and maintaining the status of cloud services.

- **Cloud Exchange (CEx):** The CEx acts as an information registry to store the costs and demand patterns. As a market maker, the Cloud Exchange provides directory services, dynamic bidding-based service clearance, and payment management services. Cloud Coordinators periodically update their availability, pricing, and Service-Level Agreement (SLA) policies with the CEx. Cloud Brokers query the registry to learn information about existing SLA offers and resource availability of member Clouds in the federation. The CEx can make use of economic models to provide match-making services.

### D. Discussion

The InterCloud [1] project is one of the first initiatives in the field of centralized federating of clouds. Buyya et al. formulate the challenges in this field and present a framework to address them. The problems faced while building a federation of clouds and constructing an eHealth architecture have many similarities. Indeed, enabling access to dynamic data, distributed between multiple sources, preserving QoS and overcoming interoperability issues are crucial for building a reliable eHealth system.

One of the components of the framework proposed in [1] is Cloud Exchange. Acting as a mediator, it makes possible information exchange and dynamic interconnections between different entities, which is coherent with our goals. This approach will serve us in our work towards overcoming an interoperability issue.

In [1], Buyya et al. do not present an implementation of the framework. Nevertheless, the research vision helps us to understand the theoretical and practical problems of distributed and heterogeneous environments. The architectural design principles proposed by the authors can lay a foundation for building an architecture for eHealth systems.

### III. Strategies for Health Data Exchange for Secondary, Cross-Institutional Clinical Research

Elger et al. [2] give an overview of technical, practical, legal, and ethical aspects of secondary data use and discuss their
implementation in the multi-institutional @neurIST project.\(^3\)

A. Terminology clarification

An electronic health record (EHR) refers to the concept of a comprehensive, cross-institutional, and longitudinal collection of a patient's health and healthcare data. In the first instance information from EHR is used in the treatment context, providing higher quality and safer care for patients. EHRs also have many important secondary uses, such as in disease-specific clinical or epidemiological research projects, healthcare research and assessment of treatment quality [8].

In the primary use of an EHR for the treatment of a single patient, the patient’s identity is necessary and protected by medical secrecy. While the patient’s identity might not be relevant for secondary uses, removal of the patient’s explicit identifier is not enough, since the data remains personal due to its specificity to the individual. This necessitates more elaborated data protection strategy.

Ambiguous definitions of terms and contradictory guidelines developed in the area of secondary use of medical data, even within European Union, raise problems from legal, ethical, and technical standpoints. Elger et al. discuss the main laws and guidelines that describe how to prepare data for use in medical research. Based on how the concepts of personal, identified and identifiable data are addressed in the various legal documents (including the EC Data Protection Directive 95/46/EC, and the Article 29 Working Party) the authors summarize the definitions of the following terms:

- **Personal data** refers to data that is about an individual who can reasonably\(^4\) be identified or identifiable.
- **De-identification** is the process of removing (or modifying) identifiers from the personal data so identification is not reasonably possible.
- **Pseudonymization** is the step where a pseudonym or code is added to this de-identified data.
- **Proportional or reasonable anonymity** applies to de-identified/pseudonymized data which can not reasonably be used to identify specific individuals.

B. Issues related to secondary use of medical data

Digital health records should facilitate the secondary use of data for improving and advancing medical knowledge. However, we have to consider that the data contained in EHRs may be distributed, involve a range of IT systems, expand over time, and are of a personal nature to the patient. Hence, the following challenges can be encountered while building a platform for research involving secondary use of health data:

- **Ensuring Privacy of the patients**
  - Collecting informed consents to respect awareness and wishes of the patient regarding secondary usage.
  - Applying pseudonymization techniques to achieve reasonable anonymity. The authors describe pseudonymization and reversal schemes they implemented for the @neurIST project.
  - Handling sensitive data such as religion, nationality, and ethnic background.
  - Removing identifying information from free-text data contained in medical records.

- **Architecture-related considerations**

Usually each clinical center is managing the data of its patients locally. Therefore, in the case of multi-institutional research we need to rely on a distributed architecture for clinical data storage. The two following models are proposed:

- an anonymized model where the data are de-personalized and stored in a repository at the center where the data are generated;
- an on-the-fly model where the live clinical information system is chosen to store the entire data that are de-personalized prior to leaving the hospital intranet.

The first model provides more control to the sites of origin of the data and reduce the amount of data that are available to a potential intruder. However, it requires efforts from all clinical centers to implement and maintain such a model. Also, if a site is off-line the response might be filled with outdated information. The advantage of the second model is that always the latest data are in use.

- **Data quality**

Since data is collected from multiple sources, it is often inconsistent and incomplete. Though, the question of data quality is essential in the case of research that affects the patient’s care. Some procedural approaches to maximize data quality presented:

- developing an appropriate training and awareness of the personnel entering data in the clinical IT system to eliminate errors coming from digitalizing the clinical records;
- post-hoc cleansing to catch inaccuracies introduced by typographic errors;
- maintaining the continuity of an individual participant record for example by using the patient’s UID based on a hospital code.

C. Hybrid access control model

An access control system for multi-institutional research should follow common patterns and principles for distributed cross-domain or cross-border information systems, in which heterogeneous environments of multiple distinct institutions need to be interconnected. Hybrid security models consistent with the health application domain proposed by Elger et al. in combine the combination of local and distributed models across various security domains. In other words, within a security domain all the security is concentrated and placed under the responsibility of this domain whereas between different security domains local credentials are mapped to inter-domain credentials that can be exchanged.
The approach proposed in [2] is to create a designated security entity (the Security Token Service, STS) in each domain that will be in charge of issuing and verifying short-term security tokens with the entities of the other security domains. The client requests a security claim from the local STS using mechanisms defined in the WS-Trust standard specification. This request contains the originator’s username and the desired relationship for the claim. The STS retrieves which attributes the user was assigned to (for the given relationship) and issues a SAML token. Finally, the token is signed by the STS.

D. Discussion

Secondary use of medical data can significantly enhance healthcare experiences for individuals [11]. Especially for rare diseases, conducting multi-center studies are advantageous, as interaction between different medical institutions enables to obtain enough statistical power quicker than studies in a single center. Elger et al. give a cohesive overview of the issues surrounding the secondary use of health data originating from multiple clinical centers based on a real-life example. The contribution presented in [2] can serve as a point of reference for the other works in the field.

The pseudonymization approach and a high-level hybrid access-control system that partially address privacy issues are presented in this paper. The authors also list security vulnerabilities, including the possibility of cracking the proposed pseudonym generation mechanism, dependence on a trusted third party and the possibility of establishing an indirect identification. However, they do not provide any solutions to these problems. Moreover, the concept of reasonable anonymity introduced by the World Health Organization “to secure genetic samples or genetic data, depending on how this is done, who is to have access and the uses which have been consented to” remains vague and leaves too much freedom in the interpretation of terms given in Subsection III–A.

Although some approaches to overcome the issues related to secondary use of medical data pointed out in Subsection B are presented, the discussion remains at a high-level. There is a lack of concrete instructions and evaluation/metrics for the suggested approaches, and the question of their implementation remains open.

IV. ANONYMIZING DATA WITH RELATIONAL AND TRANSACTION ATTRIBUTES

A. RT-datasets and their anonymity

Publishing datasets containing both relational and transaction (i.e., set-valued) attributes, RT-datasets for short, is essential in many real-world applications. Health information is one example of such datasets. Usually medical studies require analyzing patient demographics and diagnosis information together. The patients features (e.g., demographics) can be modeled as relational attributes and diagnosis as a transaction attribute.

Hereafter the formal definitions of RT-dataset, group of records $G(r)$ and $(k, k^m)$-anonymous RT-dataset are given.

- An RT-dataset $D$ is a dataset consisting of records containing relational attributes $R_1, ..., R_n$ that are single-valued, and a transaction attribute $T$ that is set-valued.
- Given a fixed number of items $m$, for each record $r \in D$ its group $G(r)$ is a set of records that contains $r$ and each record $q \in D$ such that $q[R_1, ..., R_n] = r[R_1, ..., R_n]$ and $q[T] \cap I = I$, where $I$ is any set of $m$ or fewer items of $r[T]$. $|G(r)|$, the size of $G(r)$, represents the risk of associating an individual with a record $r$.
- A group $G(r)$ is $(k, k^m)$-anonymous, if and only if $|G(r)| \geq k$ for each record $r \in G(r)$. An RT-dataset $D$ is $(k, k^m)$-anonymous, if and only if the group $G(r)$ of each record $r \in D$ is $(k, k^m)$-anonymous.

$(k, k^m)$-anonymity provides the same protection as $k$-anonymity [12] for relational attributes, and as $k^m$-anonymity [13] for the transaction attribute. In other words $(k, k^m)$-anonymity ensures that for any record $r$ in RT-dataset $D$ and any set of $m$ or less items in transaction attribute of $r$, there should be at least $(k-1)$ records that are indistinguishable from record $r$. Unfortunately, $k$-anonymity for relational attributes, and as $k^m$-anonymity for transaction attribute do not imply $(k, k^m)$-anonymity. To achieve $(k, k^m)$-anonymity authors employ generalization functions defined as follows:

- A relational generalization function $R$ maps a value $v$ in a relational attribute $R$ to a generalized value $\tilde{v}$, which is a range of values, if $R$ is numerical, or a collection of values, if $R$ is categorical.
- A transaction generalization function $T$ maps an item $u$ in the transaction attribute $T$ to a generalized item $\tilde{u}$. The generalized item $\tilde{u}$ is a non-empty subset of items in $T$ that contains $u$.

B. Problem formulation

To prevent identity disclosure (i.e., the association of patient and medical record) and attribute disclosure (i.e., the association of a patient with sensitive diagnoses) it is not enough to use the methods that preserve privacy of dataset containing only relational or only transaction attributes. Furthermore, naively applying these methods consequently to RT-datasets leads to information loss that makes meaningful analysis of both relational and transaction attributes impossible.

The authors consider two general data utility measures: Rum, for relational and Tum, for transaction attributes. Lower values of Rum and Tum imply better data utility. Rum is monotonic and Tum is anti-monotonic to subset relationships. More formally, given two groups $G$ and $G'$:

$$\text{Rum}(R(G) \cup R(G')) \leq \text{Rum}(R(G \cup G')) \quad (1)$$  

$$\text{Tum}(T(G) \cup T(G')) \geq \text{Tum}(T(G \cup G')) \quad (2)$$

In other words, data utility is preserved better while generalizing the relational values of small groups and large groups of transaction values. Any metrics that satisfied the following properties could be used as Rum and Tum.

Therefore the problem of constructing a $(k, k^m)$-anonymous version of an RT-dataset with
bounded information loss regarding relational (Problem 1) or transaction (Problem 2) attributes can be formulated as follows: given an RT-dataset \( D \), data utility measures \( \text{Run} \) and \( \text{Tum} \), parameters \( k \) and \( m \), and a threshold \( \delta \) (specified by the publisher), construct a \((k,m)\)-anonymization version of \( D \), such that \( \text{Run}(D) \leq \delta \) and \( \text{Tum}(D) \) is minimized (Problem 1) or \( \text{Tum}(D) \leq \delta \) and \( \text{Run}(D) \) is minimized (Problem 2). Solving Problem 1 or Problem 2 ensures that \( D \) preserves privacy and utility, but it is NP-hard.

### Anonymization approach

To address these problems, two anonymization frameworks, \textbf{Run-BOUND} and \textbf{Tum-BOUND} are proposed. These frameworks are based on a three-phase process:

- **Initial cluster formation**: \( k \)-anonymous clusters with respect to relational attributes, which incur low information loss, are formed.
- **Cluster merging**: the clusters are merged until the conditions set by Problem 1 or 2 are met.
- **\((k,m)\)-anonymization**: each cluster becomes \((k,m)\)-anonymous, by generalizing its items with low \textbf{Tum}.

Poulis et al. also propose a family of algorithms to implement the second phase in each framework. These algorithms operate by building clusters, which can be made \((k,m)\)-anonymous with minimal information loss, and preserve different aspects of data utility.

The function \( \text{Rmerge} \) is responsible for the merging phase of \textbf{Run-BOUND} (Step 4). It can be implemented by one of the following algorithms based on different merging heuristics:

- \textbf{Rmerge}_\text{R} merges clusters with similar relational values,
- \textbf{Rmerge}_\text{T} merges clusters with similar transaction items,
- \textbf{Rmerge}_\text{RT} takes a middle line between these two algorithms.

Similarly \textbf{Tmerge} is used at the Step 4 of \textbf{Tum-BOUND} and has 3 algorithmic realization (\textbf{Tmerge}_\text{R}, \textbf{Tmerge}_\text{T}, \textbf{Tmerge}_\text{RT}).

As concrete instances of \textbf{Run} and \textbf{Tum} authors used Normalized Certainty Penalty (NCP) [14] and Utility Loss (UL) [15], respectively. For a numerical attribute \( R \), \( NCP \) is based on the domain size, a range of generalized values and the weight that reflects the importance of the attribute. The \( \text{UL} \) is computed taking into account the number of items mapped to the generalized value \( \tilde{u} \), a weight reflecting the importance of \( \tilde{u} \), and the sum of sizes of all generalized items in the record. These data utility measures are defined for a generalized value \( \tilde{v} \) and an item \( \tilde{u} \), a record \( r \) and an RT-dataset \( D \) [3].

To improve efficiency by enforcing \((k,m)\)-anonymity with minimal \textbf{Tum}, the Bit-vector Transaction Distance (BTD) is used as optimization for cluster merging. The BTD for records \( r_1, r_2 \) is defined as follows:

\[
\text{BTD}(r_1, r_2) = \frac{1}{\text{ones}(b_1 \lor b_2)} + \frac{1}{\text{ones}(b_1 \land b_2)} + \text{ones}(b_1 \lor b_2),
\]

(3)

where \( b_1 \) and \( b_2 \) are the bit-vector based representations of \( r_1[T] \) and \( r_2[T] \); \( \lor \), \( \land \), and \( \lor \) are the Boolean operators, for XOR, AND, and OR; and the function \( \text{ones} \) counts the number of 1 bits in a bit-vector. The BTD of a cluster \( C \): \[
\text{BTD}(C) = \max\{\text{BTD}(r_1, r_2) | [r_1, r_2] \in C\}. \tag{4}
\]

Poulis et al. also propose the concept of \((k,l^m)\)-diversity that forestalls identity disclosure, and, additionally, the inference of any combination of sensitive items, based on \( l^m \)-diversity [13].

### D. Experimental evaluation

The approach (in particular, the algorithms) proposed in the paper was implemented in C++ and evaluated in terms of data utility and efficiency. The algorithms were applied to two datasets: \textit{Informs} and \textit{YouTube}" (see Table 1).

| Dataset | \(|D|\) | Rel. att. | \(|\text{dom}(T)\)| | Max, Avg # items/rec. |
|---------|--------|----------|----------------|----------------------|
| Informs | 36555  | 5        | 619            | 17, 4.27             |
| YouTube | 131789 | 6        | 936            | 37, 6.51             |

### TABLE I

**DESCRIPTION OF THE DATASETS**

The following parameters have been used \( k = 25 \) for \textit{Informs} and \( k = 100 \) for \textit{YouTube}, and for both datasets \( \delta \) varied in \([X, 1]\), where \( X \) is the NCP of the dataset produced after the Initial cluster formation step, for \textbf{Run-BOUND}, or the \( \text{UL} \), for \textbf{Tum-BOUND}.

Data utility was captured by the Average Relative Error (ARE) [15]—a measure that reflects how well anonymized data was preserved. For example, for the \textit{Informs} dataset, the ARE measures how well the anonymized data preserves the relational attributes such as age, category, length, rate, #ratings, #comments, and the transaction attribute related videos.
support query answering. Workloads, comprised of 100 queries involving relational, transaction, or both attribute types, which retrieve random values or sets of 2 items by default, were constructed.

The experimental results for achieving \((k, k^m)\)-anonymity with the \textbf{Rum-bound} approach show that:

- the data remain useful for queries involving both attribute types;
- for the queries involving only relational values the smallest \textit{ARE} is achieved with lower values of \(\delta\) (lower bounds for \textit{NCP}) by applying \textbf{RMERGE}\(_T\);
- for the queries involving only relational values the smallest \textit{ARE} is achieved with higher values of \(\delta\) by applying the \textbf{RMERGE}\(_R\) algorithm.

\textbf{RMERGE}\(_{RT}\) allows more accurate query answering than \textbf{RMERGE}\(_R\), in relational attributes, and than \textbf{RMERGE}\(_T\), in the transaction attribute, as it merges clusters based on both attribute types.

In an analogous way some observations about \textbf{Tum-bound} approach and \((k, l^m)\)-\textit{diversity} algorithms are presented.

The impact of the dataset size and of the parameter \(k\) on the runtime had been studied as well:

- \textbf{RMERGE}\(_T\) and \textbf{Tmerge}\(_T\) scale better;
- all \textbf{RMERGE} algorithms improve with increasing value of \(k\), as fewer clusters are merged.

To show that using \textit{BTD} improves efficiency without degrading data utility, the authors develop baseline algorithms that do not perform optimization of cluster merging. The experimental results show that when using the \textit{BTD} the efficiency and scalability are at least 10 times better compared to baseline algorithms.

\section*{E. Discussion}

Poulis et al. introduce the problem of anonymizing \textit{RT-datasets} and propose the first approach to tackle it. They develop two frameworks that produce \((k, k^m)\)-\textit{anonymous RT-datasets} with bounded information loss in one attribute type and minimal information loss in the other. Algorithms that preserve different aspects of data utility are proposed and experimentally evaluated.

The approach proposed in the paper can be extended by investigating relationships between \((k, k^m)\)-\textit{anonymization} and relaxed differential privacy definitions [16] to strengthen privacy protection, as it is suggested by the authors.

An essential application of this approach is healthcare, where \textit{RT-datasets} are used for the medical studies. However, usually these studies incorporate different eHealth systems [2], therefore information is distributed across multiple databases. How can one apply \((k, k^m)\)-\textit{anonymization} approach in a distributed environment? How can one select the parameters to bound utility measures of relational and transaction attributes in practice? These questions need to be addressed before it would be possible to apply the framework in a real-word scenario.

Moreover, with the approach proposed by Poulis et al. it is impossible to update the information related to a specific person as well as combine a particular person’s data originating from multiple sources. Nevertheless, bringing the proposed approach together with pseudonymization techniques could be considered in order to overcome the problems of mining dynamic data of the patients.

\section*{V. Research Proposal}

Following the Swiss eHealth strategy, by the end of the year 2015, people living in Switzerland will be able to access their electronic patient records through the health portal [9]. As a step forward, “E-toile”\(^6\), a patient-centered, multi-institutional healthcare information network, has been developed recently as a pilot project in the canton of Geneva. Current advances in eHealth allow for improvements in the patients’ care. Databases that contain information about the patients can be used in the medical studies, particularly, in pharmacokinetics and pharmacodynamics, leading to innovative approaches for TDM. However, to put these into practice many challenges in different research areas need to be overcome. In this research proposal we focus on data management. In the following we state our motivation and goals, briefly review related work in addition to the papers surveyed in the Sections II-IV, and propose research issues and themes.

\subsection*{A. Interoperability in the heterogeneous eHealth system}

In general, information about a patient is large, distributed across the healthcare network, changes over time, and might have different representations [2]. A major objective is to make this information available in its entirety from any segment of the network, where a variety of different policies and standards specify principles of data management.

Use of semantic representations to automatically interpret the shared knowledge between different healthcare systems has been recognized as a powerful approach [18], [19]. Urovii et al. [18] propose a P2P agent coordination model that combine semantic representations and coordination languages to enable dynamic interactions between eHealth systems. Recent research efforts in cloud computing present promising techniques for addressing the problems related to semantic data processing [5]. InterCloud [1] and RESERVOIR [4] are making use of a broker/mediator architecture in order to overcome semantic interoperability conflicts. Several frameworks that use cloud-based approaches for building eHealth systems [6], [7] have been already proposed.

When data about a particular patient is analyzed, one has to take into account that it was collected from multiple sources and, hence, it could be inconsistent and incomplete. We intend to build an interoperable system that enables data exchange and performs error detection and data repair based on the application domain. We intend to revisit and turn to good advantage the principles of architectural design used by Buyya et al. in the construction of InterCloud [1]. Another possible approach to establish the connections between the data sources is leveraging existing schema matching techniques, which establish matches between meta-data, and to create the mappings

\(^6\)http://www.e-toile-ge.ch
(e.g., transformation functions) between the data instances. In this approach, we could focus on the principles of emergent semantics [17] to improve the quality of the matchings.

B. Protection of Patients Privacy

The patient’s data required for the medical research includes the following information: (1) personal data such as gender, age, weight, height, etc; (2) drug intakes such as dosage per day, intervals between two doses, drug’s pharmacokinetics parameters, other drugs’ influences, etc; (3) measurement results such as drug concentration, analyzing time, other symptoms, etc. Since this information contains personal and sensitive data, there is a need to comply with the legislative requirements as to where data are stored, how it is processed and who can access it. For example, the patient’s data should be de-identified for use in clinical studies.

A variety of privacy models, (e.g., $k$–anonymity, $l$–diversity, $e$–differential privacy, etc. [12], [21]) can be used for privacy-preserving data publishing. However, Poullis et al. show in [3] that all these methods are not appropriate for the anonymization of the RT-dataset. The authors propose a new approach for the anonymization of these datasets yet preserving data utility. In [2,10] the authors employ a pseudonymization approach. An advantage of the latter is that it makes possible to link the patient data based on pseudonyms and recontact the patients if needed. However, it does not provide as strong privacy guarantees as anonymization [20].

Integration of medical data is required for conducting clinical studies. To preserve the patients’ privacy we need to control the information flow that crosses organizations [6]. We will examine possible security threats, specific to the implementation of a new approach for therapeutic drug monitoring, and will work towards building a robust access control model integrated into an interoperable eHealth system. We will investigate the possibilities of applying $(k, k^m)$–anonymization approach [3] in a distributed environment and combining it with the pseudonymization techniques. This will allow for defining different privacy views that support visualizing and searching the data while satisfying the privacy requirements.

C. Conclusion

In this research proposal we address the interrelated problems faced in the context of the development of a new approach for TDM: achieving interoperability and data integration in a heterogeneous network of many clinical institutions, while ensuring data quality and users’ privacy. Solving these problems is a crucial step towards making eHealth systems a reality to serve the daily needs of patients and improve medical practice all over the world.

REFERENCES