Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The linked medical data access control framework

Eleni Kamateri, Evangelos Kalampokis *, Efthimios Tambouris, Konstantinos Tarabanis

Information Technologies Institute, Centre for Research & Technology – Hellas, 6th km Xarilaou – Thermi, 57001 Thessaloniki, Greece
University of Macedonia, Egnatia 156, 54006 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 28 August 2013
Accepted 1 March 2014
Available online xxxx

Keywords:
Medical data
Data-cubes
Linked data
RDF
Access policy
Privacy

ABSTRACT

The integration of medical data coming from multiple sources is important in clinical research. Amongst others, it enables the discovery of appropriate subjects in patient-oriented research and the identification of innovative results in epidemiological studies. At the same time, the integration of medical data faces significant ethical and legal challenges that impose access constraints. Some of these issues can be addressed by making available aggregated instead of raw record-level data. In many cases however, there is still a need for controlling access even to the resulting aggregated data, e.g., due to data provider's policies. In this paper we present the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to medical data across distributed sources with diverse access constraints. The LiMDAC framework consists of three Linked Data models, namely the LiMDAC metadata model, the LiMDAC user profile model, and the LiMDAC access policy model. It also includes an architecture that exploits these models. Based on the framework, a proof-of-concept platform is developed and its performance and functionality are evaluated by employing two usage scenarios.

1. Introduction

Clinical research aims at finding new and better ways to understand, diagnose, prevent, or treat a specific pathological process, e.g., diseases or adverse events. It comprises three main categories: (i) the patient-oriented research that involves human subjects; (ii) the epidemiological and behavioral studies that examine the distribution of disease and the factors that affect health; and (iii) the outcomes and health services research that seeks to identify the most effective and efficient interventions, treatments, and services [1].

Clinical research often requires the integration of medical data coming from multiple datasets that are usually stored across multiple sources such as hospitals, clinical sites, research institutes and pharmaceutical companies [2–5]. Medical data may contain sensitive patient data such as demographics, diagnoses, and medication, as well as radiology images, laboratory test results, doctors' entries and comments [6,7].

In patient-oriented research, the integration of multiple medical datasets enables the identification of a sufficient number of subjects [8]. For example, clinical trial phase III, which assesses the safety and the efficacy of a studied treatment or drug, requires large groups of people matching specific eligibility criteria that cannot be found through a single clinical site. In epidemiological studies, analysis of integrated datasets improves the statistical power of results. For instance, studies of clinical effectiveness or disease biology in rare diseases are only possible through multicenter analyses [9]. The integration of multiple datasets also enables better understanding of relationships between pathological processes and risk factors, or between genotype and phenotype [10,11]. For example, recent genome-wide association studies identified 13 novel loci associated with systolic and diastolic blood pressure as well as hypertension [12,13].

At the same time however, clinical researchers face technical and interoperability [14], as well as ethical and legal [15], challenges in discovering and accessing scattered and heterogeneous medical data. Although the former challenges have been addressed by several standards [16,17], the latter still remain.

In order to overcome these ethical and legal challenges, the approach of aggregating data has been proposed and widely employed. According to this approach, only the counts of subjects having specific characteristics are reported instead of raw record-level data, guaranteeing in this way non-identification and anonymization. Despite that, there is still need for controlling access even to aggregated data, e.g., due to data providers' policies. A promising technology that facilitates data discovery and access at a Web scale is Linked Data. Linked Data refers to *"data published*

* Corresponding author at: University of Macedonia, Egnatia 156, 54006 Thessaloniki, Greece.
E-mail addresses: ekamater@uom.gr (E. Kamateri), ekal@uom.gr (E. Kalampokis), tambouris@uom.gr (E. Tambouris), kat@uom.gr (K. Tarabanis).

http://dx.doi.org/10.1016/j.jbi.2014.03.002
1532-0464/

on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets" [18]. Currently, the most promising implementation of Linked Data involves publishing structured data in RDF using URIs in contrast to the full-fledged Semantic Web vision focusing on the ontological level or inferencing [19].

The objective of this paper is to present the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to medical data across distributed sources with diverse access constraints. The framework consists of (a) three Linked Data models, namely the LiMDAC metadata model for describing aggregated medical data, the LiMDAC user profile model for describing medical data consumers, and the LiMDAC access policy model, and (b) an architecture that exploits and orchestrates the three models to enable controlling access to medical data. From a technological perspective, the framework is validated using a proof-of-concept platform that is developed for that purpose.

The remaining of the paper is organized as follows. In Section 2, we provide background knowledge that is necessary for scoping and presenting our work. Section 3 presents the state-of-the-art regarding (a) existing solutions for controlling access to medical data and (b) the use of Linked Data technologies in medical data and for access control. Section 4 describes the proposed LiMDAC framework while Section 5 illustrates a proof-of-concept implementation of the LiMDAC framework. Finally, in Section 6 the results are discussed and in Section 7 conclusions are drawn.

2. Access constraints

Ethical and legal challenges related to medical data mainly derive from (a) strict regulations that protect personal data and prevent patient re-identification by any means [20], (b) agreements that are specified in consent forms, e.g., patients approve sharing their data only in certain clinical studies [21], and (c) policies of stakeholders owing the data e.g., pharmaceutical companies do not contribute to a clinical research led by competitors, or physicians exclude data derived from studies in progress [22–24].

In general, ethical and legal challenges impose access constraints that can be categorized as follows [25,26]:

- *Secrecy:* Ensures the privacy of patients and the confidentiality of medical data preventing unauthorized disclosures of information.
- *Integrity:* Ensures the integrity of medical data and prevents unauthorized or improper modifications of data.
- *Availability:* Ensures the availability of medical data only to authorized persons and prevents the unauthorized or unintended withholding of data.

In order to overcome *secrecy* and *integrity* constraints, the approach of aggregating data has been proposed. Aggregated data includes only counts of patients having specific characteristics instead of raw record-level information. Aggregated data is usually structured in the form of multi-dimensional data cubes [27,28]. In this way, non-identification and anonymization are ensured while the original data remain safe from any modifications. Despite that, there is still a need for controlling access to aggregated data, e.g., due to data provider's policies, and thus *availability constraints* call for appropriate solutions [29].

In order to elaborate on availability constraints, a patient-oriented research and an epidemiological study scenario are described below.

2.1. Patient-oriented research

In patient-oriented research, clinical researchers search for subjects that meet certain eligibility criteria related to a clinical study. Initially they identify possible data providers and ask them whether data of relevant subjects is included in their patients' database. Data providers having such data and wishing to participate to the specific clinical study have to perform some intensive tasks. First, they check whether the identified subjects can be included according to the study's eligibility criteria. Then, they match the eligible subjects with the patients' consent forms to identify if they can be enrolled to the specific trial. Finally, they confirm that access to the patient data is permitted without violating any access constraints, e.g., when subjects have been recruited for a different trial. If the number of eligible patients is not sufficient, clinical researchers seek for additional subjects from other sources to meet the recruitment target.

2.2. Epidemiological study

In epidemiological studies, clinical researchers perform statistical analyses of medical data in order to conduct secondary clinical research and thus, identify risk factors influencing the occurrence of a pathological process. In order to have accurate and statistically significant results, they need a large number of medical data. To this end, they identify possible data providers and ask for relevant data they can access without violating any access constraints. Data providers wishing to contribute to the specific clinical research have to perform some intensive tasks. First, they modify their data in order to ensure that their data will be transferred in an anonymized and non-identifiable form. To achieve this, they delete all references to the subject and create aggregated data. Data providers confirm that access to the data is permitted without violating any policies and provide the data e.g., data that is used for studies in progress.

In addition to these scenarios, we interviewed stakeholders working in organizations that participate in the EU funded FP7 Linked2Safety project [30]. In particular, we interviewed five clinical researchers, one data manager and three clinical study managers coming from three healthcare organizations maintaining and using medical data for clinical research, namely the Institute of Neurology and Genetics in Cyprus, the Lausanne University Hospital, and ZEINCRO Hellas S.A., a private Contract Research Organisation in Greece. This exercise resulted in a list of user requirements that are related to availability constraints. This list is presented in Appendix A.

These scenarios and requirements enable us to identify that two *abstract roles* related to medical data management are important in clinical research. The *data provider* creates and keeps medical data regarding patient-specific information in order to organize patients' treatment, or conduct a clinical research. The *data consumer* discovers subjects meeting certain eligibility criteria for a patient-oriented research, or medical data to perform an epidemiological study.

In addition, these scenarios and requirements enable us to come up with an *abstract process* that delineates clinical research and consists of the following steps:

1. Data provider modifies and aggregates data in order to ensure anonymity and non-identification.
2. Data consumer searches for data providers.
3. Data consumer asks data provider for certain data.
4. Data provider checks whether the requested data is available.
5. Data provider checks whether data consumer is allowed to access the data according to some access constraint policies.
6. Data consumer receives the data.

Please cite this article in press as: Kamateri E et al. The linked medical data access control framework. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.03.002