Knowledge-Based Systems 79 (2015) 99–117

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An integrated framework for securing semi-structured health records

Flora Amato [a], Giuseppe De Pietro [b], Massimo Esposito [b,*], Nicola Mazzocca [a]

[a] *Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, University of Naples Federico II, Naples, Italy*

[b] *National Research Council of Italy – Institute for High Performance Computing and Networking (ICAR), Via P. Castellino 111, 80131 Naples, Italy*

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 30 January 2015
Accepted 2 February 2015
Available online 11 February 2015

Keywords:
Electronic health record
Role-based access control
Policy language
Information structuring
Ontology

ABSTRACT

In the last years, the adoption of Electronic Health Records (EHRs) have been widely promoted, with the final aim of improving care quality and patient safety. Yet, sharing patient data in a large distributed and heterogeneous context, such as the healthcare domain, has inherently introduced security and privacy risks, due to the great sensitivity and confidentiality of the patient data and the need of accessing such data by a large number of health care workers with various roles for the patient care. Even though various techniques have been developed to effectively implement fine-grained access control, which allows flexibility in specifying differential access rights of individual users, some unsolved problems can be pointed out with respect to the specification of complex policies over EHRs: (i) the difficulty of forcing narrative text to assume a semi-structured coded form into EHRs in order to build access control policies also working at a section-level; (ii) an overly high-level of theoretical ability required to practically use access control models and policy languages as a whole, due to a scarce integration among them; and (iii) the lack of tools for easily editing and upgrading access control policies over EHRs. In order to face all these open issues, this paper proposes a hybrid framework aimed at enabling and supporting the definition of fine-grained access control policies working on semi-structured EHRs. The key issues of the framework are: (i) a semantic-based method that hybridizes linguistic and statistical techniques in order to give a semi-structured form to a narrative text to be inserted into EHRs, by identifying its specific sections; (ii) a formal role-based authorization model, encoded as a couple of ontologies, to regulate the access to these semi-structured EHRs with respect to their sections; and (iii) a procedural policy language and a set of patterns to simply encode and update access control restrictions in the form of "if–then rules" built on the top of the ontological model formalized. A prototype implementation of this framework is realized in the form of a system offering simple and intuitive interfaces to the security administrators. Finally, an experimental evaluation over real documents contained into EHRs, i.e. discharge summaries, is described, showing the feasibility of the proposed framework and suggesting that its application could simply and proficiently secure the access to healthcare information contained into semi-structured EHRs and, thus, face security and privacy risks in real healthcare scenarios.

1. Introduction

Information and communication technologies have greatly affected the delivery of health care, in the form of computerized systems, such as health information systems, clinical information systems, and picture archiving and communication systems, with the final aim of both reducing healthcare costs and improving care quality and patient safety [47]. In particular, in the last years, hospitals and health care providers have increased the adoption of such electronic health care systems to manage patient health care data in the form of Electronic Health Records (hereafter, EHRs) [36]. According to the International Organization for Standardization (ISO) definition, EHR means a repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information, and its primary purpose is to set objectives and planning patient care, document the delivery of care and assess the outcomes of care [26]. This information included in EHRs has several different functions in patient care, management and health policy, generating many benefits for clinicians, clients, and, generally, for the healthcare system [44], such as improved decision-making at the point of care [222], safer drug administration [39] better medication management, better adoption of

* Corresponding author. Tel.: +39 0816139512; fax: +39 0816139531.
E-mail addresses: flora.amato@unina.it (F. Amato), giuseppe.depietro@na.icar.cnr.it (G. De Pietro), massimo.esposito@na.icar.cnr.it (M. Esposito), nicola.mazzocca@unina.it (N. Mazzocca).

http://dx.doi.org/10.1016/j.knosys.2015.02.004
0950-7051/© 2015 Elsevier B.V. All rights reserved.

screening programs, advanced tools and services for remote health monitoring [23,41,49], enhanced communication between a variety of healthcare professionals, and improved resource utilization [38].

The amount and quality of information available to health care professionals in EHRs has a pivotal role to support continuing, efficient and quality integrated healthcare [4], yet sharing patient data in a large distributed and heterogeneous context, such as the healthcare domain, inherently introduces security and privacy risks [36]. In particular, due to the great sensitivity and confidentiality of the patient data and the fact that such data may need to be accessed by a large number of health care workers with various roles for the patient care, a high level of secure protection for data and data access is required. One of the most challenging aspects with respect to security and privacy for healthcare organizations, however, is the amount of power given by 'Patient consent and confidentiality' to the patients in terms of access control restrictions over their individual EHRs [18]. Even though various techniques have been developed to effectively implement fine-grained access control, which allows flexibility in specifying differential access rights for individual users, some unsolved problems can be pointed out with respect to the specification of complex policies over EHRs, where access should be granted or denied according to the right and the need of the healthcare workers to perform a particular job function on specific EHR sections.

The first point, which represents one of the largest potential obstacles, regards the difficulty of forcing narrative text to assume a semi-structured coded form into EHRs in order to build fine-grained access control policies on the top of its specific sections, coupled with the defensive attitude of physicians toward the use of computerized health information systems. Clinicians have a long tradition of using paper forms and dictation services, showing an enduring preference for narrative data (i.e. clinical text written in a natural language, such as Italian or English), due to advantages, such as familiarity, ease of use and freedom to express anything they wish [32]. Indeed, natural language provides many mechanisms that augment or enrich simple facts, for example to qualify their severity or degree, convey temporal relationships, indicate patterns of causality, provide rationale, propose hypotheses, and suggest alternatives [32]. Moreover, producing data without following structural rules can result faster in many situations. Furthermore, healthcare workers are not motivated to indicate a structured or semi-structured form for the data they produce, since scarcely aware of the possible advantages, such as automatic document processing or a more effective data protection. As a consequence of that, a fundamental requisite for efficiently and accurately securing healthcare records is to automatically transform narrative data into semi-structured EHRs, suitable for machine processing.

Secondly, with reference to existing access control models and mechanisms, many efforts have been made to meet the specific security and privacy needs of the healthcare domain. In parallel, and almost separately, general-purpose policy languages for access control have been defined, without tying to a model, in order to improve the usability and simplicity for the designers. Indeed, existing models may not have the machinery to express all the policy details of a given system or may deliberately leave important aspects unspecified [21]. To face such an issue, formal access control models and policy languages should be harmonized in the sense that advanced access control concepts should be embodied into a model as well as integrated and supported by a policy language in a natural intuitive manner [3]. In particular, policy languages should allow restrictions to be described over a sharable and semantically well-defined domain model to promote common understanding among healthcare workers and support automatic reasoning about them.

Finally, since access control models and policy languages require a good expressive power, they typically suffer from a resistance amongst real users, even those entrusted with the management of access control policies, due to their highly complex syntax. Even though they are not required to be generally written for widespread use, this practical issue strengthens the convincement that different solutions should be used for encoding access control policies for healthcare systems without assuming an overly high level of theoretical ability. As a consequence of that, one prerequisite for the broad acceptance of them and their efficient application to medical settings is the guarantee of a high level of upgradability and maintainability, (i) to modify access control policies according to dynamically changing security needs, or (ii) to adapt generic, site-independent access control policies to a specific healthcare organization. Since updating policies can require a continuous intervention, it is unthinkable that it cannot be done directly by security administrators when needed. In contrast to the intensive efforts made to develop access control models and policy languages, the issue of providing solutions for easily editing and upgrading access control policies has been widely neglected thus far.

In order to face all these open issues, this paper proposes a hybrid framework aimed at enabling and supporting the definition of fine-grained access control policies working on semi-structured electronic health records. In more detail, the strong points of the proposed framework can be summarized as follows.

First, the framework provides a semantic-based method that hybridizes linguistic and statistical techniques in order to give a semi-structured form to narrative text to be inserted into an EHR, by first identifying relevant concepts in textual data and, successively, delimiting the sections composing the document depending on the concepts recognized.

Second, an authorization model for role-based access control (hereafter, RBAC) with respect to EHRs has been proposed. It extends the National Institute of Standards and Technology (hereafter, NIST) RBAC reference model to regulate the access to EHRs, working not only at a document level, but also with respect to their different sections. In other words, it is aimed at determining the authorization decisions that enable healthcare workers to carry out specific tasks on the different sections of EHRs. The framework supports this RBAC model by using Semantic Web technologies and, in particular, a high-level ontology that expresses the elements of the proposed RBAC model, and, in addition, a domain-specific ontology that captures the features of a specific application domain.

Third, this framework also includes a procedural policy language to encode access control restrictions in the form of "*if–then rules*" built on the top of the ontological formalizations of the elements belonging to the proposed RBAC model. Plus, a set of patterns has been defined for supporting the simple insertion and editing of such access control restrictions with the aim of reducing the complexity of the formalization process, by graphically guiding the definition of policies that could be functional in the context of healthcare organizations and enabling their automatic encoding into machine executable languages.

A prototype implementation of this framework has been realized in the form of a system offering simple and intuitive interfaces to the security administrators who do not have a deep technical expertise. It provides a set of facilities for structuring narrative text into an EHR and writing access control policies, by hiding the syntax constructs used in both the proposed RBAC model and policy language.

The rest of the paper is organized as follows. Section 2 introduces an overview of the state-of-the-art solutions for building access control policies on the top of semi-structured EHRs and clearly highlights the motivations for this work and its research