doi:10.3233/SHTI210953

Embedding Risk-Based Anonymization into Data Access Control for Providing Individual-Level Health Data in a Secure Way

Murat SARIYAR [a,1] and Jürgen HOLM [a]
[a] Bern University of Appl. Sciences, Department of Medical Informatics, Switzerland

Abstract. Especially in biomedical research, individual-level data must be protected due to the sensitivity of the data that is associated with patients. The broad goal of scientific data re-use is to allow many researchers to derive new hypotheses and insights from the data while preserving privacy. Data usage control (DUC) as an attribute-based access mechanism promises to overcome the limitations of traditional access control models achieving that goal. Park and Sandhu provided the usage control (UCON) model as an instance of DUC, which defines policies that evaluate certain attributes. Here, we present an UCON-based architecture, which is augmented with risk-based anonymization as provided by the R package sdcMicro and an extensible Access Control Markup Language (XACML) environment with a core policy decision point as implemented by authzforce.

Keywords. Data Security, Data access control, XACML, Anonymization

1. Introduction

Protecting data in the context of computer-assisted processing of individual-level data requires the usage and/or implementation of security mechanisms. Especially in biomedical research, individual-level data have to be protected due to the sensitivity of the information that is associated with patients, e.g., the propensity to develop breast cancer [1, 2]. Protecting privacy risks requires technical as well as organizational measures, such as the usage of terms & conditions before authorization of data users [3].

One central aim of many scientific data re-use scenarios is allowing many researchers to derive new hypotheses and insights. Such a broad goal requires high flexibility in using as much of the data as possible without compromising data privacy and security. On the one hand, authorization and control of the user activity is often not sufficient for preventing disclosure of sensitive information, as de-anonymization scandals showed [4]. On the other hand, limiting the amount and type of operations on data to ensure high protection decreases the utility of the data. One way to tackle such settings is relying on data-usage-control that considers de-anonymization risks [5].

Data usage control promises to overcome the limitations of traditional access control models. Standard access control protocols regulate the issue of granting access to an

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

object with certain rights (e.g., read only for a certain role, where a role is a named collection of users and relevant permissions). Park and Sandhu [6] provided the usage control (UCON) model for a systematic usage control. Unlike role-based access control, which assigns to certain users pre-defined roles with a set of privileges associated with them, UCON is an attribute-based access control mechanism, which defines policies that evaluate many different attributes. The evaluations are based on three decision factors: authorizations, obligations, and conditions. Authorizations are functional predicates that are evaluated to decide whether a subject is allowed to perform a request on an object (in most cases data). Obligations are requirements a subject must fulfill, and conditions relate to system-related decision factors that are independent of subjects and objects.

There are already works on integrating anonymization procedures into attribute-bases access mechanisms available (see [7]), mainly focusing on the differential privacy (DP) model, which requires that the presence or absence of any individual record must not affect the answer of a query. DP is especially designed for interactive query settings with aggregate outcomes. Even though, there are adaptation to the case of publishing data to allow much more flexibility, the required noise addition is often too high in terms of the resulting data utility [8]. One reason is the difficulty of transforming the output noise addition mechanism into an input-related one for single records, which contradicts the core idea of DP. Hence, k-anonymity models using methods such as shuffling, generalization and micro-aggregation are often more promising in terms of risk-utility trade-offs for dealing with individual-level data in a flexible way [9].

Main motivation for this work is the implementation of a health network platform for pharmacogenetic (PGx) treatments and research. An authorized user of the platform, for example a clinical researcher, should be able to use the platform for SNP-related association analyses. Here, we present the UCON-based architecture and the tools, which is augmented with risk-based anonymization as provided by the R package sdcMicro and an extensible Access Control Markup Language (XACML) environment with a core policy decision point as implemented by authzforce [10].

2. Methods

First, we apply the UCON framework to our pharmacogenetic platform, after which a risk-based anonymization approach is incorporated.

UCON allows to formulate restrictions (rules) on user's access to the data and the operations that are allowed for them. Policies or rules are tuples of the form (subject, action, resource, purpose, system-related condition, user-specific obligation), e.g. (Researcher, machine learning analysis on the data on the client, SNP data, hypothesis generation for research, anytime, ethical statement is signed). The common example for obligations is acceptance of terms of use, being a mandatory requirement that must be met before or during access. Conditions are not directly related to subjects and objects, but to environmental and/or system requirements that must be satisfied, e.g., access is only granted when the system load is under some threshold. Based on the associated attributes of the subject and the resources, the system evaluates these policies for decisions on data access. Hence, defining these attributes is central for the implementation of UCON. For implementing the UCON components, the natural choice is the extensible Access Control Markup Language (XACML).

For anonymizing data, it is crucial to determine the quasi-identifiers (QIDs), which are those attributes that have discriminatory value (i.e., they increase the probability of re-identification), can be obtained from external resources, and are potentially useful for the data user. Examples of QIDs are gender, age, postal codes, race, ethnicity, etc. These QIDs must be protected from being used for disclosing information of identifiable individuals, which is done by using data perturbation techniques, such as shuffling, generalization and micro-aggregation. From a statistical perspective, a procedure for protecting sensitive data should be based on a disclose scenario, dealing with risks and utility at the same time. Typical thresholds for maximal accepted risks lie between 0.005 and 0.01, while thresholds for utility are use-case dependent. As a general utility measure, usually the entropy measure of the loss of information is used. While definition of QIDs and risk threshold will be made together with the implementation of UCON, risk and utility estimation are part of the policy evaluation during access request.

3. Results

The relevant subject (s) and resource (r) attributes of the privacy aware UCON system are given in Table 1. Subjects are assigned to different roles according to their clearance level, which are associated with certain risk thresholds (the concrete numbers are omitted here), general data utility properties, and the related data perturbation. For example, clinical users can access data with high-risk thresholds and high utility by changing nothing of the QIDs. For researchers, the necessary changes for a risk-utility balance are dependent on the origin of the researcher. Trade-off for data utility means, that an iterative process is allowed, in which the allowed risk threshold can be changed through additional security measures (e.g., data usage is continuously monitored), if the researcher is not satisfied with the anonymization result. For public use, only highly aggregated data is provided without allowing any compromise.

Table 1. Core attributes for the subjects (s) and the resources (r) that are used by our UCON system.

Roles=Purpose (s)	Security Level (r)	Data utility (r)	Data perturbation (r)
Clinical use	High risk threshold	High	Raw or pseudonymized
Research intern	Medium risk threshold	Trade-off	1st level anonymization
Research extern	Low risk threshold	Trade-off	2nd level anonymization
Public use	Very low risk threshold	Low	Highly aggregated

For implementing UCON with risk-based anonymization, an authzforce server was installed that provides a multi-tenant RESTful API to policy administration points (PAP) and policy decision points (PDP). A web service wraps the data access request of a subject, which is sent to the server. The REST request has four parts: subject, resource, action, and an environment. In our use case, the subject is an internal researcher (role), the resources are PGx data sets, actions are they ways to access the data (e.g., reading, or on-site analysis) and via the environment part the purpose, obligation statements as well as the required data utility is specified (see Section 2 on the expected 6-element tuples). Credentials are provided by SAML tokens, which contain user IDs and roles. On the client, the request is processed by an authzforce module that extracts the various attributes to request a decision from the remote authzforce server, which enforces PDP decisions.

The PDP decisions are implemented with respect to the attributes in Table 1. If an internal researcher requires data with maximum allowed information loss in the environment attribute that cannot be achieved by anonymization for his security level, the request is denied with a proposal to adjust his required entropy value. To compute the best achievable entropy value, the sdcMicro package is run manually before any requests and the results are stored on the PDP as resources for the responses. We used minimal sample uniques for the risk estimation (SUDA) and the following four anonymizing techniques: generalization by recoding, post-randomization, micro-aggregation, and shuffling. The system is just a prototype and not productive yet, due to the lack of certification of the whole PGx platform as a medicinal product.

4. Discussion

Even though, our proposed system seems feasible for the practice, we advise to update the risk model and the risk thresholds in certain time intervals, since risks and the status of anonymity change with time. If none of the allowed risk thresholds is compatible with the required utilities, additional measures such as highly secure analysis environment could be established to allow an increase of the allowed risk threshold. In other words, the whole system cannot be assessed by certain results of performing analysis on the perturbed data, as this only evaluates the anonymization component, not the attribute-based data access system. For assessing the system in practice, it will be important to record the number of users, the number of rejections as well as trade-off rounds, and the satisfaction of the data users with the data quality.

There are many options for concretizing authorizations, obligations, and conditions. We emphasize, that data security is not only a technical issue. Should the platform be open for all kind of analysis? What kind of restrictions seems necessary to guarantee data protection? What kinds of restrictions have small and high impact on the utility of the data? What is the standard use case? Who is the typical user? Based on answers to these and related questions, the system should be adapted to different scenarios. The flexibility of the UCON model is conducive for such adaptations, even though it is not necessary to stick totally to it if the system has a limited scope as in our scenario.

References

[1] Castellani J. Are clinical trial data shared sufficiently today? Yes. BMJ 357 (2013), f1881.

[2] Eichler H-G, Pétavy F, Pignatti F, et al. Access to Patient-Level Trial Data — A Boon to Drug Developers. N Engl J Med 369 (2013), 1577–1579.

[3] Peltier TR. Information Security Policies, Procedures, and Standards: Guidelines for Effective Information Security Management. CRC Press, 2016.

[4] Janmey V, Elkin PL. Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack. AMIA Annu Symp Proc 2018 (2018), 1329–1337.

[5] Dimitrakos T, Martinelli F, Pretschner A, et al. Proceedings of the 4th International Workshop on Security and Trust Management (STM 2008): Policy Evolution in Distributed Usage Control. Electron Notes Theor Comput Sci 244 (2009), 109–123.

[6] Park J, Sandhu R. Towards Usage Control Models: Beyond Traditional Access Control. In: Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies. New York, NY, USA: ACM, pp. 57–64.

[7] Metoui N, Bezzi M. Differential Privacy Based Access Control. In: Debruyne C, Panetto H, Meersman R, et al. (eds) On the Move to Meaningful Internet Systems: OTM 2016 Conferences. Cham: Springer International Publishing, 2016, pp. 962–974.

[8] Domingo-Ferrer J, Sánchez D, Blanco-Justicia A. The limits of differential privacy (and its misuse in data release and machine learning). Commun ACM 64 (2021), 33–35.

[9] Templ M, Kowarik A, Meindl B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. J Stat Softw 67 (2015), 1–36.

[10] AuthzForce (Community Edition) - XWiki, https://authzforce.ow2.org/ (accessed 23 August 2021).