Security and Privacy Issues for enabling the Secondary use of Electronic Health Records (EHRs) in Clinical Research

Elif ERYILMAZ, SRDC Ltd./Ankara, Turkey
Agenda

- Current Situation & Objectives
- SALUS Project
- Legal Framework
  - Standards and Regularity Guidance
  - Additional Guidance for Implementation
- SALUS Security and Privacy Architecture
  - Data Level Protection Mechanisms
  - Additional Security Services
- Current Results & Future Work
- Acknowledgements
Current Situation and Objectives

- One of the most important challenges in the context of secondary use for clinical research: “Security and privacy of the medical data”
- Lack of standards-based and adaptable security and privacy mechanisms that can be used by both clinical care and clinical research parties in an interoperable manner.
- To find a balance between the privacy concerns for the use of personal data and the requirements of clinical research environments that aim to serve to the public good.
- Development of an extensible security infrastructure that supports re-use of the Electronic Health Records (EHRs) for strengthening the post-approval drug safety studies in the area of clinical research.
SALUS Project

- SALUS Project (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies)
- Aiming at to create the necessary infrastructure to enable secondary use of EHRs in an efficient and effective way for reinforcing the post market safety studies so that patient safety can be ensured through early detection of rare adverse events.
  - Functional interoperability profiles have been developed to query population based EHR data from distributed EHR systems for carrying out post market safety studies.
  - As a result of these population based queries, a set of medical summaries of the eligible patients can be shared in standard based medical summary formats.
  - From the security and privacy point of view, novel data protection mechanisms have been developed to work directly on top of queried clinical data instances for the post-market safety studies.
Standards & Regularity Guidance

- **Selected standards:**
  - International Organization for Standardization (ISO) Health Informatics - Pseudonymisation, Published Technical Specification # 25237 (ISO TS25237).

- **Regularity guidance:**
    - This is a specific reference to 45 CFR 164.502(d) which specifies the general rules for uses and disclosures of de-identified protected health information.
Additional Guidance for Implementation

- HITSP (Healthcare Information Technology Standards Panel) Anonymize Public Health Case Reporting Data Component: HITSP/C87
- IHE IT Infrastructure Healthcare Pseudonymization Handbook (Internal Draft)
- NHS (National Health Service) Pseudonymisation Implementation Project (PIP)
SALUS Security and Privacy Infrastructure

- Establishment of the security framework which is compliant with legal and ethical requirements on the national and European level after these detailed analysis
- Implementation of the Security Architecture including
  - Data Protection Mechanisms applied to the queried clinical instances represented in standard (CDA/RDF) formats and
  - Additional Security Services compatible with standard profiles that guarantees the safe use of EHRs for the clinical research studies.
Data Protection Mechanisms

- Development of novel data protection mechanisms to work directly on top of queried clinical data instances for the post-market safety studies including
  - De-identification Service
  - Pseudonymization Service

- The overall data level protection infrastructure in the context of Care and Non-care Zones is depicted in the following figure:
Data Level Protection Architecture

EHR Database → EHR extractor and Synchronizer → Clinical Data Pool

Care Zone

IDAT+ MDAT+ PID

query

MDAT+PID → CDA/RDF patient medical summary

De-identification Service

MDAT(enc)+ PID

Pseudonymization Service

PID → PSN + MDAT(enc)

Research/Non Care Zone

PID Service (optional)
De-identification Service

- De-identification
  - Process of removing personal identity revealing attributes and replacing the required identifiers and attributes required for research purposes either with pseudonyms or when possible with more generalized categories
  - No single de-identification procedure!
    - Context based
  - Degree of anonymity is important
    - To prevent linking with external source of information
  - We have carried a risk analysis for each data elements in the clinical instances:
    - We have created a flexible security architecture, where some thresholds for the uncommon cases can be configured with the Data Protection Offices of EHR sources according to their preferences.
Data requirements

- First, we have identified data requirements for each pilot application scenarios in the project.
- Sample view from this complete list is presented below:
  - “x”: must have
  - “(o)” : nice to have (optional)

<table>
<thead>
<tr>
<th>Selected SALUS Scenarios/Related EHR Sections</th>
<th>Selected SALUS Scenarios/Related EHR Data Items</th>
<th>Enabling Notification of Suspected AEs</th>
<th>Enabling Semi-automatic ADR Reporting</th>
<th>Characterising the cases and contrasting them to a background population</th>
<th>Temporal pattern characterisation</th>
<th>Exploratory Analysis Studies on EHRs for Signal Detection (Temporal Association Screening)</th>
<th>Secondary use data sources for Post Marketing safety studies (Calculating incidence rates of CMF in diabetic patients with a recent acute coronary syndrome (ACS) event)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Patient Name or Initials</td>
<td>x(o)</td>
<td></td>
<td>x Year of Birth is Adequate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of Birth</td>
<td>x</td>
<td>x(Age is to be derived)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>x</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Race</td>
<td>x</td>
<td>x(o)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Birth Place (Region or City)</td>
<td>x</td>
<td>x(o)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patient registration date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patient de-registration date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>General Practitioner Record Number</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hospital Record Number</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Investigation number</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td>x Year of Birth is Adequate</td>
<td></td>
</tr>
<tr>
<td>Pregnancy</td>
<td>YES/NO</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delivery Date</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pregnancy Status</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Last Menstrual Period Date</td>
<td>x(o)</td>
<td></td>
<td></td>
<td></td>
<td>x Year of Birth is Adequate</td>
<td></td>
</tr>
</tbody>
</table>
De-identification Methods

- IHE IT Infrastructure provides the high level description of key categories for de-identification as follows:

<table>
<thead>
<tr>
<th>De-identification Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redaction</td>
<td>Removing an atomic data element</td>
</tr>
<tr>
<td>Fuzzing</td>
<td>Adding “noise” to an atomic data element</td>
</tr>
<tr>
<td>Generalization</td>
<td>Making an atomic data element less specific</td>
</tr>
<tr>
<td>Longitudinal consistency</td>
<td>Modifying data so that it is shifted by a specific amount</td>
</tr>
<tr>
<td>Text Processing</td>
<td>Special considerations for free-format text</td>
</tr>
<tr>
<td>(Recoverable) Substitution</td>
<td>Changing one data element into another data element</td>
</tr>
<tr>
<td>Pass-through</td>
<td>No change</td>
</tr>
</tbody>
</table>
De-identification algorithms determined

- We have decided possible data protection mechanisms for each of data elements identified before.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Algorithm</th>
<th>Demographics</th>
<th>Patient Name or initials</th>
<th>DOB</th>
<th>Date of birth</th>
<th>Gender</th>
<th>Race</th>
<th>Birth Place (Region or City)</th>
<th>Patient Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Redaction</strong></td>
<td>Delete Attribute Name and Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delete Value</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fixed Length Substitute Value</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Original Length Substitute Value</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Type Labeling</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Substitution</strong></td>
<td>Substitute Attribute Name and Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Substitute meaningful value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Substitute Meaningless Value</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recoverable Substitution</strong></td>
<td>Pseudorandom Values</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequential Values</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fuzzing</strong></td>
<td>Numeric (statistical algorithm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Code Set (Random)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>DOB to age</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shift by randomoffsets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remove day/month/year</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geographical locations</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pass-through</strong></td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Text Processing</strong></td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Pseudonymization Service

- **Pseudonymization:**
  - A particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms
    - provides a means for information to be linked to the same person across multiple data records without revealing the identity of the person as a data subject which is often required for clinical research studies
- **Types:**
  - **Irreversible pseudonymization:** hash based
  - **Reversible pseudonymization:** two way cryptography
Additional Security Services

- The data level protection mechanism is supported by additional mechanisms compliant with IHE Audit Trail and Node Authentication (ATNA) profile to ensure the clinical data security with the following principles:
  - Each transaction should be audited to an audit repository to ensure accountability
  - Each node should be mutually authenticated through X509 certificates
  - Transport Layer Security (TLS) should be used for each transaction

- The applicability of the IHE ATNA profile is examined in two separate categories for these additional security mechanisms as “Auditing” (ITI-20: Record Audit Event) and “Message Level Security” (ITI-19 Authenticate Node).
Results

- Implementation of de-identification algorithms on top of sample clinical data instances represented through CDA/RDF based templates of IHE PCC and HL7/ASTM CCD is finalized.

- Both irreversible (hash-based) and reversible pseudonymization algorithms (based on two way cryptography algorithms) on top of the retrieved input from the De-identification Service are implemented.

- Initial design of additional security services are done to support data level protection.
An extract from RDF before de-identification
An extract from RDF after de-identification

```json
[]
  rdf:type salus:patient;
  salus:ID
  [ rdf:type salus:ii;
    salus:extension "17689263734";
    salus:root "2.16.840.1.113883.2.9.4.3.2"
  ];
  salus:address
  [ rdf:type salus:addr;
    salus:city "";
    salus:country "Italy";
    salus:postalCode "12345";
    salus:state "MI";
    salus:streetAddressLine ""
  ];
  salus:allergy
  [ rdf:type salus:Allergy;
    salus:adverseEventDate
    [ rdf:type salus:iVTs;
      salus:high "2010-05-01T00:00:00"^^xsd:dateTime;
      salus:low "2009-08-24T00:00:00"^^xsd:dateTime
    ];
    salus:adverseEventType
    [ rdf:type salus:cd;
      salus:code "235719002";
      salus:codeSystem "2.16.840.1.113883.6.96";
      salus:codeSystemName "SNOMED CT";
      salus:displayName "Food intolerance"
    ];
    salus:product
    [ rdf:type salus:cd;
      salus:code "303300008";
      salus:codeSystem "2.16.840.1.113883.6.96";
      salus:codeSystemName "SNOMED CT";
      salus:displayName "Egg protein"
    ];
];
```
An extract from CDA before de-identification
An extract from CDA after de-identification
Applied Techniques

As seen in the figure, we have applied the following methods for the specific data elements presented in the figures:

- **Ids** -> Replace all ids with generated random values
- **Address** -> Remove address information that is more specific than a state
- **Name** -> Substitute Meaningless Value
- **Birthdate** -> Generalize date to year
- **Birthplace** -> Remove place information that is more specific than a state
- **Date/Time values** -> Shift the values within the same amount
Results in summary

- We have implemented these methods by using Service-Oriented methodology providing extensible approach to implement additional methods with respect to the requested features of the end-users.

- As a result, we have achieved this security infrastructure with the novel aspects as follows:
  - De-identification is processed on top of the queried clinical data instances instead of all data elements in a data warehouse,
  - An extensible de-identification framework is created based on SOA principles using RESTful implementations in a modular way that makes further development easy based on the needs of the end-users,
  - A flexible de-identification framework is created where the de-identification method can be configured for each data element after analyzing and assessing the risks for each data element set that needs to be exchanged in a de-identified manner with the respective stakeholders.
Future Work

- **Additional Security Services**
  
  When we have finalized the work from the data level protection security, we will integrate the additional security services related to audit trail and node authentication compliant with the IHE ATNA Profile.
  
  These additional mechanisms will include authentication and authorization techniques to be decided later within the scope of the project.
Acknowledgements

- The research leading to these results has received funding from the European Community’s Seventh Framework Pro-gramme (FP7/2007-2013) under grant agreement no ICT-287800, SALUS Project (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies).


- Our submission about this part of the project presented here has also been accepted to be presented in the 14th World Congress on Medical and Health Informatics (MedInfo) 2013, August 20-23, in Copenhagen/Denmark.
References

- NHS (National Health Service) Pseudonymisation Implementation Project (PIP), http://www.connectingforhealth.nhs.uk/systemsandservices/pseudo
Elif Eryilmaz
Senior Software Engineer
Software Research, Development and Consultation (SRDC) Ltd., ODTU Teknokent Silikon Blok Kat:1 No:14
Middle East Technical University Campus
06531 Ankara TURKEY
Email: elif@srdc.com.tr
Phone: +90 (312) 210 1763
Fax: +90 (312) 210 1837
Thank you...