Privacy-aware analytics on healthcare data

Henk-Jan Meijer, Albana Gaba, Yeb Havinga

HL7 Security Workgroup Meeting

San Antonio TX - January 2014

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 318633

This work is released under the Creative Commons cc-by-sa license
Amsterdam, The Netherlands

Health IT supplier

European research projects:
- COMMODITY12
- AXLE

Chronic diseases:
- diabetes
- anticoagulation
- pulmonary diseases
...

Actively participating in HL7, IHE, CIMIT

10+ years experience in SaaS / eHealth / Telemedicine

Used daily by 6,000 medical professionals

Largest telemedicine provider in Europe[1]: 65,000 on-line self management patients

70 medical organizations

Treatment, Care and Prevention: 380,000 patients
The AXLE project

- European-funded innovation project
- **Analytics on eXtremely Large European data:**
  Large-scale complex analytics on real-world datasets, while addressing the full requirements of real datasets (data quality, privacy, security and auditability)[2]
- Time frame: Nov 2012 – Nov 2015
- Portavita and four partners from academia and industry:
Health data as real-world dataset

Portavita provides:
- Synthetic CDA-based dataset generator
- Complex data models (HL7 RIM-based)
- Realistic use-cases of data use

Required granular access control to patient data
- Treatment of patient data is protected by law
- Patient's consent and organization policies
Health data as real-world dataset

Portavita provides:
- Synthetic CDA-based dataset generator
- Complex data models (HL7 RIM-based)
- Realistic use-cases of data use

Required granular access control to patient data
- Treatment of patient data is protected by law
- Patient's consent and organization policies

Can we use HCS[3] for data analytics purposes?
Outline

- Portavita use cases for data analytics
- Legal requirements
- Patient consent
- De-identification techniques
- HCS components in AXLE design
Use Case 1: Public Health

In The Netherlands caregroups involve various care providers (e.g., GPs, dieticians, specialized doctors, etc.) for the treatment of chronic diseases

The National Institute for Public Health and the Environment (RIVM) wishes to monitor the process of treatment of chronic diseases

Examples of operations on the data:

- How frequently are the patients making regular check-ups
- The impact of such new process on the health of patients
- Comparison of the performance of various care groups
- ...

Portavita is required to export a de-identified dataset of patients data (EHRs)

Purpose of use: research for general interest
Use case 2: Business Intelligence

The organizations (caregroups) wish to process patients' data for business intelligence purposes

Examples of operations on the data:
- Quality assessment across various care providers
- Monitoring, reporting of various treatments
- ...

Roles within the organizations:
- Super user: access all patients' data
- Research user: access de-identified patients data

Purpose of use: internal research
Legal background: EU directives

  - Patient data is personal data => sensitive => explicit consent is required for treatment of the data
  - Anonymous data fall outside the scope of the EU Directive => no consent required
Legal background: Dutch law[6]

- Patient data can be used for health research without consent if one of the following two conditions is met:
  a) It is not reasonably possible to ask for consent and the privacy of the patient is not unnecessarily jeopardized
  b) Given the nature of research, asking for consent is not feasible and the data arrive at the researcher in such a way that re-identification is sufficiently prevented.

- In both instances three other conditions have to be met:
  1. The research serves a general interest
  2. The research cannot be carried out without those data
  3. The patient has not objected to such use of his/her data for research
Legal background: Dutch law[6]

- Patient data can be used for health research without consent if one of the following two conditions is met:
  a) It is not reasonably possible to ask for consent and the privacy of the patient is not unnecessarily jeopardized
  b) Given the nature of research, asking for consent is not feasible and the data arrive at the researcher in such a way that re-identification is sufficiently prevented.

- In both instances three other conditions have to be met:
  1. The research serves a general interest
  2. The research cannot be carried out without those data
  3. The patient has not objected to such use of his/her data for research
Guidelines from legal analysis

• Patient consent:
  – Patients can opt-out from data treatment for research purposes

• Re-identification of patients should be sufficiently prevented
  – Safe pseudonymization of data subject
  – Suitable aggregation level of the research data related to the pseudonym
Consent CDA[7]: Research opt-out

.......... 
<!-- Privacy Consent Directive Entry -->
<entry typeCode="COMP">
    <templateId root="2.16.840.1.113883.3.445.4"/>
    <act classCode="ACT" moodCode="DEF">
        <templateId root="2.16.840.1.113883.3.445.5"/>
    </act>
    <!-- Purpose of use -->
    <code code="RESEARCH" codeSystem="2.16.840.1.113883.3.18.7.1" codeSystemName="nhin-purpose" displayName="Uses and disclosures for research purposes"/>
    .......... 
    <!-- Action -->
    <entryRelationship typeCode="COMP" contextConductionInd="true">
        <templateId root="2.16.840.1.113883.3.445.8"/>
        <observation classCode="OBS" moodCode="DEF" negationInd="true">
            <code code="IDISCL" codeSystem="2.16.840.1.113883.5.4" displayName="Information disclosure" codeSystemName="ActConsentType"/>
        </observation>
    </entryRelationship>
    <!-- Information References: category, object id, sensitivity, related problem -->
    <entryRelationship typecode="COMP" contextConductionInd="true">
        <templateId root="2.16.840.1.113883.3.445.9"/>
    </entryRelationship>
</entry>
Consent CDA[7]: Research opt-out

.........
<!-- Privacy Consent Directive Entry -->
<entry typeCode="COMP"/>
<templateId root="2.16.840.1.113883.3.445.4"/>
<act classCode="ACT" moodCode="DEF">
<templateId root="2.16.840.1.113883.3.445.5"/>
<!-- Purpose of use -->
<code code="RESEARCH" codeSystem="2.16.840.1.113883.3.18.7.1" codeSystemName="nhin-purpose" displayName="Uses and disclosures for research purposes"/>

.........
<!-- Action -->
<entryRelationship typeCode="COMP" contextConductionInd="true">
<templateId root="2.16.840.1.113883.3.445.8"/>
<observation classCode="OBS" moodCode="DEF" negationInd="true">
<code code="IDISCL" codeSystem="2.16.840.1.113883.5.4" displayName="Information disclosure" codeSystemName="ActConsentType"/>
</observation>
</entryRelationship>
<!-- Information References: category, object id, sensitivity, related problem -->
<entryRelationship typecode="COMP" contextConductionInd="true">
<templateId root="2.16.840.1.113883.3.445.9"/>

.........
De-identification of patient data
Identifying Patients Health Data[8]

• Direct identifiers:
  – Variables that directly identify an individual
  – Names, telephone numbers, Social Security Numbers etc.
  – Not useful for analytics

• Indirect identifiers (quasi-identifiers):
  – Gender, age, locations (e.g., postal code), profession etc.
  – Useful for analytics

• Health data
  – Diagnosis, medication, dates
  – Useful for analytics
  – May re-identify patients
Masking*

• Common techniques:
  – **Suppression**: removes fields
  – **Randomization**: replaces fields with random fake values
  – **Shuffling**: shuffles real values
  – **Hashing**: pseudonyms via one-way hashing

• Significantly reduces utility of the data

• Typically applied only to **direct identifiers**

* Not to be confused with Masking within the Security Labeling Service (SLS) module
Safe Harbor

- HIPAA approved de-identification technique
- Masks 16 direct identifiers (e.g., names, email address, SSN etc.)
- Generalizes two quasi-identifiers:
  - First three digits of postal code when population size > 20 000, otherwise postal code is 000
  - Dates related to an individual (e.g., birth date, admission date, discharge date) represented only by year
- Pros:
  - Simple
  - The most common way to de-identify patient data
- Cons:
  - High risk of re-identification
  - Other fields, such as health data (e.g. diagnosis, number of visits etc.) and quasi-identifiers can be used to re-identify a patient
De-identification: statistical method

Maintain analytical value:

- Mask direct identifiers
- Generalize other fields
  - when risk of re-identification greater than a certain threshold

Re-identification risk threshold depends on the risk of exposure

- Public vs internal use of the dataset
Privacy vs. data utility

Assessing re-identification risk: example

- **Equivalence classes (similar records)**

<table>
<thead>
<tr>
<th>Patient gender</th>
<th>Handedness</th>
<th># of records</th>
<th>Re-identification risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>L</td>
<td>2 records</td>
<td>.5</td>
</tr>
<tr>
<td>F</td>
<td>R</td>
<td>3 records</td>
<td>.33</td>
</tr>
<tr>
<td>F</td>
<td>L</td>
<td>3 records</td>
<td>.33</td>
</tr>
<tr>
<td>M</td>
<td>L</td>
<td>2 records</td>
<td>.5</td>
</tr>
<tr>
<td>F</td>
<td>R</td>
<td>3 records</td>
<td>.33</td>
</tr>
<tr>
<td>M</td>
<td>L</td>
<td>3 records</td>
<td>.33</td>
</tr>
<tr>
<td>M</td>
<td>R</td>
<td>2 records</td>
<td>.5</td>
</tr>
<tr>
<td>F</td>
<td>R</td>
<td>3 records</td>
<td>.33</td>
</tr>
</tbody>
</table>

- **Threshold:** \( \tau = .4 \)
- **Above threshold:** \( 2 \ (F-L) + 2 \ (M-R) = 4 \) records
- **Conclusion:**
  40% of all records has a re-identification risk above threshold
HCS for analytics on healthcare data
1. HL7 RIM Database (MGRID*)
2. Access Control Service (XACML)
3. Security Labeling Service
4. Obligation Service
5. Access policies
6. Response

Labeling rules

Request

*) www.mgrid.net
1: Request:
- User/requester
- Purpose of use
- Query

Example:
- John Smith wants to do research and asks for demographic data of all female diabetes patients
2: Query

Example:
- *demographic data of all female diabetes patients*
3: Query result:
- Tagged data

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna Anderson</td>
<td>1966-12-11</td>
</tr>
<tr>
<td>Barbara Brown</td>
<td>1939-11-15</td>
</tr>
<tr>
<td>Carol Clark</td>
<td>1965-02-02</td>
</tr>
<tr>
<td>Dorothy Davids</td>
<td>1954-05-04</td>
</tr>
<tr>
<td>Elizabeth Evans</td>
<td>1948-11-02</td>
</tr>
</tbody>
</table>
4: Labeled result:
- Tagged data
- Security label: sensitivity, obligations, purpose of use, etc.

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date of birth</th>
<th>Sensitivity: low</th>
<th>Purpose of use: treatment, research</th>
<th>Obligation: de-identify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna Anderson</td>
<td>1966-12-11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barbara Brown</td>
<td>1939-11-15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Labeled result:

Transformations applied by SLS (vs Obligation Service)

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna Anderson</td>
<td>1966-12-11</td>
</tr>
<tr>
<td>Barbara Brown</td>
<td>1939-11-15</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity: low
Purpose of use: treatment, research
Obligation: de-identify
4. Labeled result:
   - Tagged data
   - Security label: sensitivity

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna Anderson</td>
<td>1966-12-11</td>
</tr>
<tr>
<td>Barbara Brown</td>
<td>1939-11-15</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

**Expression of conditional obligations**
E.g. only de-identify when purpose of use is 'research'

Sensitivity: low
Purpose of use: treatment, research
Obligation: de-identify
5: Policies
- Legal policies
- Organizational policies
- ...
- Patient consent
6: Response
- Tagged data: de-identified

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>*******</td>
<td>1966</td>
</tr>
<tr>
<td>*******</td>
<td>1939</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
6: Response
• Tagged data: de-identified

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>*******</td>
<td>1966</td>
</tr>
<tr>
<td>*******</td>
<td>1939</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Taking into account risk of exposure
Conclusions and future work

HCS seems adequate for our data analytics use cases
- Access control for primary and secondary use
- Security label vocabulary
- Consent CDA

Next steps:
- Design and implement our SLS
- Access control mechanism
References

1. FIEEC, ASIP Sante’. “Lessons learned from the FIEEC/ASIP study on telemedicine and telehealth”, March 2011.

2. http://axleproject.eu


5. Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). January 2012.

6. Act on the Treatment Contract (art. 7.458 BW) and the Data Protection Act (art. 23.2) combined.

