Security and Privacy of P4 Medicine: Challenges and Possible Solutions

Jean-Pierre Hubaux

With gratitude to the many biomed and CS researchers with whom I have been fortunate to collaborate on this topic
Growing Concern: Medical Data Breaches

Around 1 declared breach per day, each affecting 500+ people

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

<table>
<thead>
<tr>
<th>Name of Covered Entity</th>
<th>State</th>
<th>Type of Breach</th>
<th>Individuals Affected</th>
<th>Date of Breach Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Central Kansas Area Agency on Aging</td>
<td>KS</td>
<td>Business Associate</td>
<td>8750</td>
<td>10/1/2017</td>
</tr>
<tr>
<td>Texas Children's Health Plan</td>
<td>TX</td>
<td>Health Plan</td>
<td>922</td>
<td>10/27/2017</td>
</tr>
<tr>
<td>Catholic Charities of the Diocese of Albany</td>
<td>NY</td>
<td>Healthcare Provider</td>
<td>4024</td>
<td>10/27/2017</td>
</tr>
<tr>
<td>Arvada Dental, LLC - Drs. Bailey and Donahue</td>
<td>CO</td>
<td>Healthcare Provider</td>
<td>1716</td>
<td>10/26/2017</td>
</tr>
<tr>
<td>MSA Home Healthcare Colorado, Inc.</td>
<td>CO</td>
<td>Healthcare Provider</td>
<td>2098</td>
<td>10/25/2017</td>
</tr>
<tr>
<td>TJ Samson Community Hospital</td>
<td>KY</td>
<td>Healthcare Provider</td>
<td>803</td>
<td>10/24/2017</td>
</tr>
<tr>
<td>Brevard Physician Associates</td>
<td>FL</td>
<td>Healthcare Provider</td>
<td>7078</td>
<td>10/24/2017</td>
</tr>
<tr>
<td>Aetna, Inc.</td>
<td>CT</td>
<td>Health Plan</td>
<td>1980</td>
<td>10/23/2017</td>
</tr>
<tr>
<td>Recovery Institute of the South East P.A.</td>
<td>FL</td>
<td>Healthcare Provider</td>
<td>669</td>
<td>10/21/2017</td>
</tr>
</tbody>
</table>
“WannaCry” Ransomware Virus (May 2017)

Do state institutions have the resources to fight hackers?

Public sector has lessons to learn as hospital trusts and GPs struggle to recover from ransomware attack

![WannaCry Ransomware Virus](image)

A ransomware attack brought computers to a standstill across the world on Friday. Photograph: Ritchie B. Tongo/EPA

The Guardian, 14 May 2017
Ransomware Attack against German Hospitals

20. März 2016, 10:05 Uhr  Klinikum Neuss

Wenn Cyberkriminelle ein Krankenhaus lahmlegen
Another Major Concern: Re-identification Attacks against Genomic Databases
Security / Privacy Requirements for Personalized Health

• Pragmatic approach, *gradual* introduction of new protection tools
• Different *sensitivity levels* of the data
• Different *access rights*
• Exploit *existing* data (electronic health records) and tools
• Be *future-proof* (no short-sighted “bricolage”)
• Awareness of *patient consent*
• Secure also the *collection* of health data (via smartphones, wearable sensors,...)
Privacy-Enhancing Technologies

Two main approaches:

- Protect the data themselves:
  - Use of **cryptography**
  - Symmetric / asymmetric encryption
  - Property-preserving encryption
  - (Partially) homomorphic encryption
  - ...

- Avoid that responses leak “too much” information:
  - Provide only **global**
  - (e.g., statistical) **results**
    - K-anonymity, l-diversity, t-closeness
    - Differential privacy
    - For genomics, see “Homer attack” and subsequent ones
MedCo: System and Threat Models

Honest-but-curious adversary:
- honestly follows the protocol
- tries to infer sensitive data from the different steps of the protocol

Malicious-but-covert adversary:
- can tamper with the protocol
- tries to infer sensitive data from the query end-result
Main Concerns

• **Attribute disclosure** due to illegitimate access to the data
  o External (hacker) or internal (insider) attacker stealing the data

  ➔ Standard encryption can protect data ONLY at rest or in transit BUT NOT during processing (e.g., in the memory)

• **Patient re-identification** due to legitimate access to the data
  o Malicious users performing “smart” data requests in order to re-identify patients in a specific dataset (e.g., patients with HIV)

  ➔ De-identification or anonymization is ineffective with genomic data
Main Requirements

Functionality:

\[ \text{COUNT}(\text{patients})/\text{SELECT}(\text{patients}) \]
\[ \text{FROM database} \]
\[ \text{WHERE } * \text{ AND/OR } * \]
\[ \text{GROUP BY } * \]

* represents any possible concepts in the ontology

Security/Privacy:

- Protection of data confidentiality at rest, in transit and during computation
- no single point of failure
- only the investigator can obtain the query end-result
- (optional) unlinkability
- (optional) differential privacy
MedCo: Combining the Best of Both Worlds

Biomedical Informatics:

• Data model from *i2b2* *(Informatics for Integrating Biology and the Bedside)*

• Interoperability layer from *SHRINE*

IT Privacy and Security:

• Privacy-preserving distributed protocols from *UnLynx*
Use Case: Tests on Clinical Oncology

Public Data from cBioPortal

- 121 patients (later scaled to 121,000) with 9 clinical attributes and 1,978 mutations on average per site and patient

- **Query 1**: “Number of patients with skin cutaneous melanoma AND a mutation in BRAF gene affecting the protein at position 600.”
  ➔ (2 clinical attributes, 4 mutations)

- **Query 2**: “Number of patients with skin cutaneous melanoma AND a mutation in BRAF gene AND a mutation in (PTEN OR CDKN2A OR MAP2K1 OR MAP2K2 genes)”
  ➔ (2 clinical attributes, 77 mutations)

Hardware and Software Setting

- 3 servers: 2.5GHz Intel Xeon E5-2680 v3 CPUs with 12 cores

- memory: 256GB of RAM

- network: 10 Gbps link

- OS: Ubuntu

- crypto: ElGamal encryption on Ed25519 elliptic curve with 128 bit security

- database: PostgreSQL

- deployment technology: Docker
MedCo: Core Architecture & Protocol

MedCo Protocol:
A, B) ETL & Encryption Phase
1) (user) Query Generation
2) (local) Query Analysis & (distributed) Tagging
3) (local) Query Processing
4) (local) Result Aggregation
5) (local) Result Obfuscation
6) (distributed) Results Shuffling
7) (distributed) Results Re-Encryption
8) (user) Result Decryption

ONT: Ontology
ETL: Extract, Transform, Load
CRC: Data Repository (Clinical Research Chart)
Aggr: Broadcast and Aggregator (hub)
Performance Results:
Query Runtime vs. Database Size

**Query 1**

<table>
<thead>
<tr>
<th>Size of the database</th>
<th>Insecure i2b2</th>
<th>MedCo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x</td>
<td>1.8</td>
<td>2.4</td>
</tr>
<tr>
<td>2x</td>
<td>1.9</td>
<td>2.4</td>
</tr>
<tr>
<td>4x</td>
<td>2.0</td>
<td>2.6</td>
</tr>
</tbody>
</table>

**Query 2**

<table>
<thead>
<tr>
<th>Size of the database</th>
<th>Insecure i2b2</th>
<th>MedCo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x</td>
<td>6.1</td>
<td>6.7</td>
</tr>
<tr>
<td>2x</td>
<td>6.4</td>
<td>7.1</td>
</tr>
<tr>
<td>4x</td>
<td>6.6</td>
<td>7.3</td>
</tr>
</tbody>
</table>
User Experience

• MedCO is transparent for the investigator (through a SHRINE/i2b2 frontend plugin)
Demo of MedCo
DPPH: Data Protection in Personalized Health

• P4 (Predictive, Preventive, Personalized and Participatory) medicine
  – Revolutionize healthcare by providing better diagnoses and targeted preventive and therapeutic measures

• Challenges:
  – Scalability/Big Data
  – Efficiency and usability in data sharing
  – Mitigating privacy risks and complying with data protection

• Centralized vs Distributed Approaches:
DPPH: Key Facts

- 5 research groups across the ETH domain + SDSC (Swiss Data Science Center)
- Funding: 3 Millions CHFRs
- Duration: 3 years (4/2018 - 3/2021)
- Funding Program: ETH PHRT (Personalized Health and Related Technologies)
- Related event on February 15th, 2018
  - Workshop on Secure, Privacy-Conscious Data Sharing
  - [http://dpph18.epfl.ch](http://dpph18.epfl.ch)

**Project goals:**
- Address the main privacy, security, scalability, and ethical challenges of data sharing for enabling effective P4 medicine
- Define an optimal balance between usability, scalability and data protection
- Deploy an appropriate set of computing tools
DDPH: Key Enablers

The integration of the shown tools and frameworks provides crucial benefits for medical research.

- **Scalable distributed scientific computing infrastructure**
  - Scalability and Reproducibility

- **Ease querying and aggregating distributed medical data**
  - Accessibility and Usability

- **Secure and privacy-conscious data sharing and processing for medical data**
  - Privacy and Accountability

- **Robust support for secure operations on distributed data**
  - Security and Access Control

- **Framework for inference risks and countermeasures, and ethical analysis of distributed platforms for medical data sharing**
  - Legal and Ethical compliance
Envisioned Nation-Wide Deployment

Integration with I2b2 and TranSMART

Events on Genome Privacy and Security

- **Dagstuhl** seminars on genome privacy and security 2013, 2015
- **Conference on Genome and Patient Privacy (GaPP)**
  - March 2016, Stanford School of Medicine
- **GenoPri**: International Workshop on Genome Privacy and Security
  - July 2014: Amsterdam (co-located with PETS)
  - May 2015: San Jose (co-located with IEEE S&P)
  - November 12, 2016: Chicago (co-located with AMIA)
  - October 15, 2017: Orlando (co-located with Am. Society for Human Genetics (ASHG) and GA4GH)
- **iDash**: integrating Data for Analysis, Anonymization and sHaring (already in previous years)
  - October 14, 2017: Orlando
- **Inst. For Pure and Applied Mathematics (IPAM, UCLA)**
  - Algorithmic Challenges in Protecting Privacy for Biomed Data
    10-12 January, 2018
- Workshop at EPFL, Switzerland, February 15, 2018

⇒ Lots of material online
“genomeprivacy.org”

Community website

- Searchable list of publications on genome privacy and security
- News from major media (from Science, Nature, GenomeWeb, etc.)
- Research groups and companies involved
- Tutorial and tools
- Events (past & future)
Conclusion

• Worldwide, medical confidentiality is **in jeopardy**
• P4 medicine requires collecting and sharing **many more data**
• Presence of **genomic data** and health-data collection with **wearable devices** will further increase the risk
• Several solutions, including **advanced cryptography**, are usable to protect genomic (and more generally medical) data
• We are working on **fully decentralized tools** (UnLynx, MedCo)
• We have **operational prototypes**, currently in deployment phase (at Lausanne University Hospital (CHUV))
• We aim at **nation-wide** deployments
• There is a tremendous need for **standardization**, especially for multi-site studies
• Our contributions to the topic: http://lca.epfl.ch/projects/genomic-privacy/