Distributed Analytics on Sensitive Medical Data

Oya Beyan
Fraunhofer Institute for Applied Information Technology (FIT)
RWTH Aachen University, Informatik 5, Germany
Personal Health Train Implementation Network

An open ecosystem for innovation

PHT is a distributed infrastructure to support:

- data management
- data analysis
- medical decision making

Empowers citizens / public and private organisations:

- Gain control over their own data
- Manage, safeguard, and share their data
- Partner with researchers, healthcare providers

enable the reuse of health data for the benefit of individuals and society
Guiding Principles:

- Control over data
- Reusable health data
- Distributed and federated solutions
- Responsible use of health data
- Ethics-by-design
- Machine-readability at the core

Goals:

- a common core infrastructure
- set of standards, guidelines, specifications
- reference implementations
Current state of art in data sharing

Why current data sharing practice is not sufficient to support data driven medicine?

Perspectives:

- Citizens / Care Givers are eager to use data analytics services
- Researchers / Data Scientist demand more data to develop analytics models
Data for Predictive, Preventive and Personalized medicine

Cleveland Clinic

SUSPECTED COLORECTAL CANCER

For Patients Without History of Colorectal Cancer, Inflammatory Bowel Disease, Familial Adenomatous Polyposis, Lynch Syndrome, or Adenomatous Polyp

Predicts 10-year colorectal cancer risk.

EPILEPSY

For Patients Eligible For Epilepsy Surgery

Predicts 2-year and 5-year probabilities of freedom from seizure recurrence and probabilities of obtaining an Engel Score of 1.

FEMALE PELVIC MEDICINE AND RECONSTRUCTIVE SURGERY

For Women with Urinary Incontinence Considering Midurethral Sling Surgery

Predicts risk of any adverse event, bothersome stress, urge incontinence or positive stress test with 12 months of surgery.

http://riskcalc.org/
Data for Predictive, Preventive and Personalized medicine

### UR-CHOICE Pelvic Floor Disorders Risk Calculator

<table>
<thead>
<tr>
<th>Risk Factors</th>
<th>12-Year Risk for Women with One or More Previous Births.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pelvic Organ Prolapse</td>
<td>Vaginal: 13% Any: 8%,Bothersome: 3%,Treatment: 19%,Bothersome or Treatment: 9%</td>
</tr>
<tr>
<td></td>
<td>C-Section: 13% Any: 8%,Bothersome: 4%,Treatment: 19%,Bothersome or Treatment: 7%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>Vaginal: 40% Any: 22%,Bothersome: &lt;0.5%,Treatment: 25%,Bothersome or Treatment: 28%</td>
</tr>
<tr>
<td></td>
<td>C-Section: 30% Any: 22%,Bothersome: &lt;0.5%,Treatment: 25%,Bothersome or Treatment: 24%</td>
</tr>
<tr>
<td>Fecal Incontinence</td>
<td>Vaginal: 23% Any: &gt;10%,Bothersome: 1%,Treatment: 6%,Bothersome or Treatment: 5%</td>
</tr>
<tr>
<td></td>
<td>C-Section: 23% Any: &gt;10%,Bothersome: 4%,Treatment: 6%,Bothersome or Treatment: 5%</td>
</tr>
<tr>
<td>Any Pelvic Floor Disorder</td>
<td>Vaginal: 49% Any: 28%,Bothersome: 1%,Treatment: 21%,Bothersome or Treatment: 37%</td>
</tr>
<tr>
<td></td>
<td>C-Section: 41% Any: 28%,Bothersome: 1%,Treatment: 16%,Bothersome or Treatment: 32%</td>
</tr>
<tr>
<td>Two or More Pelvic Floor Disorders</td>
<td>Vaginal: 16% Any: 6%,Bothersome: &lt;0.5%,Treatment: 7%,Bothersome or Treatment: 5%</td>
</tr>
<tr>
<td></td>
<td>C-Section: 16% Any: 6%,Bothersome: &lt;0.5%,Treatment: 7%,Bothersome or Treatment: 5%</td>
</tr>
</tbody>
</table>
Data for Predictive, Preventive and Personalized medicine

Data Analytics Services

What happens to your data?

- Computing power is required to execute models
- Data is transferred mostly to a cloud service
- You have granted some rights to service providers but you are not sure about consequences
- You are not part of the data governance anymore

Oya Beyan, GO FAIR Personal Health Train Implementation Network German Chapter Workshop, 12th February 2019, TMF, Berlin
Sharing Data to Develop Analytical Models
Personal Health Train Approach

*bring algorithms to the data*

*...rather than bringing all data to a central place*
Sharing Algorithms to Develop Analytical Models
PHT Approach

- Persistently identifiable, auditable trains
- Traceability and provenance
- Reproducibility
- Standardized interfaces
PHT Approach

Data Consumers

- Intends to access privacy-sensitive data from multiple curators
- formulates queries to describe data requirements
- specifies the analytics algorithm
- executes it remotely

I designed a great algorithm to predict heart diseases but need data to train it.

What is the average weight of patients diagnosed with cardiovascular diseases?

Are there any patients aged 65+ who use beta-blockers in combination with verapamil?

TOOLS & SERVICES

- Metadata discovery
- Model repositories
- Metadata definition
- Shipping request
PHT Approach

Trains

Adaptable, reproducible, identifiable, versioned, self containing

```python
import ml_lib as ml

def main():
    model = ml.load('/model1')
    model.fit_transform('/data')
    model.save('/model2')
```

metadata.rdf

Algorithm.py

query.{cql/sparql}
PHT Approach

Data Stations

- Provides computing resources
- Integrates data from multiple data sources
- Maintains private data repositories
- Publishes schemas/metadata describing available data sets
- Acts as FAIR data points
PHT Approach

Data Stations: Execution of Tasks

- Data Integration Engine
  - aggregates associated Data Banks
  - exposes data in a standard (FHIR, RDF, ..)
  - evaluates Train’s data queries (CQL, SPARQL ...)

- Docker execution engine
  - Executes analytic tasks in secure enclave

- Permission validation
**PHT Approach**

**Handling Station**

- Acts as a gateway between consumer and the curators
  - Broker between parties
  - Evaluates & monitors trains
  - Facilitates train routing
  - Registry & Indexing
  - Maintains metadata
  - Aggregates results

- Centralized point of trust or network
PHT implementations

Best practices
### PHT implementations

#### First Implementation: euroCAT Distributed Learning for Personalized Medicine, 2016

**Aim:** predicts dyspnea (a common side effect after radiotherapy)

**Data:**
287 lung cancer patients, treated with curative intent with chemoradiation or radiotherapy
- MAASTRO (Netherlands, Dutch)
- Jessa (Belgium, Dutch)
- Liege (Belgium, Dutch and French)
- Aachen (Germany, German)
- Eindhoven (Netherlands, Dutch)

**Learning:**
Bayesian network model
PHT implementations

Varian Learning Portal

distributed machine learning infrastructure
two self-contained components: Master & Site

http://vlpdashboard.jvsoest.eu/run/21363
PHT implementations

Phenotyping Algorithms Prototype, SMITH Project 2018

![Phenotype Definition](image1.png)

![Monitor](image2.png)
How does it work?

Main concepts and architecture
PHT Main Concepts and Architecture

Training Phase

Discovery Phase

Consumer

Handler

Curator

Oya Beyan, GO FAIR Personal Health Train Implementation Network German Chapter Workshop, 12th February 2019, TMF, Berlin
PHT Main Concepts and Architecture

- Sent during computation phase
- Study Description, Authentication, Buildfile, Version, Provenance (Stations visited, Evaluation, ...)
- Data Extraction and Cleaning
- Machine Learning, Data Mining, Record Linkage, ...
- Combine multiple results from distributed curators
- Metadata
  - Train PID, Query Type, ...
- Query
  - CQL/SPARQL
- Analytics
  - Docker Container
- Aggregator
  - Federated Learning

Detached at the Handling Station
PHT Main Concepts and Architecture
How to move forward?

PHT Implementation Network
GoFAIR Implementation Networks

Added value

- Joint implementation based on FAIR Principles for research data, algorithms, processes, software etc.

- Cooperation stimulates convergence and global interoperability

- Further fragmentation will be prevented

- Opportunity to help shape the Internet of Fair Data and Services
Working Groups

Session A: Strategic Vision

- Control over data
- Reusable health data
- Responsible use of health data
- Ethics-by-design
- Distributed and federated solutions
- Machine-readability at the core
Working Groups

Session C: Use Cases

- Elixir: a distributed infrastructure to share and store bioinformatics resources
- FAIR4Health Project
- Rare Diseases Communities
- National Research Data Infrastructures
- MII Projects: Cross border data reuse between Data Integration Centers
- ........
We are working together

........... be part of it!

Please contact us:

Oya Beyan
beyan@fit.fraunhofer.de

Fraunhofer Institute for Applied Technology FIT,

Schloss Birlinghoven, Konrad-Adenauer-Straße, 53754 Sankt Augustin, Germany