UiT

THE ARCTIC UNIVERSITY OF NORWAY

Towards privacy preserving comparative effectiveness research

Kassaye Y. Yigzaw

Johan Gustav Bellika

Anders Andersen

Gunnar Hartvigsen

HelseIT 2013, Trondheim

Overview

- Motivation
- Comparative effectiveness research
- Barriers
- Identifiable data
- Deidentified data
- Secure multi-party computation
- Discussion

Motivation

- Demography change (i.e. aging population, multiple chronic conditions)
- Infectious diseases
- Health care system is under serious challenges

Motivation (2)

- An increased use of electronic health records (EHRs)
- Detail and diversity of healthcare and biomedical data is collected
- Health care systems´ effectiveness and efficiencies
- Patient outcomes and safety

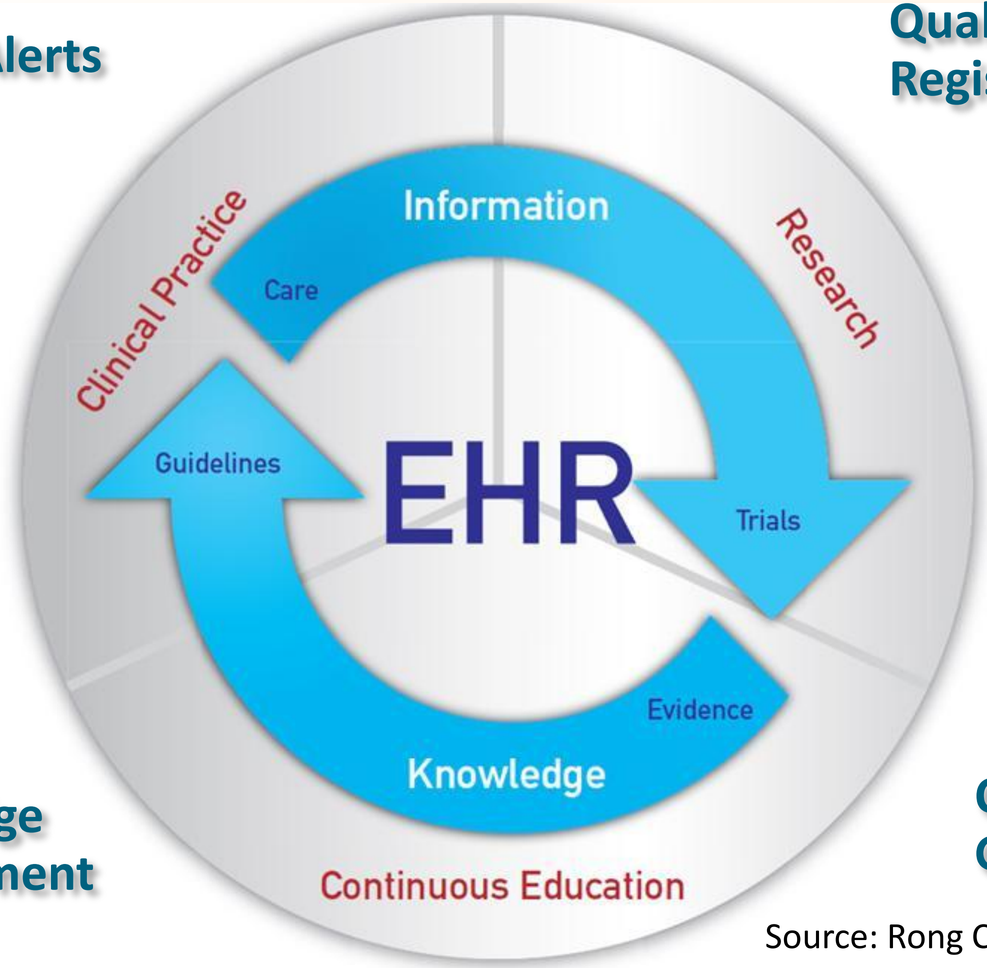
Knowledge generation and use in medicine

Alerts

Quality Registries

Clinical Decision Support

Clinical Trials



Order sets

Compliance Checking

Knowledge Management

Computerized Guidelines

Source: Rong Chen, MedInfo2013 tutorial on OpenEHR Guideline Definition Language

Comparative Effectiveness Research

- Generate evidence on the effectiveness, benefits, and harms of different treatment options in real life
- Study designs: systematic reviews of existing studies, RCTs, and observational data analyses
- Observational studies use existing data sources

Comparative Effectiveness Research (2)

- Lab test result, treatment and outcome, outpatient visits, hospitalization, primary care visits, pharmacy, and/or other information
- Patients receive care from multiple institutions
- Strong statistical power
- Population heterogeneity
- Horizontally and vertically partitioned dataset
- Link data distributed across multiple institutions

What is the problem?


Privacy

Objective

Enjoy the benefits of both the privacy and research worlds!

Identifiable Data

- Use of identifiable data requires individuals´ consent
- Except under limited circumstances
- Difficult to obtain consent from some patients, such as severely ill, demented and pediatric patients
- Often, it is not practical to collect consent (i.e. large study size)

Identifiable Data (2)

- Consenter Vs. non-consenter difference
  - Demographic and
  - Socio-economic characteristics
- Biased samples

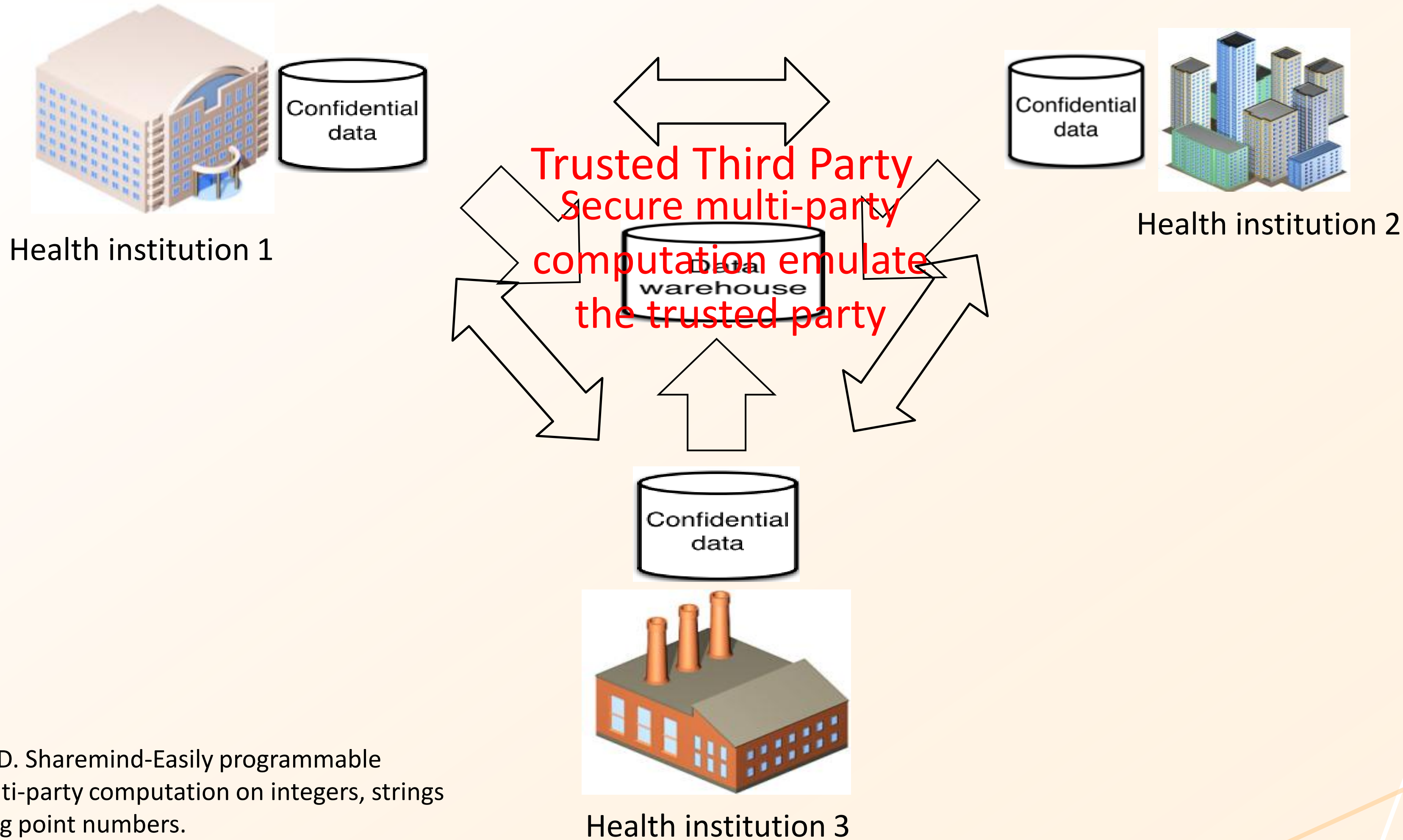
De-identified Data

- De-identified data can be used for research
- Health data can be deidentified:
  - Removing identifiers (e.g. Safe harbor and limited dataset)
  - Statistical methods
- The HIPAA safe harbor method involves removal of 18 identifiers including biometric or genetic data
- Limited dataset removes 16 identifiers (except date and zip code) and obtain data use agreement

De-identified Data (2)

- ↑De-identification ≈ ↓data usefulness ≈ ↓re-identification
- Causal relationship between events
- Link data from multiple source to individual record
- Sub-populations level study

Secure multi-party computation (SMC)



Bogdanov D. Sharemind-Easily programmable secure multi-party computation on integers, strings and floating point numbers.

Secure multi-party computation (2)

- A set of two or more parties with private inputs, $x_1,..,x_n$ wish to jointly compute a function, $f(x_1,..,x_n)$, of their inputs
- Parties wish to preserve some security properties. E.g. privacy and correctness.
- Even in the face of adversarial behavior by some of the participants, or by an external party.

Yehuda Lindell. Presentation "Tutorial on Secure Multi-Party Computation". IBM T.J.Watson

History

- Introduced by Yao in 1982 (two-party computation)
- Goldreich et al. in 1987 (Multi-party computation)
- No practical implementation until the last decade

SMC techniques

- Generic techniques (i.e. Garbled circuit, Homomorphic encryption, Secret sharing)
- Specialized techniques (i.e. secure sum, scalar product)

SMC protocols

- All to all communication
- Representative based approach
- Considered not efficient and not scalable to hundreds and thousands of distributed data sources

Distributed SMC

- Decompose the computation problem in a way that can be computed by neighbor peers in parallel
- A peer only jointly compute with neighbors
- ONLY combined statistics of neighbor peers´ private data will be learned
- Reasonable to hide private data in combined statistics of neighbor peers

Distributed SMC (2)

- Constant communication complexity
- Enable parallel computations
- Execute asynchronous algorithms

- Hypothesis:

“Distributed SMC enables more efficient and scalable solutions.“

Discussion

- Data sources maintain autonomy over their record
- No new information can be discovered after a computation
- Preserve patients´ and data owners´ privacy
- Increased data owners motivation to participate

Reference

- “Towards Privacy-Preserving Computing on Distributed Electronic Health Record Data” Middleware 2013 (submitted)

Acknowledgement

- Gro Berntsen, Norwegian Center for Integrated care and Telemedicine, University Hospital North Norway
- Tromsø Telemedicine Laboratory (TTL)
- University of Tromsø
- Norwegian center for integrated care and telemedicine (NST)

Thank you!

Contact Information
Kassaye Y. Yigzaw
PhD student
University of Tromsø
kassaye.y.yigzaw@uit.no
+4796747253