Bringing it all together: How to join and analyze sensitive data from multiple sources

Florian Endel$^1,2$

$^1$IMEHPS.research, florian@endel.at
$^2$Vienna University of Technology

Farr Conference,
27 August 2015,
St Andrews, Scotland
Setting & Objective

- data everywhere
  - routinely collected claims data
  - administrative data & registries

- linking, merging & integrating: bring it together!
  - technology available and in development
  - public benefit and interest possible
  - funding achievable

- domain specific characteristics of
  - data collections
  - patients, healthcare systems ("data generating processes")
  - research question & goals
  - legislation & privacy concerns

- 4 examples from Austria
Background: health insurance in Austria

- ~ 8.5 million inhabitants
- mandatory insurance
  - (nearly) everyone has to be insured
  - by a specific insurance provider
  - depending on occupation, region
- 19 social security institution (insurers)
- grown historically
- organized differently
- differing payment systems, legislation, scopes
- unique personal identifier (UPI) available
- separate data collections
1) Different information, similar sources

- Routinely collected claims data from these social security institutions:
  - Primary care
  - Specialized outpatient care
  - Prescriptions

- Including
  - Personal information from patients and providers
  - Accounting details
  - ~95% of Austrian population

- Not allowed to be linked due to privacy concerns

- Objective: Merge these sources
1) database “GAP-DRG”

- selected subset (tables, variables)
- censored details: e.g. year of birth
- encrypted UPI: pseudonymization
- linked by common pseudonym(s)
  - linkage on personal level:
    - data about same patient from different sources
  - integration without direct relationship

- nevertheless, issues occurred, e.g.:
  - definition of a standardized data model
  - quality issues depending on source
  - patients with multiple insurances
  - harmonization of accounting and coding systems
  - mix of various systems in same database
  - differing commitments of data providers

- limitation: data from 2006 + 2007; subsets from 2008-2011
Background: inpatient care in Austria

- organized differently in comparison to outpatient services
- not paid directly and only by insurance institutions
- DRG (related) payment system
- data collection
  - centralized
  - cleaned
  - used routinely
- no UPI: only episodes\(^1\)
- limited personal information

\(^1\)2015: change of legislation; availability of UPI yet unknown
2) similar information, different sources

- objective: integrate inpatient data into “GAP-DRG”
- details: e.g. poster 1401, SHIP conference 2011, 2013
- short story:
  - insurance institutions get limited information on hospital episodes
  - differences between sources remain
  - deterministic record linkage
  - by utilizing
    - varying personal details
    - episode specifics
- whole procedure fitted to various peculiarities of
  - data (quality)
  - reporting systems
- result: “GAP-DRG” covers all sectors of the healthcare system
3) data from different domains

- **objective:** link GAP-DRG with data from other domains
  - example project: unemployment data

- **both databases hold pseudonymized UPIs**
  - from the same identification number
  - encrypted in different ways

- **challenges & requirements**
  - do not link sources (healthcare & unemployment data) directly
  - establish link between pseudonymized databases
  - do not leak any previously unknown information to anyone during linkage
  - prepare link but transfer data later
  - do not break encryption
  - get everyone to agree on procedure
3) data from different domains: flowchart
4) international cooperation

- EU FP7 project CEPHOS-LINK
- cooperation of several countries...
- ... differing in many aspects
- objective 1: perform comparable data analysis in parallel
- objective 2: perform pooled data analysis
  - to include system specific effects
  - compare countries in more detail
- cross-border data transfer
- not “record linkage” on personal level
- alternative: distributed data analysis \(^2\)
  - e.g. stratified Cox regression model

4) cross-border data transfer

- very hard to get it right
- privacy
  - approvals from ethical committees and data owners
  - anonymisation and secure data handling
- anonymisation
  - k-anonymity (l-diversity)
  - disclosure risk estimation
  - huge loss of information but multiple datasets possible
- data transfer
  - asymmetric cryptography
  - secure research environment
- harmonized variable definition
  - same structure and formatting
  - same content and meaning
  - the most complicated part!
Conclusion

- various applications of “bringing data together”
- huge spectrum of possibilities
- generalizeable methods but very specific utilization
- depending on e.g.
  - data sources and providers
  - “history” of data: from generation to researcher
  - purpose of data and project
  - objective and research question
Bringing it all together: How to join and analyze sensitive data from multiple sources

Florian Endel\textsuperscript{1,2}

\textsuperscript{1}IMEHPS.research, florian@endel.at
\textsuperscript{2}Vienna University of Technology

Farr Conference,
27 August 2015,
St Andrews, Scotland