Plausible Adversaries in Re-Identification Risk Assessment

Lukasz Kniola
Biogen, Maidenhead, UK
The scope of this presentation is to present the **opinions and suggestions of the author**. The interpretations of standards and procedures presented are those of the author and are not necessarily correct. Any views, conclusions and recommendations are those of the author, and they do not represent the positions of their employer.
Agenda

- Context
- Methodology
- Survey
- Trends and Outlooks
Adversaries

Possible

Plausible
Context Of Data Release

**Internal Secondary Research**
(Data Re-use)

**External Secondary Research**

**Public Release**
Classification (1/2)

Direct Relation

Family
Acquaintances
Doctors
Employers

Lawyers
Insurance Companies
Journalists

Academic Research
Regulatory Bodies
Other Sponsors

Hackers
(Demostration Attacks)

Patients Themselves

No Relation
Classification (2/2)

Adversary’s Background Knowledge

**Likely** / **Unlikely**

**Target of Re-id:**

**Specific** / **Random**

Adversary Knows Target is in Data Set?

**Definitely** / **Doubtfully**
Prosecutor

Journalist

Marketer
Prosecutor

Matching Person To Record
Journalist

Matching Person To Record?

Matching Record To Person
Prosecutor ≥ Journalist ≥ Marketer
Maximum and Average risk

\[ \Sigma = 0.074 \]
Uniqueness in the data set
Harm ≠ Risk

Identity Disclosure vs Attribute Disclosure

Intent To Harm & Potential Harm
Survey

23 Responders

6 Questions
12 Adversary categories

Testing perceived:
Motivation
Means/Tools
Background knowledge
Intent to harm

Likelihood of knowing:
Target in the data set
Full list of trial participants
Survey

Likelihood that adversary knows target is in data set

- Patients themselves
- Family
- Doctors
- Acquaintances
- Insurance Companies
- Hackers
- Employers
- Journalists
- Lawyers
- Other sponsors
- Regulatory bodies
- Universities

Legend:
- Low
- Medium
- High
Survey

Likelihood that adversary has full list of participants

- Hackers: Low (75%), Medium (25%), High (5%)
- Doctors: Low (80%), Medium (20%), High (10%)
- Regulatory bodies: Low (75%), Medium (25%), High (10%)
- Universities: Low (80%), Medium (20%), High (10%)
- Employers: Low (80%), Medium (20%), High (10%)
- Insurance Companies: Low (80%), Medium (20%), High (10%)
- Journalists: Low (80%), Medium (20%), High (10%)
- Other sponsors: Low (80%), Medium (20%), High (10%)
- Lawyers: Low (80%), Medium (20%), High (10%)
- Acquaintances: Low (80%), Medium (20%), High (10%)
- Patients themselves: Low (80%), Medium (20%), High (10%)
- Family: Low (80%), Medium (20%), High (10%)

Legend:
- Low
- Medium
- High
Survey - Conclusions

- Hackers
- Insurance
- Journalists

- Family
- Lawyers
- Patients

- Doctors
- Employers
- Acquaintances

- OtherSpon
- Universities
- RegBodies

Journalist Risk
Maximum

Prosecutor Risk
Average

Prosecutor Risk
Average

Journalist Risk
Maximum
EMA Policy 0070

Guidance

Demonstration Attack

Participants of Public Interest

Embarrass Data Controller or Process

Acquaintance Recognition

Vulnerable Patients & Sensitive Info

Maximum Risk

Across All Records
EMA Policy 0070
Trends

Prosecutor

Easier to Defend
Easier to Apply
More Conservative
Lower Data Utility
Greater Detail Retention
Higher Data Utility

Requires:
Better Guidelines
Define "Similar Studies"
More Evidence
Thank You

Questions?

Lukasz.Kniola@Biogen.com