Anonymization methods of structured health care data: A literature review

Olga Vovk, Gunnar Piho, Peeter Ross

School of Information Technologies

Tallinn University of Technology

This work has been conducted in the project ”ICT programme” which was supported by the European Union through the European Social Fund

General info

Anonymization Methods of Structured Health Care Data:

A Literature review

- Wide scope of data in health care
- Different forms, in Electronic Health Record (diagnosis with code, medical images, doctor's notes in free text).
- Digitalization:

  Opportunities - analyze data and make data-driven decisions

  Challenges – data privacy and security
- Data anonymization – opportunity to share data while preserving privacy

Problem

- Multiple methods exist to anonymize data, however, certain methods may be vulnerable and unreliable
- Discover new methods introduced in recent years
- The paper gives an **overview of existing methods**, used for structured health data anonymization, discuss the **advantages and disadvantages**

Aim

The aim of the research is to provide a **systematic literature review** of the most recent literature **(2017 -2020** years) about the existing anonymization methods in health care

RQ1 What are the methods of data anonymization?

RQ2.What are the challenges and issues in using those methods?

Method

Guidelines Barbara Kitchenham and Stuart Charters

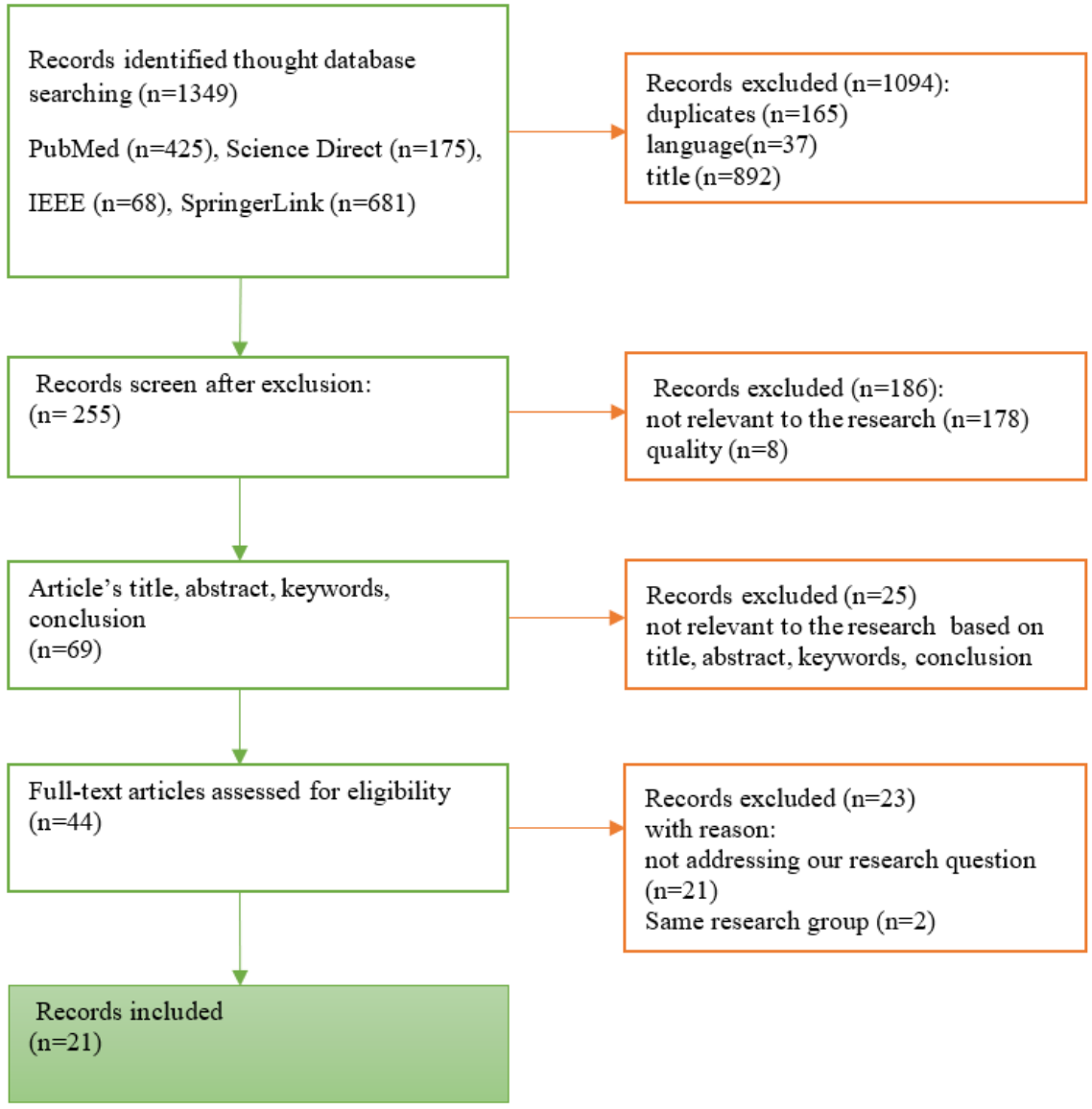
Steps:

- A. Planning the review (research interest and questions)
- B. Conducting the review (databases)
- C. Reporting the review (results)

Method

- PubMed, IEEE Explore Digital Library, Science Direct, and SpringerLink

- Records identified thought database search - **1349** records,
- Included in final review – **21** records



Results

- 33 methods identified
- 12 out of those methods are presented as the main methods in articles

Among those methods:

- k-anonymity
- l-diversity
- t-closeness
- e-differential privacy
- (k, k^m)-anonymity
- Cryptographic algorithms

Results

Identified problems

- Privacy risk and data utility
- Vulnerability to different attacks (e.g. homogeneity attack, background knowledge attack, linkage attack)
- Different methods for different types of data (micro-data, big data, transaction data)
- Trustfulness of data
- Not only technical problem
- Computational resources requirements
- Difficult to implement in real-life data

Conclusion

- All of the identified methods have their **benefits and limitation**.
- Most of the mentioned methods are built based **on generalization and suppression techniques,** cryptographic **techniques**.
- At least **7 new anonymization methods** for health data were presented in 2017-2020, although most of them are the improvement of existing methods
- This tendency shows that the field is developing and more **improved algorithms** can be expected in upcoming years.
- The main issue - finding the right **balance between data privacy and utility**
- **Cryptographic algorithms** show promising result in anonymization field, however, require certain computational power and resources to be implemented.

Future research

- Explore more databases (Scopus and Web of Science)
- Practical applicability of existing methods

THANK YOU FOR YOUR ATTENTION!
Questions?