RESEARCH DATA ALLIANCE

Health Data Interest Group @VP16

Transparency and Trust in Health Data

11th November 2020, 5:00 - 6:30 PM UTC

Co-Chairs: Edwin Morley-Fletcher, Yannis Ioannidis, Leslie McIntosh

research data sharing without barriers

rd-alliance.org

HEALTH DATA INTEREST GROUP
AGENDA

1. **Current difficulties in trustfully and lawfully sharing anonymous and pseudonymous health data ,** Edwin Morley-Fletcher, HDIG Co-Chair
2. **Special complexities in dealing with brain data**, Yannis Ioannidis (HDIG Co-chair)
3. **How to guarantee effective transparency on data used by bio-medical researchers in their publications and by pharma companies in clinical trials**, Leslie McIntosch (HDIG Co-chair)
4. **Q&A** and **open discussion on further themes in view of P17**

Transparency in healthcare is a huge issue

- 1. Quality of healthcare: transparency of providers' performance measures (quality of outcomes and processes)
- 2. Patient experience: patient perceptions of their experience and outcomes
- 3. Finance: costs transparency (DRGs) in public healthcare
- 4. Governance: open decision making, rights and responsibilities, resource allocation, accountability mechanisms.
- 5. Personal healthcare data: access, "ownership", personal data protection
- 6. Communication of healthcare data: the extent to which all the above is currently accessible, reliable, and usable by all relevant stakeholders.

Potential of AI and Big Data for healthcare innovation

- Starting from a large and varied amount of data, artificial intelligence algorithms are able to identify complex patterns of relationships that can escape human researchers
- Algorithm machines enable to automatically perform millions of operations per second
  - minimising human error
  - hugely reducing costs
  - once a rigorous logical definition of what is the problem at stake has been attained

The data-driven analytical paradigm

- Big Data contribute not only to verify theoretical hypotheses with statistical techniques, but also to explore new scenarios and derive new theories, as well as, more generally, to discover new knowledge.
- Some scholars speak of a real scientific revolution compared to the classic "hypothesis, model, experiment" approach

Unpredictability by design and black boxes

- The process of algorithmic production of knowledge is quite often "unpredictable by design"
  - being normally based on big data analytics testing large numbers of algorithms on the data
  - in view of discovering meaningful correlations, on which ML causality or DL infer
- This may produce a "black-box" effect
  - with the risk of rendering automated decision-making inscrutable or prone to hidden biases
  - though apparently functioning
- Hence the request for algorithms that respect the FAT principles
  - Fairness
  - Accountability
  - Transparency
- Hence the whole ongoing debate on AI Ethics and algorithm explainability

The inconvenient truth

- Data sharing in healthcare remains rare
- It is characterised by high transaction costs
- Happening mostly under private agreements typically enacted by large corporations.
- Although available data is continuously expanding, it largely sits idle
  - fragmented in siloes
  - carefully guarded by data controllers to reduce legal exposure.

Anonymisation

- The GDPR draws a dividing line between personal data and non-personal data
- This is paramount to determine the scope of application of the European data protection law
  - Personal data are subject to the Regulation
  - Non-personal data are not
- Anonymous data fall outside of the scope of the GDPR
- There is a discrepancy in the way anonymisation is referred to in the GDPR
  - on a risk-based approach, as “personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable by all the means reasonably likely to be used”

  and how it is defined by the European Data Protection Board
  - where “anonymisation results [only] from processing personal data in order to irreversibly prevent identification”.
- The subsequent regulatory uncertainty makes it extremely difficult to obtain anonymised data from clinical institutions
- Also because of the new heavy sanctions falling on non-compliant Data Controllers

Pseudonymisation

- Pseudonymisation is the other method of protecting privacy introduced by the GDPR,
- It relates to the processing of health data in ways by which they can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that data are not attributed to an identified or identifiable natural person.
- As re-identifiable, even encrypted data are pseudonymous.
- Given their re-identifiability, and therefore qualifying as personal data, all pseudonymous data require on principle a specific legal ground, such as an explicit personal consent, as well as the provision of specific information to the data subject, for being lawfully shared with third parties, even for research purposes.

A burdensome exercise

- The research or pharma company receiving the downloaded pseudonymous data becomes to all effects the Data Controller of those data with all implied accountability obligations
  - Transparency
  - Data retention
  - Minimization
  - Privacy impact assessment
  - Data security
  - “Right to be forgotten”
- In these conditions, the sharing of pseudonymous data likely results in a burdensome exercise

The Visiting Mode

- Pseudonymous data can be dealt with in the “visiting mode”
  - by bringing the algorithms to the data
  - without disclosing neither the data nor the algorithms
  - allowing to perform secure computations
  - whose results are the only released outcome.
- Three tools
  1. Homomorphic Encryption
  2. Secure Multiparty Computation
  3. Federated Learning

Current limitations

- The “visiting mode”, and more generally Privacy Preserving Machine Learning, is an emerging field in data science.
- As yet, the foundation on either HE or SMPC still implies a large communication and computation overhead
- Which makes it hard to use where very large amounts of data are required
  - since communication and computation costs are greatly affected by the increase of the number of involved parties or of the model’s complexity.

Synthetic Data

- Synthetic data are fully artificial data, which achieve anonymity by breaking the link between private information and data's information content.
- They are automatically generated by making use of Generative Adversarial Networks (GANs), based on two models playing recursively against each other.
- High-quality synthetic data closely resemble the real data and are a suitable substitute for processing and analysis.

Synthetic Data

- Synthetic data are fully artificial data, which achieve anonymity by breaking the link between private information and data's information content.
- They are automatically generated by making use of Generative Adversarial Networks (GANs), based on two models playing recursively against each other.
- High-quality synthetic data closely resemble the real data and are a suitable substitute for processing and analysis.

Generating differentially-private synthetic data

- Differential privacy provides an until-now lacking mathematical foundation to privacy definition:
- “Differentially Private Synthetic Data Generation is a mathematical theory, and set of computational techniques, that provide a method of de-identifying data sets—under the restriction of a quantifiable level of privacy loss. It is a rapidly growing field in computer science”

[National Institute of Standards and Technology Differential Privacy Synthetic Data Challenge 2019: Propose an algorithm to develop differentially private synthetic datasets to enable the protection of personally identifiable information while maintaining a dataset's utility for analysis]

AI in Healthcare Whitepaper

Big Data Value Association – BDVA Task Force 7 - Sub-group Healthcare - November 2020:

- “Many researchers find unclear the approaches that are required to collect anonymized data to ensure final users’ privacy”.
- “Generative Adversarial Networks (GANs) can generate meaningful synthetic data which do not suffer from the confidentiality constraints of the source data”.
- “EU-funded projects are encouraged to make available synthetic data sets that sufficiently resemble source data while avoiding privacy issues”.
- “Generative Adversarial Networks (GANs) have demonstrated potential, but their general applicability has not been established yet”.

A possible next step

- The time is ripe for an initiative that clarifies legal and technical definitions and eases the difficulties of coping with health data sharing while fully implementing the GDPR.
- Aim of prompting private and institutional centres to work on the potential of synthetic data and secure computation systems
- Suggesting a roadmap for their further implementation and market adoption
- A stepping-stone to foster the activation of a thriving digital ecosystem for the biomedical sciences.