RESEARCH DATA ALLIANCE

# Health Data IG @ P12

## Genomic data and privacy preserving technologies

Gabarone, Botswana, 5th November 2018

Edwin Morley-Fletcher

research data sharing without barriers
rd-alliance.org

# Sharing data regarding genomic information

- In April 2018, the European Commission has launched the goal of *Sharing data to personalise healthcare*.
- The Declaration for sharing health data was signed by multiple European governments with the specific aim of delivering cross-border access to genomic information.
- As stated by the EU Digital Commissioner Mariya Gabriel on that occasion, the possibility that genomics meet privacy technologies will represent a “real breakthrough”, and cross-border access to genomic information will act as a “game changer for European health research and clinical practice”.
- “Sharing more genomic data will improve understanding and prevention of disease, allowing for more personalised treatments (and targeted drug prescription) […] reaching a larger cohort that will provide a sufficient scale for new clinically impactful research”.

# How to allow genetic association testing in a privacy-preserving and secure environment ?

- The extremely sensitive nature of genomic data mandates the use of controlled data sharing policies to ensure that the privacy of individual subjects is not compromised.
- At the same time, it is clear that flexible data sharing and analytics will play a critical role in fuelling the next era of large-scale genomic association studies and scientific discoveries.
- To effectively address privacy concerns, it is crucial that state-of-the-art cryptographic security and privacy tools (such as Secure Multi-Party Computation, Homomorphic Encryption, and Differential Privacy) be appropriately adapted and deployed for large-scale genomic data analysis and sharing.

# Secure Multiparty Computation

- Secure Multiparty Computation (SMC) is a subfield of cryptography with the goal to create methods for parties to jointly compute a function over their inputs, keeping these inputs private.
- SMC allows a set of distrustful parties to perform the computation in a distributed manner, while each of them alone remains oblivious to the input data and the intermediate results.
- The computation is considered secure if, at the end, no party knows anything except its own input and the results.

# Homomorphic Encryption

- Homomorphism describes the generic property of an encryption scheme that allows to perform operations directly on encrypted data.
- Several existing encryption schemes are available that are homomorphic to any or certain operations.
- Deep Learning based solutions are adapted to work on homomorphically encrypted input-output data, i.e. supervised training techniques are employed on encrypted input-output value pairs.
- The deep learning model outputs encrypted results which are sent back to the client and decrypted by the client with the symmetric key.

# TUB, within MyHealthMyData, selected for the EC Innovation Radar

- Thus, the secured distributed processing of the medical data is performed in such a way that the third party does not learn anything about the data, and the user does not learn anything about the machine learning model.
- The approach ensures that both data and predictions remain private and data analysis is performed only on the encrypted version of the data.
- One Partial Homomorphic Encryption solution has been developed within the MyHealthMyData project by the Transilvania University of Brasov, and is now showcased by the EC Innovation Radar at: https://ec.europa.eu/futurium/en/industrial-enabling-tech-2018/transilvania-university-brasov
- The Innovation Radar Prize will select the top innovators supported by the @eu_h2020 during the past year.

# Output privacy

- Data (or, output) privacy, addresses the orthogonal question of what is revealed and ensures that the function output itself will not reveal "sensitive information" and cannot be used to learn about any individual data instance(s) possessed by any contributing party.
- Depending on the definition of privacy, or the type of problem at hand (static vs. interactive data publishing) different approaches may be used based on either static data publishing (i.e., "sanitization" or anonymization) techniques, or on Differential Privacy (DP).

# Anonymisation

- For data privacy, a common solution is based on anonymized data publishing.
- The key idea lies in removing identifying information from the published data, as well as generalizing/obfuscating secondary information (quasi-identifiers, e.g., age, zip code), that might lead indirectly to the identity of an individual.
- This is typically accomplished through combinatorial techniques for aggregating quasi-identifiers, such as k-anonymity or l-diversity. Still, such techniques offer no real privacy guarantees against all possible attacks (e.g., through linkage with other data source(s)).
- Differential privacy (DP) is one of the most popular (and, practical) definitions of privacy today. Intuitively, it requires that the mechanism outputting information about an underlying dataset is robust (with high probability) to the modification of any one sample.

# Data traceability and leakage control

- Data traceability and leakage control, deals with the important issue of tracing the routing and usage of outsourced personal data.
- Data providers may want to know if their data is used as a-priori foreseen or has been illegally rerouted, resold, or simply leaked.
- Data watermarking schemes offer possible mechanisms for controlling malicious personal data leakages.

# Watermarking

- Watermarking can be used to identify exactly the user who leaked some pieces of information, even if the data was modified, re-encoded or merged them with some other data.
- To do so, data just needs to be watermarked with the ID of the user accessing it.
- As defined, watermarking is an “a posteriori” protection mechanism due to the fact it allows access to the data while guaranteeing protection through the dissimulated message.
- It completes encryption, which is an “a priori” protection mechanism, as encrypted data typically needs to be deciphered before being accessed.
- It also provides an effective deterrence mechanism, as informing of the presence of invisible tracers that uniquely identify them, makes them less prone to reroute of leak the data.

# Verifiable computing

- Verifiable computing addresses the critical question of correctness of outsourced data computations.
- Outsourcing data services (e.g., analytics, AI model training/inference) to third-party, potentially untrusted data processors requires support for verifiable/authenticated computing mechanisms, where short, easy-to-verify cryptographic proofs are provided to ensure that the outsourced computation was carried out correctly

# The AI challenges

- Supporting complex, state-of-the-art AI models (e.g., deep neural networks) introduces new privacy requirements.
- Due to their complexity, such models have huge capacity for “memorizing” arbitrary information and, thus, can easily leak private information on sensitive training data.

# What type of data sharing ?

- Genetic data represent very sensitive data that can allow the identification of the individuals present in a dataset and therefore they need to be protected.
- In order to evidence genes involved in diseases, large datasets are needed that cannot be gathered by a single team but involve collaborative efforts.
- Data sharing is thus required but should not compromise individual privacy.

# Sharing strategies

- Strategies have been developed where instead of sharing the data, teams share the results of the tests performed at each locus and a meta-analysis using these summary statistics is then performed.
- However, it is also necessary to test more sophisticated models where interactions between genetic variants located in different genes could be involved.
- Utilising privacy preserving AI models, it is possible to provide user-friendly services to analyse exome or whole genome data of patients and compare them to data on population controls in order to evidence genetic disease association.

# Synthetic data

- Synthetic data are fully artificial data, automatically generated by making use of machine learning algorithms, based on recursive conditional parameter aggregation, operating within global statistical models.
- By definition, synthetic data do not allow any personal re-identification of original individual datasets (they belong to no really existing persons), while at the same time retaining information usefulness.

# Towards holistic synthetic virtual cohorts (1)

- With recent rapid technological developments, different types of proteomic and genomic data with different formats, structures and sizes are currently accessible.
- It is possible to focus on gene expression, single nucleotide polymorphism, copy number variation, and protein-protein/gene-gene interactions, and employ AI based techniques, making use of deep neural networks (GAN and InfoGAN) to generate these different types of genetic data.

# Towards holistic synthetic virtual cohorts (2)

- It is possible to apply these techniques to generate both SNP-chip like genotypes (containing only common genetic variants) and sequencing data, where all the genetic variants present in individuals are generated.
- With these techniques, a multi-modal set of synthetic patients can be developed, where broadly consistent structured, imaging, and genetic data are created for specific clinical phenotypes.

# Liberate data and preserve privacy

- Personalized medicine needs advanced analytics and big data,  but this need will keep undermining traditional privacy preserving approaches
- In the new ecosystem, trust, liabilities and responsibilities will be distributed
- In such a context, no unique tool or solution can prevent breaches, but privacy can be accomplished by design integrating multiple solutions in the same platform (as in MyHealthMyData)
- Synthetic data are one of the main vehicles to liberate data and preserve privacy

# For further contacts

- Edwin Morley-Fletcher emf@lynkeus.com
- Davide zaccagnini d.zaccagnini@lynkeus.com
- http://www.myhealthmydata.eu/