AVICENNA ALLIANCE

Avicenna Alliance
Association for Predictive Medicine

Avicenna Alliance Position Paper

AI and Big Data effective readiness:
A privacy-enhancing pathway to data access

Contributions

Edwin Morley-Fletcher, Prof – Lynkeus, VPH Institute; Avicenna Alliance

Cécile F. Rousseau, PhD - Voisin Life Sciences Consulting; Avicenna Alliance

Emmanuelle M. Voisin, PhD - Voisin Life Sciences Consulting; Avicenna Alliance

Liesbet Geris, PhD – University of Liège & KU Leuven; VPH Institute; Avicenna Alliance

Markus Reiterer, PhD - Medtronic, PLC; Avicenna Alliance

Maria Cristina Jori, MD - Mediolanum Cardio Research; Avicenna Alliance

Martha De Cunha-Burgman, **MSc** -Medtronic, PLC; Avicenna Alliance

Michaël Auffret, MSc - Voisin Life Sciences Consulting; Avicenna Alliance

Payman Afshari, PhD - Johnson and Johnson; Avicenna Alliance

Wen-Yang Chu, MSc - Virtonomy.io; Avicenna Alliance

Alicia Waterkeyn, LLM - RPP Group; Avicenna Alliance

I. About the Avicenna Alliance

The Avicenna Alliance is an association of industry and academia and healthcare organisations who have a commercial or research interest in the development of *in silico* medicine.

The Alliance, established in 2016, has its origins in the Virtual Physiological Human Initiative, a European Commission endorsed research area on computer modelling and simulation. Tasked by the European Commission with developing a "Roadmap for *in silico* medicine", the Alliance now seeks to put this roadmap into policy and ensure the development of a well-functioning framework for the *in silico* medicine ecosystem.

This Alliance bridges the gap between the scientific community, industry and policymakers by advocating for policy changes that take scientific and market developments into account.

II. Introduction

The Avicenna Alliance warmly welcomes the European Commission's ambitious policy package, including the European strategy for data and the White Paper on artificial intelligence aimed at making the EU fit for the digital age. In particular, the Avicenna Alliance Members welcome the European Commission's Data Strategy and its various references to the promising opportunities brought by health data for personalised medicine and the benefits and advances that computer modelling and simulation technologies can bring in healthcare.

To make the best of the new EU strategy, the Members of the Avicenna Alliance have drafted this position paper which applies both to the Public Consultations on a European Strategy for Data and the White Paper on Artificial Intelligence. This document has been developed by members of the Avicenna Alliance from medtech, pharmaceutical, software and life science industries and scientific community from academia. This paper is the starting point for a much-needed discussion on how the policy framework governing AI & Data in the health care sector can be reformed to adapt to the technologies of the digital age that can, in turn, be put to use to solve some of the greatest healthcare challenges facing society.

Twenty-first-century medical research and product development heavily rely on healthcare data, which include, but is not limited to, diagnosis, lab results, medical imaging, computational representation of anatomy and physiology. The Avicenna Alliance recognises the importance of patient privacy and data security. Nevertheless, there is mounting friction between technology solutions and regulatory constraints which makes effective data sharing in healthcare still very rare and conditioned by high transaction costs. Although available data is continuously expanding, it largely sits idle, fragmented in silos, and the advent of big data and AI environments is still remarkably delayed in European healthcare systems.

III. Privacy protection barriers to the efficient development of robust AI Solutions

The European General Data Protection Regulation (GDPR) mandates the use of privacy protection measures, such as anonymisation and pseudonymisation. In particular, anonymisation is outlined as the process by which personal data – including those referring to any individual's health status – are irreversibly altered in such a way that a data subject can no longer be directly or indirectly identified (by "all the means reasonably likely to be used").

This non-re-identifiability can, however, reduce information in the data to the point of making them of little use for scientific discovery or realistic AI-systems training. At the same time, the emergence of AI-based tools for re-identification pushes even further the amount of information that needs to be removed from a given data set to make it actually "anonymous".

Pseudonymisation, on the other hand, 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. Such additional information, therefore, needs to be kept separately. It is also 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 explicit personal consent, for being shared with third parties.

Both anonymous and pseudonymous health data end up by being, either inherently inadequate or extremely hard to scale up to an aggregation level of big data sets, that can allow efficient development of robust AI solutions in healthcare.

Neither of these approaches is technically and economically sustainable to boost data-driven R&D at industrial scales. There is, therefore, a crucial and still unsolved challenge around the sharing of big health data, and an additional challenge of applying AI on these resources in a lawful GDPR-compliant way.

IV. Promising results of the visiting mode and the generation of synthetic data

Recent outcomes of a Horizon Europe 2020 EU-funded project, MyHealthMyData (MHMD), have shown two ways for solving the challenge around big health data sharing.

The visiting mode

One is the so-called “visiting mode”, in which data are not physically accessed by third parties, but “algorithms are brought to the data” and only the outcomes of secure computations are released. The visiting mode operated in MHMD through mechanisms like homomorphic encryption, secure multi-party computation, and federated learning with untrusted black-box. These mechanisms were realised in conjunction with a permissioned blockchain system for recording transactions, an off-chain storage of health data, a metadata catalogue to view and request available data assets, smart contracts for automatically handling individual consent and institutional permissions.

If the implementation of the “visiting mode” is more envisaged, the Avicenna Alliance Members strongly call the European Commission to devote the investments and the research necessary to put in place a sustainable economic model or financing instruments to support SMEs, academic centres and other stakeholders. Transition is necessary to build, implement and maintain the necessary algorithms enabling the “visiting mode”.

The generation of synthetic data

Another approach, also indicated by MHMD, is the generation of synthetic data. Synthetic data are created from real-data, by machine-learning generative model to produce realistic, yet artificial data that maintain the same statistical properties as the original dataset. They reflect the nuances of the original data, without endangering leaking personal information. The difference between traditionally anonymised data and synthetic datasets is that the latter protect privacy by adding statistically similar information, rather than stripping away unique identifiers.

The statistical characteristics of the real population are "learned" during synthetisation, from the original data, while the synthesis process uncouples identifiable information from the data information content, attaining anonymity though still preserving information richness. This directly responds to the 26th recital of the GDPR which underlines that "*personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*", while the overarching objective is to create high-quality synthetic data that closely resemble the real data and are a suitable substitute for processing and analysis.

The benefits of using synthetic data include:

I. reducing legal constraints when using sensitive or any type of regulated data;

II. tailoring the data needs to certain conditions impossible to achieve with authentic data; and

III. generating datasets, which can be leveraged as a stand-in to validate mathematical models and, increasingly, to train machine learning models.

More particularly, the use of synthetic data in health care is particularly significant for *in silico* clinical trials, where the effect of a medical intervention is studied on a virtual patient population that closely and accurately reflects the population of interest.

V. The way forward

Synthetic data follows the distribution or other attributes of the original raw dataset. It therefore allows, under a regulatory perspective, personally identifiable information to be transformed into anonymised data that still demonstrate the natural relationships of the original variables at stake, with the highest statistical reliability.

Hence, synthetic data offers a new way of striking a balance between the risks of data leakage and of information value loss, by ensuring strong privacy guarantees while maintaining statistical properties of the original data. This is even more true when synthetic data is combined with other privacy preservation techniques, such as differential privacy, to reconcile utility for machine learning applications and legal compliance thresholds.

The developments and potentialities of data synthetisation and augmentation are becoming part of a broader field of investigation. For instance, in "radiomics"[1], new AI-based image reconstruction tools are applied at the stage of raw data decoding and transforming, with the effect of hugely accelerating image acquisition processes with minimal compromise on final image quality, despite what would traditionally be considered as data "under-sampling" and patient "under-exposing".

Investigation on such algorithms on diagnostic data augmentation still needs to pass through dedicated machine training, consequent proof of concept and extensive clinical testing-validation and therefore needs to receive adequate attention and funding from the EU. The promises of safer and faster diagnostic images to be acquired thanks to AI-based data augmentation and partial synthetisation is strongly appealing and should be given the support and attention necessary to make synthetic data a reality.

Today, technology advancements open the way to real-world production and use of synthetic data at industrial levels. However, in the absence of validated frameworks for their creation, validation and use, commercial and institutional players are not likely to adopt them in time for placing Europe at the forefront of this potential revolution. Standards, guidelines and user workflows for the selection and management of most appropriate data generators, for their configuration, are needed to enable non-expert data analysts to confidently assemble generative pipelines to serve their organisations' research and development goals. Likewise, robust evidence of statistical reliability will foster their adoption by key decision-makers and regulators to solve some of the biggest health challenges facing in the EU today.

[1] Radiomics correspond to methods extracting high-dimensional data and features from radiographic medical images using data - characterisation algorithms and aiding clinical decision-making and outcome prediction.

Rizzo, S., F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti and M. Bellomi (2018). "Radiomics: the facts and the challenges of image analysis." European radiology experimental 2(1): 36-36.

VI. Conclusion

The time is ripe for an initiative by the European Commission that will fill the void of legal and technical definitions, as well as of policy, and ease the solution of current difficulties with health data sharing while fully implementing the compliance with both the spirit and the letter of the GDPR.

The Avicenna Alliance recommends that the European Commission **set up a Coordination and Support Action Call**, bringing together leading academic centres with experts from industry and healthcare organisations to work on the demonstrated potential of synthetic data, as well as of the "visiting mode" solutions.

The Avicenna Alliance Members are convinced that a coordinated approach is necessary and therefore ask the European Commission to **define a general regulatory framework** and **draft a roadmap for the mainstream implementation and market adoption of synthetic data and "visiting mode technologies"** as this constitutes a first milestone in the activation of a thriving Digital Single Market for the biomedical sciences.

This Position Paper is endorsed by the 22 Members of the Avicenna Alliance on Friday 29 May 2020: