#02
NEWSLETTER
-

MY HEALTH
MY DATA

February 2020

DEEP LEARNING DYNAMIC CONSENT
SYNTHETIC DATA ANALYTICS
DISTRIBUTED LEARNING GDPR
BLOCKCHAIN SECURE COMPUTATION

AT THE FRONTIER OF HEALTHCARE BIG DATA

NEWSLETTER INFORMATION

Dear Readers,
Welcome to the second and final issue of the *MyHealthMyData* (MHMD) newsletter, which has been conceived as a sort of 'legacy' publication, recapitulating the work carried out throughout different research areas, illustrating scientific results and achieved innovations and providing hints on future research directions stemming from these outcomes. After a general overview of MHMD (page 4) summarising its mission and major achievements, and a perspective on its adopted legal approach and data privacy impact assessment (page 7), you will be guided through our project along with the different sections dedicated to the work we have been conducting in altogether 38 months.

The first section, **'The patient at the centre' (page 11)**, is dedicated to the actions planned to realise our use case for individuals (page 12), enabling patients to have access to their health data in digital form, employ it for personal use and share it for research and innovation, including our MHMD app beta testing campaign, and to the social study exploring users' ethics, concerns and behaviours on eHealth (page 14).

The second section, **'Ensuring privacy and security of data' (page 17)**, is dedicated to illustrating all innovations aimed to safeguard data and system security and patients' privacy in the context of MHMD, including its use of blockchain and smart contracts for implementing data transactions and users' consent (page 18), privacy-preserving data publishing (page 21), advanced solutions for secure computation on encrypted data and distributed learning (page 23), generation of synthetic datasets (page 27) and the public penetration challenge (page 29) carried out to assess and verify the overall system security.

The third section, **'Leveraging the value of big data in healthcare' (page 31)** is dedicated to the MHMD metadata catalogue (page 32), designed to enhance cohort discoverability without risking to release any personal data, and some use-case solutions provided to researchers and clinicians to take the best out of de-identified health data for enhanced medical research and care, including the Deep Reasoner/Deep Explorer tool for case-based reasoning (page 34), the personalised physiological modelling for clinical decision support (page 36) and the data value estimation model (page 39).

To conclude, the newsletter will provide some insights on some public events, conferences and workshops, dedicated to distributed ledger technology, big data, artificial intelligence, eHealth, personal data, GDPR compliance, and privacy-enhancing technologies (page 41), that the Consortium attended in the last months.

Hope you enjoy the reading!

HIGHLIGHTS

INNOVATION RADAR PRIZE 2019

Transilvania University of Braşov has been awarded as **Category Winner** of the **EC's Innovation Radar 2019**, in the **Industrial & Enabling Tech** category, for its Homomorphic Encryption secure computing framework, during Research and Innovation Days (Brussels, 24-26 September 2019)

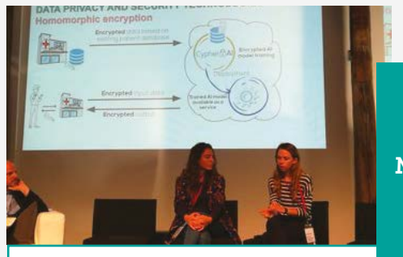
CONVERGENCE – THE GLOBAL BLOCKCHAIN CONGRESS

This flagship event (Málaga, 11-13 November 2019) was dedicated to blockchain in its multiple applications including AI, IoT, Finance, Mobility, Energy and much more. MHMD held the panel ***"Blockchain and healthcare: How is blockchain facilitating a secure, scalable, data-sharing infrastructure in the healthcare industry?"*** moderated by our former Project Officer Saila Rinne. A project **exhibition booth** was organised for dissemination and networking.

DATA PRIVACY IMPACT ASSESSMENT

As an additional project deliverable, *Panetta & Associati* has produced a **Data Privacy Impact Assessment (DPIA) of MHMD**, which has assessed and ultimately certified the **compliance of the MHMD system to GDPR privacy and security constraints and requirements**. The DPIA is a tool required in the GDPR in the case of large-scale processing of special categories of data, to describe the processing, assess its necessity and proportionality, and help manage the risks to the rights and freedoms of data subjects with appropriate measures, and has been performed for the very first time in the context of a research project.

MYDATA 2019

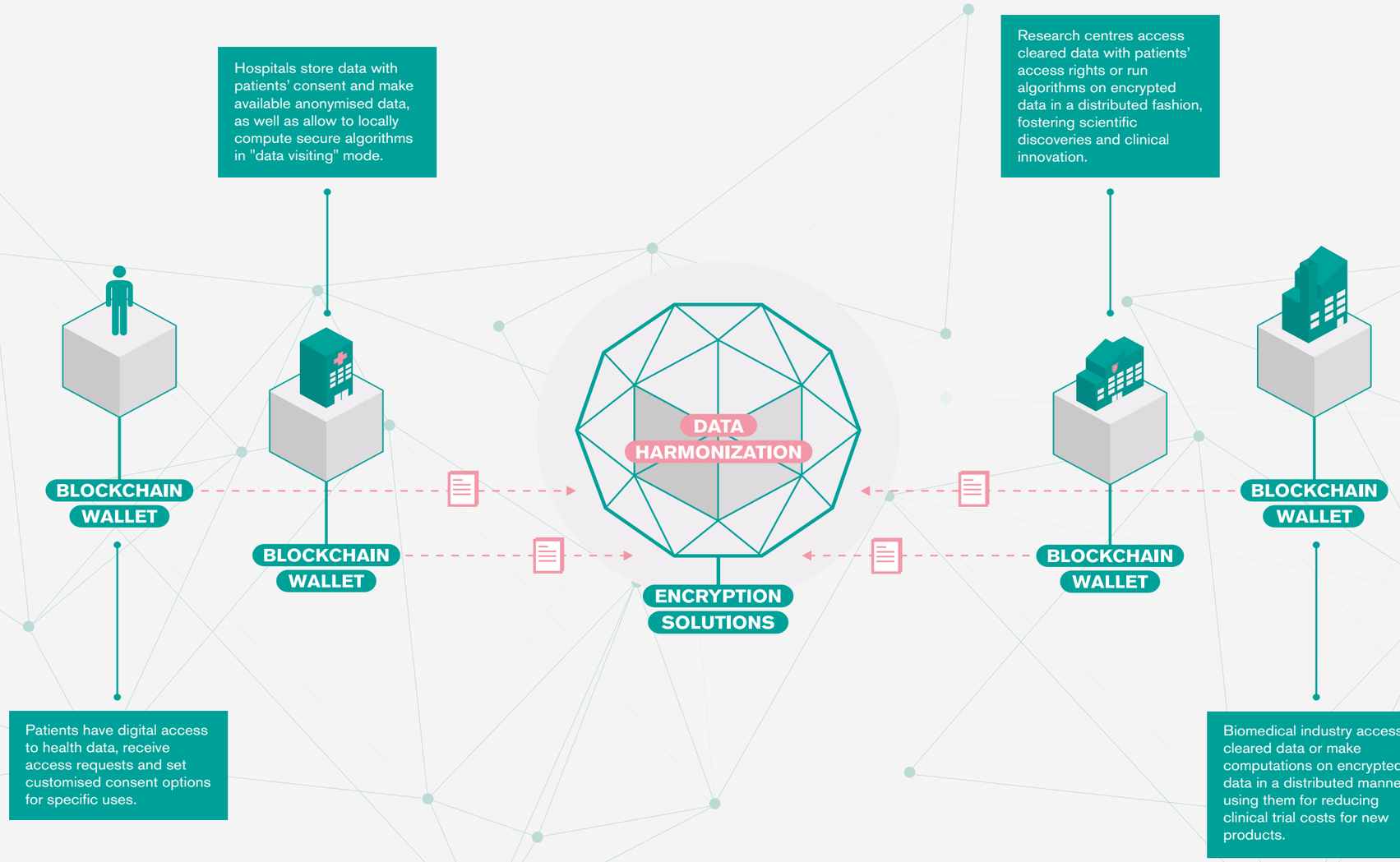


MyData (Helsinki, 25-27 September 2019) is the flagship event of the namesake movement, investigating and promoting individuals' empowerment in regard to personal data, for a more trustful and transparent digital society. MHMD took part in the session *"**Keeping control and minimizing risk in secondary usage of health data**"* with focus on MHMD data security and privacy approach and innovations, with special regard to **homomorphic encryption.**

TABLE OF CONTENTS

MHMD: AN OVERVIEW

Edwin Morley-Fletcher, *Project Coordinator*
// LYNKEUS



MyHealthMyData (MHMD) started with a relatively small budget of less than 4 million € EU and Swiss funding at the end of 2016. At that time, most of the challenges surrounding the sharing of data, in what is now currently defined as the European Health Space, were still unclear, with major uncertainties around key choices like whether to opt for relative centralisation, with what cloud or multi-cloud approaches, or whether the then emerging distributed ledger technology would actually prove to be a viable solution.

Given its pioneering goal to define a blockchain-based system for health enriched by privacy-enhancing technologies, in the past 38 months MHMD has exerted the function of a technological, ethical and legal sandbox for testing the feasibility, robustness and meaningfulness of a new privacy paradigm allowing to develop *"trust-less trust"* to facilitate data transactions between people, hospitals, research centres and businesses, leading to an open biomedical information network centred on the connection between organisations and the individual.

The core aim of MHMD has thus been dealing with health data through a distributed individual and institutional empowerment system, aimed at ensuring secure access from anywhere, on any device. This goal has implied a number of key features:

- a *private permissioned blockchain* based on Hyperledger Fabric for controlled data access;
- an *off-chain data storage* participated by multiple hospitals and by individuals;
- a *metadata catalogue* allowing to safely inspect what data are available on MHMD;
- *dynamically and automatically managing consent* by smart contracts;
- a resulting, successful *privacy by design and GDPR compliance assessment*.

MHMD has, in other words, moved from the unqualified expectation of developing a compliant system for sharing data, to the goal of guaranteeing shared control, providing trust by computation and, in addition, highly innovative ways to generate, use and share synthetic health data.

The MHMD approach to computational trust is based on three solutions: *homomorphic encryption, secure multi-party computation*, and *federated learning with an untrusted black box*. The peer-to-peer nature of the blockchain makes it possible to allow a potentially limitless number of clinical institutions and of individual subjects to safely interact with their and others' health data, either through the "visiting" mode, based on secure computation "*bringing the algorithms to the data*", or through the sharing mode, by generating and publishing anonymous, including synthetic, data, controlled with differential privacy.

According to various scenarios of trust and privacy-preserving needs, MHMD health data can also be published as pseudonymous data. A semi-automated tool, AMNESIA, is used for both pseudonymisation and anonymisation. Synthetic data, nevertheless, have proven to be a powerful solution to scale up data sharing in privacy-constrained environments such as healthcare.

No technology has yet emerged to bridge GDPR privacy requirements and the growing demand for health data, specifically for big data. Synthetic data are a technology with the potential to bridge this gap by providing realistic data while not exposing any identifiable information, in support of both medical-AI technologies and traditional biomedical products development.

They in fact overcome the crucial challenge of achieving full anonymity by breaking the link between private information and data information content. The underlying principle is that values in the original database are algorithmically substituted with those taken from the database statistical distributions, to create entirely new records, with as little traceable relation to the originals as possible. In MHMD they have been successfully used to publish health data and health imaging data, to train machine learning tools and to test clinical decision support applications.

Differential privacy (DP) provides a mathematical foundation to then substantiate privacy assessment and its legal definition. DP offers a reliable tool to control the risk of re-identifiability from anonymous data at each stage of the data life cycle, which is particularly instrumental in the modern big data ecosystem. In this sense synthetic data, while already documented in literature, represent a conceptual breakthrough in the new context of the GDPR, as they allow to respond to the elusive quest for broadly sharing health data in full compliance with this regulatory framework.

In its unrolling, MHMD has engaged also in an initially non-foreseen privacy-by-design and compliance assessment, with the aim of checking whether all the fundamental principles of the GDPR were duly fulfilled, that the risks to data subjects' rights and liberties were appropriately addressed and minimized and that the entire range of processing operations underlying the MHMD platform were in line with applicable laws and regulations. This additional effort was accomplished conducting a detailed legal analysis of how all the elements composing the MHMD system eventually can be considered fully compliant, so that conclusively the developed system proves to be "*secure, interoperable, accountable, traceable, trustable, resilient, scalable, distributed, non-repudiable, transparent and unlinkable*".

MHMD OBJECTIVES AND INNOVATIONS

CITIZENS' EMPOWERMENT

DATA PRIVACY AND SECURITY

DATA VALUE ENHANCEMENT

PERSONAL DATA ACCOUNTS and DYNAMIC CONSENT
A personal data wallet, implemented through the **MHMD app**, to **manage personal data from disparate sources** (medical records, mobile apps, IoTs), establishing access conditions through a dynamic consent module supported by a dedicated smart contract

BLOCKCHAIN and SMART CONTRACTS
A private, permissioned blockchain architecture that **manages and authorizes the access and exchange of data** according to user-defined conditions enforced through smart contracts

DATA PRIVACY IMPACT ASSESSMENT (DPIA)
Legal assessment describing actors and relevant roles, obligations and responsibilities, data categories and processing operations, system components, data usage modalities, data de-identification measures and system security

DATA PRIVACY AND SECURITY TECHNOLOGIES
Basic PERMISSIONED DATA PUBLISHING (with pseudonymised or anonymised data), allowing the hospital to activate a semi-automated k-anonymisation tool

VISITING MODE (getting the outcomes of "bringing the algorithms to the data" without providing data access)
- **SECURE MULTIPARTY COMPUTATION:** performing computation in a collaborative manner among mutually distrustful parties
- **HOMOMORPHIC ENCRYPTION:** executing computation on encrypted data and decrypting results at the source
- **FEDERATED LEARNING:** training AI-based algorithms in a distributed fashion among local distrustful parties

SYNTHETIC DATA GENERATION: generation of synthetic datasets through machine learning, controlled by differential privacy

METADATA CATALOGUE and OFF-CHAIN STORAGE
All datasets are stored off-chain (local repositories, personal clouds), yet are indexed on the blockchain by persistent identifiers, populating a metadata catalogue which describes data available in the network without revealing any identifiable information

BIOMEDICAL ANALYTICS
Applications leveraging de-identified and encrypted data
- **DeepExplorer/DeepReasoner:** deep learning for AI configuration and case-based reasoning
- **Personalised physiological models for clinical decision support** (blood circulation model)
- **Models for data value estimation**

WHAT HAS HAPPENED AFTER THE GDPR: THE MHMD LEGAL APPROACH

Lorenzo Cristofaro, Rocco Panetta and Marta Fraioli
// PANETTA & ASSOCIATI

Almost two years have passed since the long-awaited application of the *General Data Protection Regulation n° 2016/679* ("GDPR" or "Regulation"). As a matter of fact, such novel legislation has increasingly spread the knowledge and upraised the interest of individuals on data protection issues, contributing to general awareness. Subject to fundamental safeguards, Member States have been granted the possibility to integrate the provisions of the GDPR by establishing further and more specific rules in a number of pre-determined areas, including medical and scientific research, with particular regard to the identification of the conditions for lawfully processing or re-using personal and health data for this purpose and for the exercise of individual rights under the GDPR.
This scenario has made the task of identifying common rules governing *MyHealthMyData* (MHMD) extremely more complex either because *(i) there are no best practices, case laws or binding guidance to help understand how streamline and foster scientific research through personal and sensitive data; and (ii) the applicable obligations may greatly vary from a Member State to another, making unfeasible to implement the project legal framework homogeneously and triggering some unwanted operational local inconsistencies*. However, despite these uneven initial conditions, a concerted, collective work between all MHMD partners allowed to identify technical and organisational measures to implement data protection principles and integrate the necessary safeguards into data processing functions, to meet GDPR requirements and protect data subjects' rights.
The challenges faced by such a cutting-edge project have been tackled in a privacy-enhancing and security-nourishing perspective, attaining a number of innovative goals, such as – *inter alia* – enabling stakeholders to search and appraise MHMD datasets by

Figure 1. *EU contries impacted by the GDPR since 25 May 2018. Photo credits: "Europe privacy law GDPR" (CC BY 2.0) by Frank Buschman (SmedersInternet) via Flickr*

performing descriptive statistics on the underlying sources without revealing any identifiable information; enabling the concept of 'data visiting', thanks to newly-designed application of homomorphic encryption, secure multi-party computation and federated learning, which allow to apply computation on clinical datasets without any of them being pulled from the hospital repositories or exposed in other ways (researchers may only see unidentifiable aggregated outputs); designing and implementing a private and permissioned blockchain infrastructure based on Hyperledger Fabric in order to implement only credentialed data access and make consent-based data exchanges tamper proof; clarifying roles and responsibilities of any party involved in a smart contracts-orchestrated processing ecosystem.
Together with an in-depth analysis of the security and technical features adopted within the project, these achievements have been scrutinized in detail in the "*Privacy by design and compliance assessment*". MHMD partners pioneered and put into effect the principles of privacy by design and by default since the very early stages of the project, anticipating many of the recommendations made by the *European Data Protection Board* ("EDPB") in the recent draft of *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default* (currently under public consultation).

The compliance assessment: some insights
The privacy-by-design and compliance assessment was designed to evaluate if (i) all the fundamental principles of the GDPR were duly fulfilled in the MHMD platform, (ii) the risks to data subjects' rights and liberties were appropriately addressed and minimized (or eliminated, where possible) and (iii) the entire range of processing operations underlying the system was in line with applicable laws and regulations. Have these results been achieved?
All datasets available in MHMD are indexed by means of non-reputable, *persistent*, unique and standard *identifiers* (PIDs). The resulting catalogue is populated by metadata which describe data assets available in the network without revealing any identifiable information and can be browsed by advanced semantic-enabled engines, allowing to segment, group and therefore create specific cohorts of data. Persistent identifiers are leveraged *in lieu* of real and row data during transactions, ensuring that no personal information is leaked or exposed at any time. Once researchers have identified the cohort, datasets are made available after the appropriate pseudonymisation or anonymisation techniques are applied, depending on whether all the conditions set forth by the applicable legislation, with particular regard to explicit individual consent, have been duly satisfied.
Alternatively, queries made by researchers can be resolved in a 'data visiting mode', namely applying secure computation on the clinical datasets outputting only unidentifiable results, leveraging either federated learning with untrusted black box, or homomorphic encryption, or secure multi-party computation. Such computations are run directly on the datasets held by data controllers as part of the federated data storage to which hospitals are connected as blockchain nodes.
A further crucial innovation implemented in MHMD is the generation of fully synthetic data using a combination of aggregate statistics from a known population. Virtual patients are created from scratch by drawing from original distributions, so that realistic replicas are generated with no privacy disclosure risk.

Roles in MHMD: different outfits under the GDPR umbrella
The project can be defined as the sum of the legal and technological rules applicable to all possible interactions between patients, hospitals, researchers, app users and the MHMD platform/interface operator ("Platform Operator"). Among these parties, some roles are somehow bound and conditioned, as in the case of hospitals, which shall be considered autonomous data controllers, as they are free to determine the purposes and the means of respective data processing, and individual data subjects, i.e., MHMD app users.

GDPR ROLES IN MHMD

Data subjects
Personal subjects data are referred to

PATIENTS

MHMD APP USERS

Data Controller
THE BRAIN
Determines purposes and means of the data processing alone or jointly with other controllers

HOSPITALS
- MHMD blockchain nodes
- Datasets stored in local repositories

RESEARCHERS
- Academia - Industry

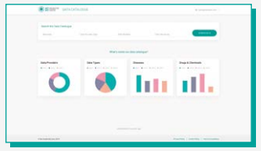
Data Processor
THE ARM
Acts on behalf of the Controller on data made available by the Controller

MHMD
- Register individual datasets on the Catalogue through the blockchain
- Register medical datasets on the Catalogue through the blockchain
- Regulates access to personal data according to Data Controllers-defined conditions

MHMD INNOVATIONS FOR DATA PRIVACY AND GDPR COMPLIANCE

BLOCKCHAIN AND SMART CONTRACTS

- A **private and permissioned blockchain** based on **Hyperledger Fabric** was implemented to **make consent-based data exchanges tamper proof**, as any request of access to a cohort of data transits through the blockchain. **All personal data is stored exclusively off chain**, to assure that the exercise of the individual rights cannot be hindered or limited in any manner.
- The **consent given by data subjects is implemented through smart contracts** to operationalize the associated usage permissions in the context of the blockchain architecture and make the **management process transparent, semi-automatic and tamper proof**.



METADATA CATALOGUE

- The catalogue is populated by **metadata which describe data available in the network without revealing any identifiable information**, and can be browsed by semantic-enabled engines, allowing to segment, group and create specific cohorts of data.

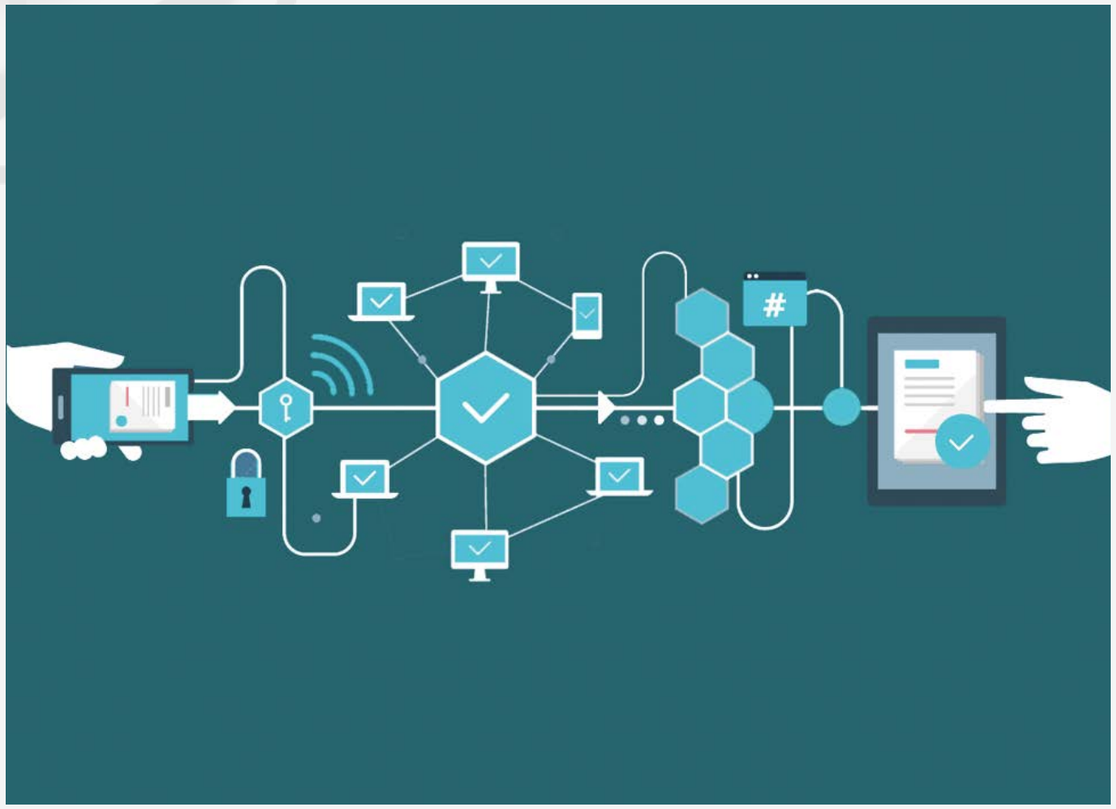
LOCAL COMPUTATION

- Data queries can be responded in a **'data visiting mode', i.e., applying appropriate computation on the clinical datasets without any of them is pulled out from the hospitals' repositories**. Analytics are run directly on the data held locally by data controllers, and researchers only see unidentifiable aggregated outputs.

SYNTHETIC DATA

- Synthetic data are generated using **a combination of aggregate statistics from a known population**. Drawing form data distributions through machine-learning algorithms, virtual patients are created from scratch, so that a significant amount of realistic data can be generated with no risk of being able to single out the original data subjects.

The Platform Operator plays the role of a trusted technological service provider capable to apply all those measures which are needed to allow users to lawfully access and process such data, thus acting as a data processor on behalf of both hospitals and researchers.
The Consortium evaluated all possible combinations of roles and responsibilities, including which parties directly liaise with data subjects on a case by case basis (sometimes the hospitals that collect clinical data, while other times the Platform Operator which establishes and manages the contractual relationship with the users of the MHMD app).



The MHMD blockchain
MHMD relies on a decentralized, blockchain-based data access control infrastructure that provides a new mechanism of trust and direct, value-based relationships between hospitals, data subjects, researchers and businesses that monitors and securely orchestrates any processing of personal data, either when relying on local computation or on secure sharing through the catalogue.
In order to meet the highest standard from both a data protection and security standpoint, a private and permissioned blockchain was designed and implemented based on Hyperledger Fabric.
This infrastructure makes consent-based data exchanges tamper proof, as any request of access a cohort of data transits through and is registered on the blockchain and must then be validated by all ledger nodes. No personal data is stored "on-chain", to assure that the exercise of individual rights – which is the core of the applicable data protection legislation – cannot be hindered or limited in any manner.
Only a metadata description of the information registered on the blockchain appears safely in the catalogue open to authorized stakeholders. This process allows the blockchain to maintain the records of available data and its associated history without the need to store any personal data in it.
The consents given by data subjects are implemented through smart contracts to operationalize the associated usage permissions on the blockchain architecture and make their management transparent, automatic and tamper proof. Smart contracts automatically control data access criteria against data users credentials and allow, or not, to execute the data exchange or computation when the preconditions defined by the data subject are met by the data access query by the researcher.
Furthermore, the ledger is protected by one-way cryptographic algorithms, to describe data and transactions results in an anonymous fashion and k-anonymity like models, which also prevent statistical inference attacks to locate data or individuals.
Following two years of intense prototyping, a GDPR-compliant permissioned blockchain was finally deployed in pioneering hospitals and research centres in Europe, after successful outcome of both internal and public hacking challenges to validate the soundness and robustness of the infrastructure.

THE PATIENT AT THE CENTRE -

ENABLING PATIENTS TO ACCESS PERSONAL DATA AND TAKE CONTROL IN THE DISTRIBUTED ECOSYSTEM

MHMD FOR INDIVIDUALS: PERSONAL DATA HANDLING AND CONSENT MANAGEMENT AT A CLICK

Ludovica Durst and Davide Zaccagnini
// LYNKEUS

Modern healthcare systems are increasingly leveraging patient-centred digital health tools which are fostering direct involvement of individuals in the management of their care. The need of 'putting patients at the centre' is coming into sight as also highlighted in the GDPR, with special regard to the so-called 'data portability', the right of receive data *"in a structured, commonly used and machine-readable format"* and have it transmitted from a controller to another without hindrance. Nevertheless, making patients' medical records accessible in digital format is still a difficult undertaking, not only for lack of viable technical solutions, but also in regard to privacy protection and still lukewarm users' involvement in their own health education and awareness. That's why, since the very beginning, MHMD made *patient centricity* and motivation one of the main challenges in the construction of the future biomedical data ecosystem.

The MHMD solution: personal data accounts, dynamic consent and smart contracts

MHMD implements patient centricity by equipping individuals with a user friendly mobile app to control data access rights (and delegate them where necessary), allowing to keep a digital copy of their medical records into a cloud-based wallet, namely the *personal data account* (PDA), along with the possibility to aggregate any IoT and mobile app data related to health, fitness and well-being. The PDA can be hosted on Dropbox, Google Drive or any other cloud-based data management service: in this way, patients

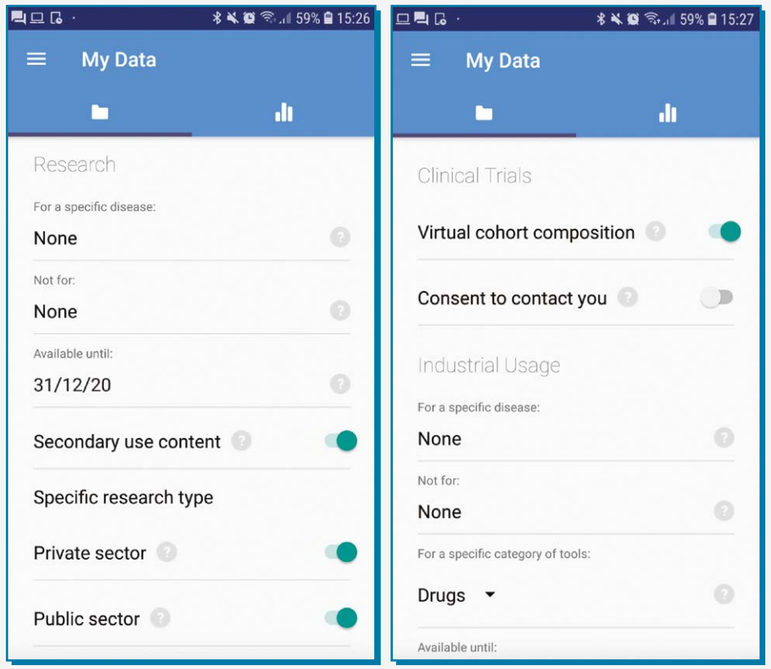


Figure 1. *Preview of the data usage settings page of the MHMD app user interface.*

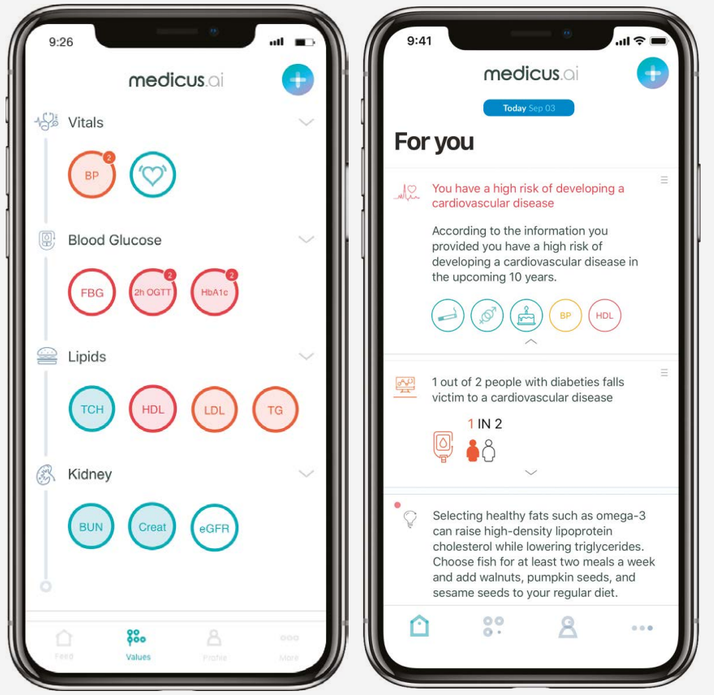


Figure 2. *Preview of the Medicus.Ai mobile app.*

can carry their data and share it as they prefer and need to. Besides medical care, users can make data available for biomedical research in exchange for services or reward.

Privacy preservation and consent were critical enablers of this model. A dynamic consent system gives individuals the ability to customise consent preferences including who is authorized to use it and for what purpose, which can be monitored from the app itself. Parameters can be modified at any time, including also the right to be forgotten. These preferences are then enforced through a smart contract, an executable logic embedded in the blockchain, that automatically matches data access requests to user-defined consent settings and allow the data transaction only if conditions are met, i.e. natively enforcing compliance with the GDPR and local policies. Data transactions are registered and traced on the immutable, tamper-proof, fully auditable MHMD blockchain.

This process has been made as intuitive as possible in the MHMD app, which supports the creation of a personal user account, the synchronisation of data from existing repositories and the setting of user's consent preferences (Figure 1). Data requests are delivered to the user in the form of notifications, in which the requester (e.g., university, research institution, private company), research objectives, requested data type and envisaged processing are detailed. The MHMD app also allows to monitor personal data usage in time, through high-level, descriptive graphical displays.

Individual users' engagement actions and the MHMD app onboarding campaign

Several initiatives have been launched to engage individual users, leveraging different channels. The *Ospedale Pediatrico Bambino Gesù* in Rome has organised meetings on digital health innovation, where the MHMD app was presented to an audience of physicians and caregivers. A patients' focus group has been organised by the University College of London at the Great Ormond Street Hospital with a small group of users, for a preliminary evaluation of app functionalities and feedback gathering.

Digi.me leveraged its expertise, user network and industry collaborations in other sectors (with e.g., Barclays, the BBC, BT, Centrica, Facebook) to boost awareness on the MHMD app as an enabler for individuals to engage in the biomedical data and innovation ecosystem. In particular, Digi.me held a number of focus groups with UK-based users to explore opportunities and challenges in patient centricity.

Most of all, direct enrolment of MHMD app users was accomplished by the involvement of Medicus Ai, a healthcare technology company based in Austria and serving hundreds of thousands of EU users providing laboratory data exchange apps and services (Figure 2). Thanks to this collaboration, a dedicated marketing campaign was organised, towards the end of the project, to recruit Medicus and Digi.me users on the MHMD app, supported by the integration with the Medicus app.

The MHMD app was leveraged in a study to assess opinions, preferences and actual behaviour around personal data control, ultimately allowing to cast some light on the complex psychological and cultural dynamics which make privacy, and its protection, one of the most complex issues facing modern societies.

COMPARING STATED PREFERENCES WITH ACTUAL BEHAVIOURS IN REGARD TO PRIVACY AND DATA USAGE CONTROL: THE MHMD SOCIAL STUDY

Ludovica Durst, Anna Rizzo and Davide Zaccagnini
// LYNKEUS

In a highly dynamic information ecosystem, citizens should be aware of the sensitivity and value of health data while, on the other hand, researchers, public health officials and businesses should be able to access those data efficiently as they pursue their legitimate agendas. As yet, though, people diffusely deplore the lack of control over personal data, while paradoxically showing little interest in taking direct responsibility in managing it, mostly due to the effort and time the data protection process requires. All this while uncontrolled aggregation of personal data by large corporations has taken the centre in public debate. Mistrust at times translates into a general, undirected reluctance to share data, and will increasingly do so, impeding research and innovation. These issues seem to be laying in poor understanding of how the personal data ecosystem works and, despite the GDPR has strongly emphasized individuals' empowerment, citizens still lack effective tools enabling them to make the most out of their data.

The MHMD social study, namely *"Platform-driven assessment of attitudes and sensibilities with regard to ethical, privacy, and data security issues"* aimed at assessing public perceptions and attitudes towards privacy and data security, in contrast with actual behaviours as observed in the use of the platform and app. The theorical framework was obtained from an extensive literature review exploring questions around trust, privacy concerns, control and health data sharing, leading to the perspective that stated preferences, views and concerns often do not match actual individuals' behaviours, following the so-called "privacy paradox". To verify this hypothesis, the Consortium decided to adopt a twofold approach: on one side, to investigate users' views through a dedicated questionnaire; on the other, to analyse users' behaviour "in practice" when utilising the MHMD app, in order to assess them comparatively.

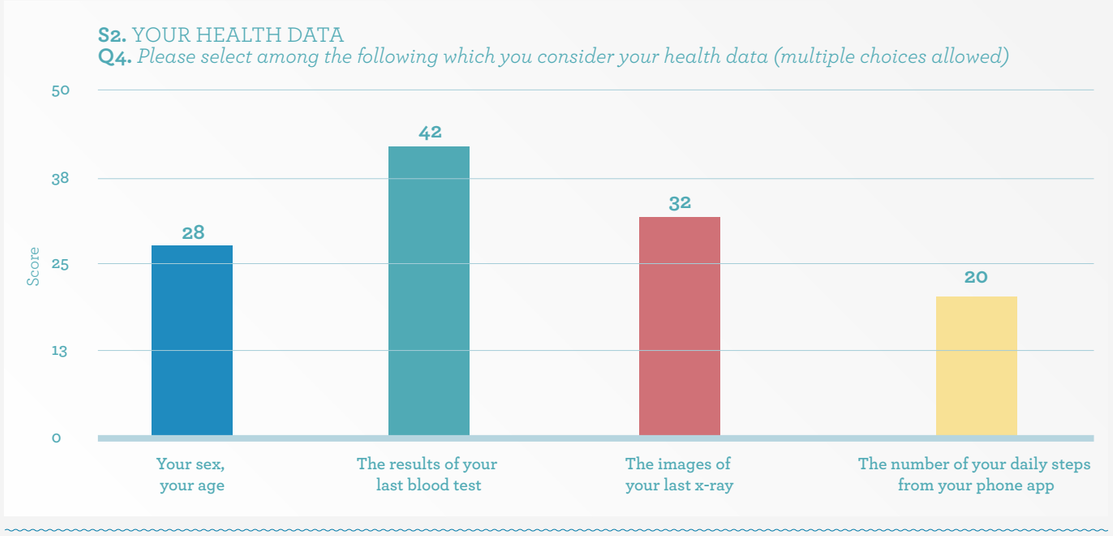


Figure 1. *MHMD user questionnaire, "health data" section. Results (%) show that there is quite an heterogeneous perception, among users, of what shall be considered health data, or not.*

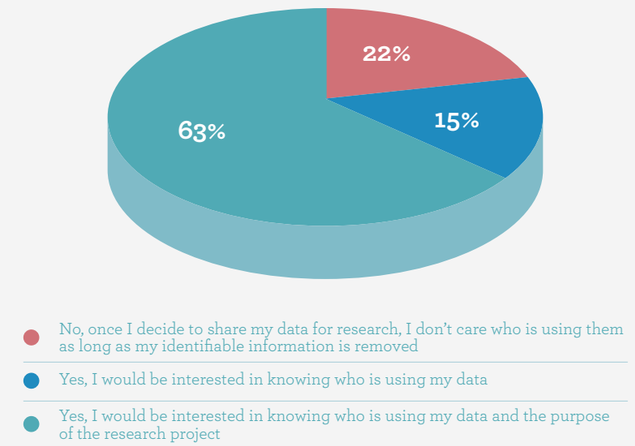


Figure 2. *MHMD user questionnaire, "MHMD app" section. Results (%) show that the vast majority of users feel the need to know who is using their data, and most of those feel necessary to know both that and the purpose of precessing.*

Assessing users' perspectives: the MHMD survey

An app-based survey helped to quantify the prevalence of certain attitudes, stated values and preferences in four areas : (1) *the value of health data, (2) privacy demands in regard to health data, (3) value of MHMD app features* and (4) *feedback on the MHMD app experience.*

The first focus area aimed at understanding the level of awareness about what can be considered 'health data' and uncovered, among other things, that while a higher number of users agreed on considering results of medical examinations as health data, only a small proportion of them (20%) deemed their physical activity medically relevant, identifying a gap in public awareness on the value of this data in the biomedical data system (Figure 1).

The second group of questions (i.e., *"Your health data and your privacy"* section) aimed at assessing the relation between personal data sharing and privacy concerns. Users were asked how happy they would be to share data for medical research, and how concerned they would be about their privacy. Results show that users have a good propensity to share their data for medical research (mean score = 3.7/5) but, unsurprisingly, most of them were rather concerned about their privacy (mean score = 3.3/5) and nearly all of them (98%) would feel more confident if it was fully anonymised. This finding confirms the strategic focus identified during the project on developing solutions which can by-pass the complexities and risks of pseudo-anonymisation, like synthetic data, as a tool to decouple individual identities from the information content of the data, which can be therefore distributed at scale.

Questions in the third section (i.e., *"The MHMD app"*) were primarily meant to observe how individuals would actively engage in managing their data. Only a very small proportion of them (15%) would not be interested in knowing who is using their data, providing it is de-identified, while the vast majority (85%) is eager to know who is using the data, and for most of those (63% of the total) it is important to know both who is using the data and for what purpose, showing the patients' will to be effectively engaged and being able to decide who to share with, which is also driven by ethical considerations (Figure 2).

With regards to the feedback on users' experience (i.e., *"Your experience feedback on the MHMD app"* section), users seem have appreciated the ability to set consent options, in particular the possibility to revoke data access or to extend them at any moment (mean score = 3.7/5), while the app has achieved less success with reference to the clarity of information and options and user-friendliness (mean score = 3.3/5) and the feeling of being in control of the data (mean score = 3.2/5). This highlights the challenge of simplifying in app-based or even desktop-based workflows the legal and ethical complexities of patient empowerment and suggests that new approaches to information management should be explored to streamline users' interaction with the consent and data management process.

Data control in practice: data study requests notifications and consent settings

In parallel with the survey, an analysis of users' activity was performed by assessing chosen consent option settings and response to specific data usage requests with different features (i.e., institution or company requesting the data, type of requested data, scope of the study and envisaged processing). To this aim, data usage requests were delivered to users as notifications in the app to test their reaction on specific cases (Figure 3). By comparing behaviours with the opinions stated in the questionnaire, the study aimed at assessing gaps between intentions and actions, to understand underlying motivations and what ultimately drives users' actions.

With regard to research options, 52% of registered users have specified to allow public sector research, almost as much have also consented to private sector research (42%) showing that there is no substantial difference from the users' perspective between sharing data with public or private bodies; also, 40% of users consented to secondary use of data: a not very high threshold, which evidently suffers a diffidence towards research when its 'boundaries' are less determinate. This partially contrasts with opinions stated in the questionnaire: in fact, when asked to express their will through multiple-choice questions, almost all users (96%) decided to make their data available to hospitals and research centres, a majority of users (59%) also allowed no profit organizations, while few only (28%) consented to for profit organizations. When coming to specific requests, though, most notifications were accepted (on average, 91% accepted, 7% declined and 2% pending over a total of 392 notifications, see Table 1), with similar scores between public and private bodies, regardless of the subject requesting

the data and usage purposes. Drawing general considerations, it seems that individuals were positively conditioned by the transparency of the information provided, knowing 'who is going to do what' on their data. It appears thus that it is not only a matter of trust in a specific institution, but also of the dynamic of the interaction. As organizations present themselves in a transparent way (with name, study scope and data usage description) users seem to be inclined to respond positively.

On the whole, the "lesson learnt" is that no single feature would realize the goal of a conscious and active data sharing under clearly understood ethical and legal parameters. Rather, a delicate and complex balance between control features, simplicity and, more importantly, transparency is needed to positively engage users. Supporting research represents a valid goal for most users, who in the overall are not too concerned by privacy risks, as long as subjects requesting data and research purposes are clearly stated. People are also inclined to trade personal information as long as control can be exercised efficiently. At the same time, de-identification is considered indispensable.

On the basis of the social study conducted in MHMD, we concluded that three key principles should always be implemented when trying to maximise engagement around privacy preserving solutions: the paramount need for transparency, the guarantee of full anonymisation and the assurance of carefully-designed user experiences. Following these principles proves to be crucial for developing a digital healthcare ecosystem where the empowerment of citizens and the enhancement of citizens' trust can be fully achieved.

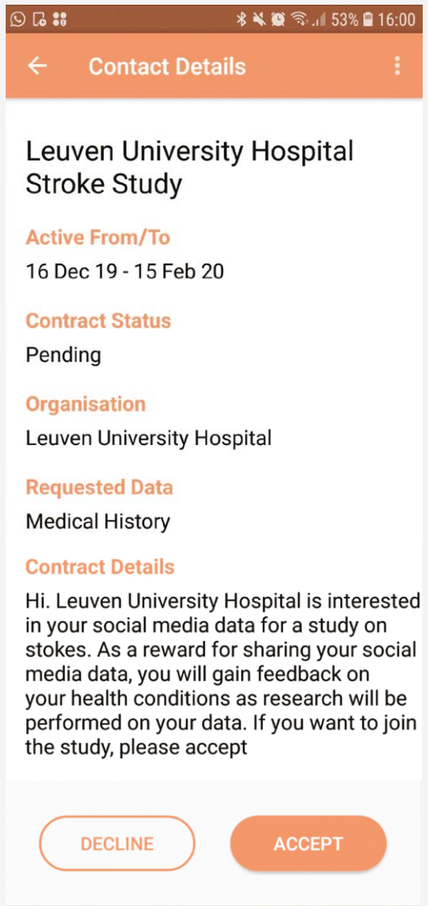


Figure 3. *Example of MHMD app notifications exemplifying specific usage requests, describing the requestor, type of requested data, intended usage and reward.*

Study notification organisations (overall users)	Total notifications			Per organisation		
	Organisation	Notifications	% out of total notifications	Accepted	Declined	Pending
	University of Navarra Medical School	80	20%	93%	6%	1%
	Rome European Hospital	73	19%	93%	3%	4%
	Boston Scientific	63	16%	89%	10%	2%
	Institut Gustav Roussy	56	14%	86%	13%	2%
	University of Freiburg Medical Center	45	11%	87%	13%	0%
	Charité Berlin	14	4%	100%	0%	0%
	Leuven University Hospital	10	3%	90%	10%	0%
	Istituto Europeo di Oncologia	9	2%	100%	0%	0%
	Medicine for Europe	9	2%	100%	0%	0%
	Ospedale San Raffaele	9	2%	100%	0%	0%
	Università Cattolica	8	2%	100%	0%	0%
	Danish Pain Research Center	4	1%	100%	0%	0%
	Klinik Hirslanden Zurich	4	1%	75%	25%	0%
	Novartis	4	1%	100%	0%	0%
	European Heart Health Institute	2	1%	50%	50%	0%
	Pharma Company	2	1%	100%	0%	0%
	Total	**392**	**100%**	**91%**	**8%**	**1%**
				Mean		

Table 1. *Overall responses to MHMD app notifications relevant to specific usage requests, showing that the majority of users accepted usage requests, irrespective of the type of institution requesting the data.*

ENSURING PRIVACY AND SECURITY OF DATA

ENFORCING DATA AND SYSTEM SECURITY THROUGH BLOCKCHAIN AND SMART CONTRACTS, PRIVACY PRESERVING DATA PUBLISHING, SECURE COMPUTATION AND DISTRIBUTED LEARNING, SYNTHETIC DATA GENERATION, PENETRATION CHALLENGES

BLOCKCHAIN FOR HEALTH DATA: THE MHMD PIONEERING EXPERIENCE

Mirko De Maldè
// LYNKEUS

When the MHMD project started, back in November 2016, it was still unclear how blockchain technology could have been used for supporting critical data transactions in the healthcare environment. The first experiments and initiatives were starting then, in parallel with our project. In three years, the landscape has evolved rapidly, and now several initiatives are actively exploring blockchains in different contexts, from health data management to pharmaceutical supply chain optimisation, drug anticounterfeiting and clinical trial management. Not surprisingly, the potential of blockchain technology in healthcare has been recognized by several stakeholders around the world, including the *Report of The Joint Economic Committee Congress of The United States or the 2018 Economic Report of the President*, which indicates blockchain as a potential solution, among other things, for coordination and portability of medical records.

Health data management: a critical landscape

Health data management is one of the most promising and challenging fields in which to experiment the use of blockchain technology. The need for innovation comes from a variety of shortcomings in current ways data are stored and controlled. Centralised systems, composed by different data siloes, are costly, unsecure and inefficient, and due to technical shortcomings and regulatory constraints they make it very difficult to mobilise and integrate sparse data sources (e.g., clinical data produced in hospitals, patients-generated data, etc.) in a meaningful way. Such an outdated model brings us to a "data-rich but information-poor" paradox, as existing data cannot be leveraged to support health providers, researchers and patients. There is an urgent need to find new models for data mobilisation and integration from various sources, in particular taking into account the emerging category of patient-generated "real-world" data from medical-IoT systems and apps. New consent management and direct access and control tools are therefore needed for guaranteeing regulatory compliance and enabling individuals' engagement, paving the way toward data self-sovereignty and patient-centric healthcare.

WHAT ROLE DOES BLOCKCHAIN PLAY IN MHMD?

- >> Acts as a "traffic light" that **manages and authorizes the access and exchange of data** according to user-defined conditions
- >> Ensures complete **traceability and auditability** of data transactions
- >> Automates the application of **privacy-preserving technologies**
- >> Facilitates **compliance with the GDPR**, in particular with respect to the *right to erasure/correction* and relevant reporting obligations

Healthcare blockchain: the way forward?

Blockchain technology enables trust, accountability, traceability and integrity of data in health record management, as it introduces a decentralized mechanism for controlling and accessing data so that each healthcare organization manages its own data, while enforcing data time-stamping and robust audit trail mechanisms. It is therefore not surprising that various blockchain studies and experiments focus on data management (Hölbl, M., Kompara, M., Kamišalić, A., & Nemec Zlatolas, L. 2018). At the same time, a number of private initiatives have started, such as MedicalChain, Patientory, HealthBank, Longenesis, HIT foundation, just to mention some examples with varied degrees of success.

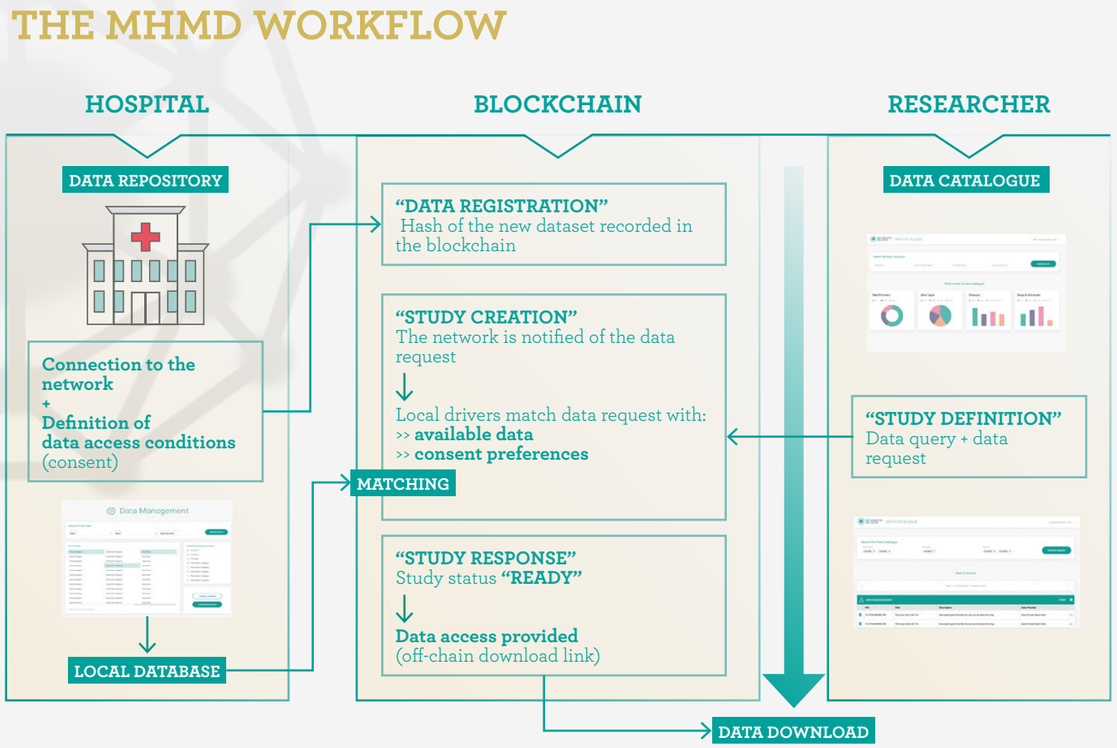
Thanks to these features, in addition to decentralized management and immutable audit trail, blockchain provides additional benefits such as more robust data provenance models (improving both ownership control and traceability of the origin of a specific data asset), increased robustness and availability (thanks to the high level of redundancy provided by the technology) and improved privacy and security (thanks to the associated cryptographic algorithms which are now being optimized as dedicated components for these types of infrastructures). This makes the blockchain technology particularly suitable for improving the efficiency of medical research, providing researchers with a transparent, reliable infrastructure to efficiently exchange permissioned data, allowing aggregation of longitudinal health information and supporting data interoperability.

MHMD: an innovative approach

In such a context, MHMD has pioneered the usage of blockchain in healthcare, realizing a platform facilitating data management, providing full control over data for data controllers and data subjects, and establishing a trust ecosystem for data to be easily mobilised and made available to researchers and innovators in the life science domain, in compliance with the European *General Data Protection Regulation* (GDPR). First, MHMD uses the blockchain as an orchestration layer in charge of managing and authorising data exchange and access (a sort of " traffic light" for data), on the basis of user-defined permission/consent settings through a dedicated smart contract. At the same time, MHMD provides novel ways of collecting and operationalising the patient consent, automating its enforcement and guaranteeing its respect at each step of the data access and mobilisation process. At the same time, the MHMD platform provides full traceability and auditability of data access permissions and exchanges, enforcing GDPR compliance, particularly with regard to the right to erasure/correction, through an automated notification system. Finally, the MHMD blockchain is responsible for automating data pre-processing (i.e., data "sanitisation") and for triggering and orchestrating specific computation processes through *secure multi-party computation* (SMPC), allowing researchers to "ask data-driven questions" to the MHMD network, receiving directly the results of a computation, rather than the data.

The MHMD data access pipeline

The scheme below shows how the data access pipeline is construed within the MHMD architecture, leveraging three fundamental components: at the data provider's side (hospital), (1) *the local MHMD driver*, is responsible for the initial data mapping, semantic harmonisation, registration and permission control functions. The driver is also responsible for sorting the proof of matching at the local level, providing the link between the information recorded in the blockchain and the underlying real data. On the data user's side (academia/industry), the system leverages (2) the *MHMD central metadata catalogue*, which allows the user to browse the data available in the network and request data access. The middle layer is (3) *the blockchain*, where the proof of data existence/registration is first recorded, and the smart contract is activated when a study



request is submitted. The MHMD smart contract includes several functions, such as *data registration*, *study creation/data access request*, *privacy preserving technologies* to be applied over the data transaction, *study update*, *response to study* by indicating available data matching the study request. Once the study request is submitted, the *create study* function is triggered, distributing the data request to all nodes in the network. The query is resolved separately and locally by each node. Once all responses are collected, the study response function is triggered, responding to the study request with the confirmation of data availability. As last step, a link is finally provided to the data user for downloading the permissioned dataset. In alternative to data publishing, the research can also submit a request for computation, getting as a result the outcome of the analytics computed over the distributed environment.

Benefits for patients, hospitals and researchers

The blockchain as developed in MHMD brings value and clear benefits. For *hospitals*, it represents a way to enforce and guarantee compliance, to improve consent management and to encourage permissioned data sharing for research and innovation. For *academia & industry*, it facilitates access to relevant and high-quality data sets for basic research, drugs and device development and testing, as well as AI training and validation. Finally, for *patients*, it provides the tool to control their personal data toward more empowered health and self-management, and better communications with care providers. Through the dynamic consent tool implemented as a mobile application, patient can provide or revoke consent and manage data access rights, getting full visibility on data access, and eventually be able to directly extract value from their own data.

> The MHMD blockchain in a nutshell

Based on a dedicated requirement analysis, the Consortium adopted *Hyperledger Fabric*, which offers a **permissioned blockchain** that ensures **high transaction rates, low network latency and low energy demands**, while providing a **flexible, modular and secure architecture with a pluggable consensus mechanism**. In Hyperledger Fabric, permissioned entities are not only known, but their encrypted **identities and roles are registered and verified**.

The traceability model includes two levels of data proofs that are stored in the blockchain:

- the ***proof of existence***, a secure hashing process for the data registered into the system;
- the ***proof of matching***, linked to the real data item stored in the Data Controller's repository. The link between proof and data is stored outside of the blockchain to comply with the GDPR.

The **MHMD smart contract** includes a number of functions for *data access, privacy preservation, consent enforcement and regulatory compliance*.

PRIVACY-PRESERVING DATA PUBLISHING IN MHMD: THE AMNESIA DATA DE-IDENTIFICATION TOOL

Manolis Terrovitis
// MANOLIS TERROVITIS

Anna Rizzo
// LYNKEUS

In the MHMD system, personal health data are stored off-chain, in the remote or cloud-based repository of the Data Controller's choice; hospital datasets, for instance, are generally stored in local facilities. When exposing those datasets to data-user third parties, though, personal data needs to be de-identified (or "sanitized") before sharing to protect data subjects' privacy.

Pseudonymous vs anonymous data: what is the difference?

This modality, called *Privacy Preserving Data Publishing* (PPDP), foresees in first place the removal or key-substitution of subjects' *direct identifiers* (i.e., information specifically related to an individual, e.g., name and surname, social security number), leading to the generation of pseudonymous data. According to the GDPR, such kind of data represents *personal data* that *"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 the personal data are not attributed to an identified or identifiable natural person"* (Article 4, GDPR). This kind of process, however, is not enough to guarantee an adequate protection of individuals' privacy. The "additional information" mentioned in Article 4 is represented by *quasi-identifiers* (e.g., postal code, gender age, date of birth), i.e., information that is not personally identifying information in itself, but is correlated enough with a subject that can lead to its re-identification if combined with other quasi-identifiers. When even this information is removed through specific anonymisation techniques such as transformation or statistical noise addiction, data can be considered *anonymised (or anonymous) data*. Anonymised data is considered free from the risk of re-identification, and thus is no longer considered personal data, falling out of the provisions of the GDPR.

Amnesia: a k-anonymisation tool for data providers in MHMD and beyond

ID	AGE	ZIPCODE	DIAGNOSIS
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer

ID	AGE	ZIPCODE	DIAGNOSIS
1	[20-30]	130**	Heart Disease
2	[20-30]	130**	Heart Disease
3	[20-30]	130**	Viral Infection
4	[20-30]	130**	Viral Infection
5	[40-60]	148**	Cancer
6	[40-60]	148**	Heart Disease
7	[40-60]	148**	Viral Infection
8	[40-60]	148**	Viral Infection
9	[30-40]	13***	Cancer
10	[30-40]	13***	Cancer
11	[30-40]	13***	Cancer
12	[30-40]	13***	Cancer

Figure 1. *Example of k-anonymisation, obtained through generalisation of quasi-identifiers (age and zipcode) of a medical database subset.*

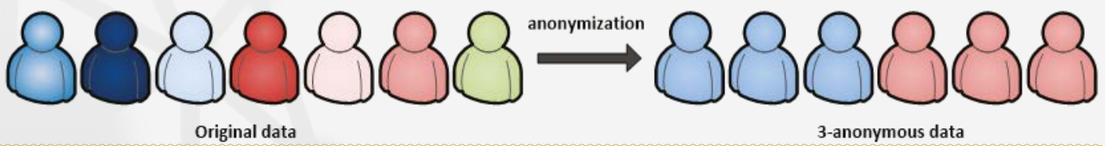


Figure 2. *A piece of data has k-anonymity properties if the information for each subject in the dataset cannot be distinguished from at least k - 1 individuals, being k a parameter of choice.*

To support Data Controllers in such a fundamental step to data sharing, MHMD needed to provide an easy way to sanitize personal information and share their datasets in the form of anonymised data to third parties. To this aim, the team of Athena RC developed and implemented in the system a customizable and user-friendly anonymisation tool: *Amnesia*. Amnesia enables users to remove personal information from a health-related dataset and transform the rest of the information in a way that prevents from re-identification of the original data subjects. The tool allows, on a preliminary level, the removal of *direct identifiers*, generating pseudonymous data; on a second level, it enables the transformation of quasi-identifiers through k-anonymisation, a type of transformation obtained by generalization (or aggregation) of certain attributes. Age, for instance, can be easily generalised by grouping subjects in age groups; similarly, zip codes can be aggregated by erasing the last code numbers. A piece of data can be claimed to have k-anonymity property if the information for each person in the dataset cannot be distinguished from at least k - 1 individuals, being k a parameter of choice. Amnesia exploits a unique k^m-anonymity algorithm. k^m-anonymity is a variant of k-anonymity that is able to efficiently anonymize multidimensional data, without significantly reducing their quality. In k^m-anonymity we assume an upper bound m on the number of quasi identifiers an adversary might know of. This assumption limits the impact of multiple data dimensions and numerous quasi identifiers, as the anonymisation algorithm has always to examine combinations of up to m quasi identifiers.

The data anonymisation features of Amnesia are aimed at releasing data to wide untrusted audiences, since the transformation on the data is irreversible. Naturally, the anonymisation parameters (i.e., the parameter k), has to be chosen carefully to minimize risk of information leak.

An open-source, easy-to-use tool for online and remote use

Amnesia is equipped with a user-friendly graphical interface that helps users to easily parameterize the anonymisation process. Also, Amnesia supports the user in assessing the impact of different parameters and preview different possible solutions, so to allow to tailor anonymisation to the actual analytics use of a certain type of dataset.

In addition, the tool offers a ReST and a command line API to facilitate its integration to complex information systems. These APIs allow the complete automation of the anonymisation by supporting *templates*. Templates are a complete set of parameters required for achieving the desired solution. By using templates, the system does not require any additional input from the user (Figure 2).

In the secure data sharing ecosystem of MHMD, Amnesia can be used locally, at the level data providers' repositories (e.g., hospital repositories), *before* transferring them to any recipient. Amnesia has been successfully tested by the anonymisation of synthetic medical records generated within the project, which contain, among others, an arbitrary number of ICD9 and ICD10 diagnosis codes.

The Amnesia anonymisation engine is implemented in Java and works both on Windows and Linux. Its interface is coded in html and javascript to support both a local application and an on-line service. It is provided freely and open source through OpenAIRE and the European Open Science Cloud at https://amnesia.openaire.eu.

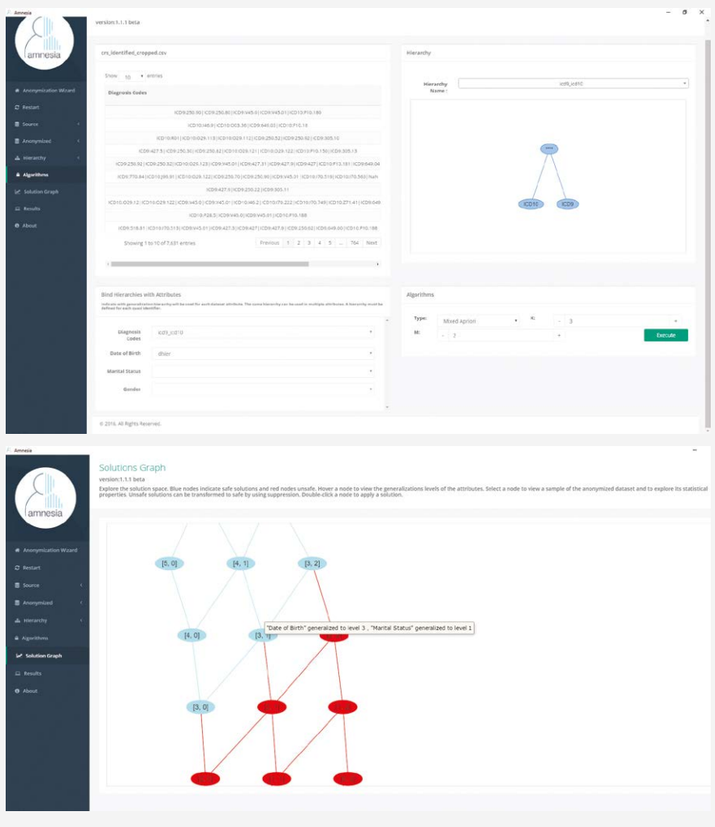


Figure 3. *Preview of the Amnesia anonymisation tool interface.*

BEYOND ANONYMISATION: SECURE AND DISTRIBUTED COMPUTING IN MHMD

Anna Rizzo
// LYNKEUS

Minos Garofalakis
// ATHENA RESEARCH CENTER

Lucian Mihai Itu, Cosmin Ioan Nita and Anamaria Vizitiu
// TRANSILVANIA UNIVERSITY OF BRAŞOV (UTBV)

Andre Aichert
// SIEMENS HEALTHINEERS

Besides authority-set obligations to data subjects' rights prescribed in the GDPR and other relevant regulations, data security and individuals' privacy against data breaches still constitute one of the biggest challenges for big data-driven research and AI. *De-identification* (both through *"pseudonymisation"* and *"anonymisation"*) of personal data before sharing currently represents the usual approach for data publishing. However de-identification, despite being a fundamental processing to guarantee a minimum standard level of protection, presents some serious pitfalls, and even the so-called "full" anonymisation of data is not always sufficient to preserve individuals from the risk of re-identification.

Sharing data securely: reality or myth?

According to the GDPR, published data can be addressed as "anonymised" if it is *"stripped of sufficient elements such that the data subject can no longer be identified. More precisely, that data must be processed in such a way that it can no longer be used to identify a natural person by using 'all the means likely reasonably to be used' by either the controller or a third party"* (Opinion 05/2014 on Anonymisation Techniques by The Article 29 Working Party). This model, also addressed as "release-and-forget" model, also foresees that this process must be "irreversible". But is that really the case? According to the most recent literature, the short answer is "no". Several studies have demonstrated the researchers' capability of successfully re-identifying data subjects and their personal information from various types of supposedly anonymous datasets, including browsing histories, medical records, taxi, subway or bike sharing trajectories, mobile phone or credit card datasets. A 2019 study by Nature Communications showed that, over a 3-million people US population, about 78% of individuals could be correctly re-identified from an incomplete dataset just by using four attributes, i.e., date of birth, location (PUMA code), marital status and gender (Rocher, L., Hendrickx, J.M., & De Montjoye, Y.A., 2019). Of course, the legal definition is subjected to the constant advancement of technological innovation and must be contextualised. In theory, the computational efforts made to derive information from a dataset could be limitless. In reality, even the most ill-intentioned hacker would give up if the required computational effort to re-identify data subjects was far disproportionate to the actual value of the dataset. The clue might be, then, to make data re-identification harsh enough to make the effort not worthwhile anymore.

Secure, but useless: the inevitable trade-off between privacy protection and data utility

Even if full, irreversible de-anonymisation was achievable, privacy preservation comes at a price. In data-driven research, *micro-data* (i.e., the information included at the level of single individuals) allows to derive important information benefiting the society as a whole, such as factors underlining a specific disease onset,

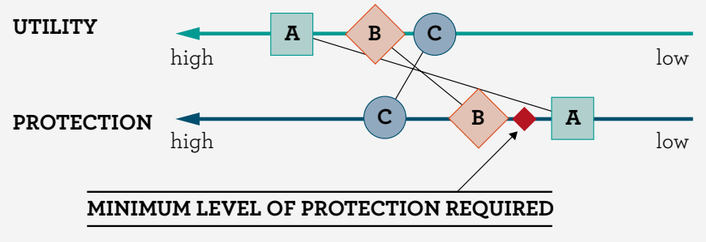


Figure 1. *Trade-off between data utility and privacy protection in k-anonymisation. Elaboration from Loukides, Grigorios, and Jianhua Shao. "Data utility and privacy protection trade-off in k-anonymisation." Proceedings of the 2008 international workshop on Privacy and anonymity in information society. 2008.*

drug efficacy or social-economic patterns. To cope with the risk of individuals' privacy loss in data publishing, anonymisation typically transforms microdata making it imprecise or distorted (e.g., by generalisation/aggregation, statistical noise addiction, permutation), limiting it to an acceptable level. This, however, causes an inevitable, substantial loss in its data mining utility, if compared to the original data. As maximising both data utility and privacy protection is computationally unfeasible, one of the biggest areas of research in privacy-preserving data publishing is about finding an optimal trade-off between privacy and utility of data. The answer is not univocal and must be evaluated on a case-by-case bases, together with the anonymisation algorithm of choice, depending on the privacy requirements of the specific dataset and research scopes. A popular approach, for instance, is to find a processing that retains as much data utility as possible, while satisfying a minimum required level of protection (Figure 1). Overall, though, it seems evident that anonymisation procedures cannot be considered neither fully resolutive for privacy-preservation, from a Data Controller's standpoint, neither fully desirable for the scopes of data-driven research, from a data user's perspective.

Privacy-preserving data flow execution: "bringing algorithms to the data"

For all these reasons, MHMD has proposed an additional data usage modality to serve as an alternative to privacy-preserving data publishing, collectively indicated as *privacy-preserving data flow execution*. The principle of this model is plain: instead of providing researchers with de-identified data for running data mining algorithms, we can go the other way around by *'bringing algorithms to the data'*. But what does it practically mean? Privacy-preserving data flow execution foresees running a given computational function, in a collaborative fashion, between mutually untrusted parties, and revealing each other just the computation results, making this approach privacy-preserving by default. In MHMD, we developed different kinds of privacy-preserving functions, namely *secure multi-party computation* (SMPC), *homomorphic encryption* (HE), together also addressed as *secure computation* techniques, and *distributed (or federated) learning* (DL).

Secure multi-party computation: joint computation over private inputs

In SMPC, a mathematical function is jointly computed among multiple distrustful parties in a distributed fashion, keeping respective data inputs and intermediate results private, and revealing nothing but the function output, a guarantee often referred to as "input privacy". The computation is considered secure if, at the end, no one knows anything except its own input and the final result. SMPC protocols are typically built assuming one of two "threat" models: (1) *honest-but-curious* (i.e., where parties are assumed to follow the protocol while attempting to learn other parties' private information) and (2) *active-malicious* (i.e., where parties can act maliciously and deviate arbitrarily from the protocol).

SMPC can be very useful for numerous practical applications where data is naturally distributed across multiple parties and data privacy is a major concern (e.g., analytics over sensitive data). Although the demand for input privacy guarantees has grown a lot in recent years, general-purpose SMPC protocols are still computationally expensive and quite limited in their ability to scale to real-world problems; still, modern SMPC protocol implementations are fast enough for certain types of computations (e.g., addition and multiplication by a public constant) often needed in practical application scenarios.

Utilizing one of the most popular open-source SMPC protocols, SPDZ, we have built a general-purpose platform where researchers can request secure data analytics over medical records spread over a federation of data controllers (e.g., hospitals), while at the same time ensuring patients' input privacy (Figure 2). SPDZ is secure in the active-malicious sense, i.e., it preserves input privacy even in the presence of an active attacker. We have also implemented a complementary functionality based on a secure importer protocol to allow hospitals to only participate in the computation as data providers and not as SMPC compute nodes. This functionality allows removing most of the computational burden from the hospitals. Our SMPC implementation supports many of the fundamental use cases of the MHMD platform. Notably, it

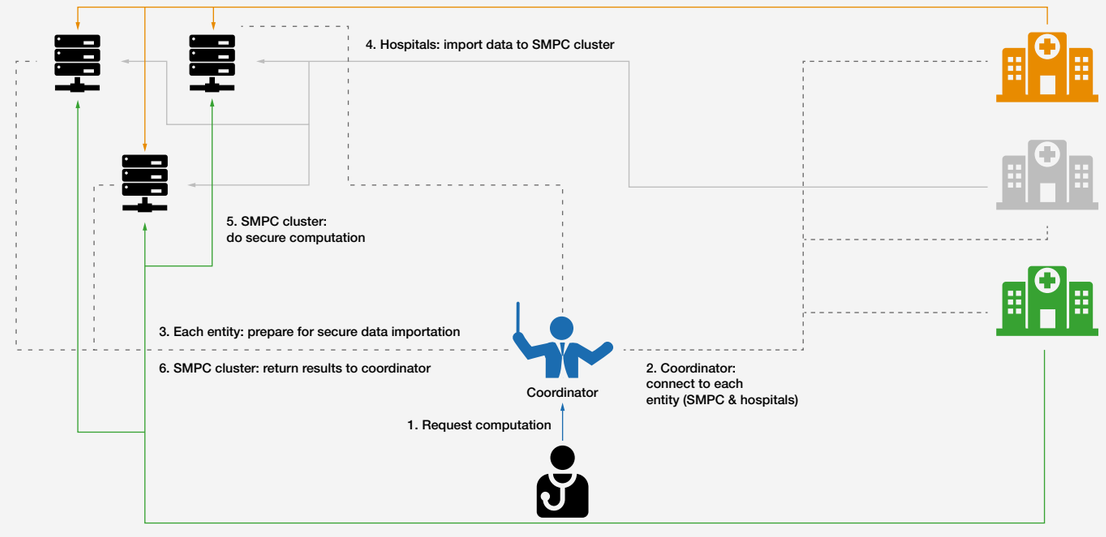


Figure 2. *SMPC workflow in MHMD.*

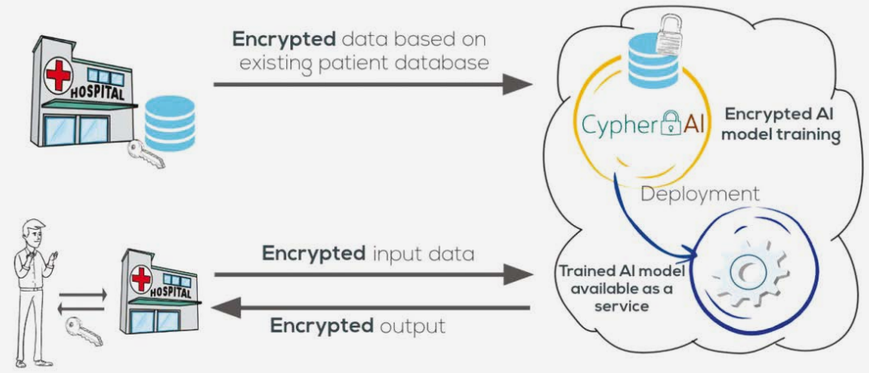


Figure 3. *HE framework developed in MHMD.*

enables the secure computation of statistics (e.g., histograms) over sensitive patient data, as well as the computation of deep learning models in a private and secure way within the "black-box" federated learning framework, introduced in the following paragraphs.

Homomorphic encryption: computation on encrypted data

Encryption is a process to encode a message or file so that it can be only be read by certain people, by using an algorithm to scramble (*encrypt*) data, while the receiving party is provided with a specific key to unscramble (*decrypt*) the information. Encrypted data can be compared to jewels that have been placed in a safe: while in there they are protected from theft, you cannot wear them, so they are kind of useless. Conventionally encrypted data is safe but cannot be used, even by legitimate parties for agreed upon purposes, until it is decrypted.

In HE, the issue is solved by using an encryption scheme allowing for computations on encrypted data, where data is encrypted before being sent to the computing service, and computations are performed on encrypted data. Once the results are available, they are sent back and decrypted at the source. Since the decryption key is not available to the computing service, the service has access only to the encrypted data, and no personal or useful information can be extracted (Figure 3). The property of an encryption scheme that allows operations on encrypted data is called homomorphism and includes *fully homomorphic* (i.e., where any function can be evaluated), and *partially homomorphic* (i.e., homomorphic for specific operations) schemes. *Fully homomorphic encryption* (FHE) has the disadvantage of increasing the computation time by around seven orders of magnitude, being thus impractical for most applications. *Partially homomorphic encryption* (PHE) is much faster than FHE and is currently available only for simple operations (e.g., summation, multiplication, summation and multiplication, searching, sorting, and equality checks). A more complex encryption strategy, that is homomorphic with respect to multiple operations, can be obtained by combining several PHE encryption schemes in a layered structure. Specifically, within MHMD we have extended an existing homomorphic encryption scheme, called the MORE scheme. The original MORE scheme can be safely applied on integer numbers but provides lower security on real numbers. Thus, we have introduced the *Hybrid MORE encryption scheme*, allowing for safely performing operations on both integer and real numbers. It relies on an additional obfuscation layer based on polynomial evaluation maps and is fully homomorphic with respect to algebraic operations: addition, subtraction and multiplication.

UTBV Category Winner of the 2019 Innovation Radar Prize

The *Innovation Radar* (IR) is a European Commission initiative to identify high potential innovations and innovators in the context of EU-funded research and innovation projects. The competition is based on a pre-selection of the most relevant innovations within four categories, namely (1) *Tech for Society* (i.e., technologies impacting society and citizens), (2) *Innovative Science* (i.e., cutting-edge science underpinning tomorrow's technological advances), (3) *Industrial & Enabling Tech* (i.e., the next generation of tech and components supporting industry), (4) *Women-led innovations* (i.e., recognizing dynamic women developing and leading great innovations with EU-funding). Then, a list of 3 finalists for each category is selected through a public online vote, who present their innovation in front of a jury of experts, including investors and entrepreneurs, that ultimately selects the winners.

Figure 5. *Awarding ceremony of the EC's Innovation Radar Awards 2019, awarding UTBV as Category Winner 2019 for the Industrial & Enabling Tech category, during Research and Innovation Days in Brussels (24-26 September 2019).*

In 2019, the HE framework developed by UTBV, already finalist in the 2018 competition, was awarded as Category Winner 2019 for the *Industrial & Enabling Tech* category at the Research and Innovation

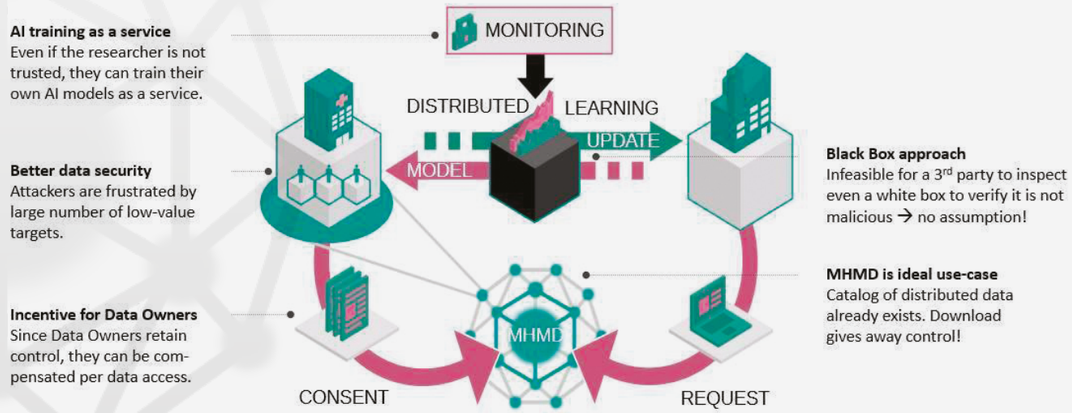


Figure 6. *Distributed deep learning with untrusted blackbox developed in the context of MHMD.*

Days in Brussels (24-26 September 2019), with the statement: *"This solution implements a software framework for developing personalized medicine solutions based on homomorphically encrypted data and artificial intelligence (AI). The framework ensures that the data remains private, and the performance of the AI models is not affected by the encryption".*

Distributed learning: shared computation among mutually distrustful parties

The precondition to machine learning is the availability of large amounts of data, usually unified into a centralised source. This requires handling, besides consent, privacy and security, data subjects' rights and other requirements, tedious and expensive tasks such as storage, transfer and curation, which all together make the sharing of data across different legislations arduous and sometimes even impossible.

To cope with such criticalities, MHMD developed an alternative approach based on shared, distributed computation among mutually distrustful parties, namely "distributed" or "federated" learning. Unlike traditional machine learning, that performs computation in a centralized location, distributed learning relies on multiple data providers who each hold a small part of the data and are willing to collaborate to perform machine learning with joint resources. This setup proves particularly suitable for MHMD, as its blockchain ensures to keep track of data access and handle consent, while the global metadata catalogue provides means to identify data in a large network of distributed data.

Traditionally, a data provider copies the data to a third party for machine learning directly. Once copied however, the provider has no physical control over what happens to the data. He must trust the party to follow contractual and legal obligations and to protect the data. By contrast, in distributed learning the data never leaves the premises of its provider, giving it full control over every single access.

In this manner, distributed learning also mitigates *the risk of data leaks*. A malicious party could target one large data repository with the intent to compromise the entire data set: that party is likely to invest many resources into the attempt, since the possible value of the compromised data is high. In a distributed setting, however, even a successful attempt on a data provider would expose only a fraction of the whole data set. Fortunately, the value of data increases exponentially with the amount of data: the malicious party is therefore much less likely to invest or concentrate resources on a single small data provider.

There is also an *economical advantage* of distributed learning. Acquisition, curation and annotation of data often represents a challenging cost for data users. In the distributed learning model, this cost is cut, but the data provider can bill a researcher for accessing the data and for the computation power spent for distributed learning. This has two positive effects. Firstly, it creates an incentive for data providers to build high-quality data sets, since those would be accessed more frequently. Secondly, it leads to continuing effort in growing and harmonizing existing data sets across data providers.

In addition, distributed learning allows for the so-called *"differential privacy"* guarantees: in other words, data providers may perturb data before each query, protecting the privacy of every individual in the data set while harming data utility only slightly. In MHMD, we developed a novel approach to distributed learning with a "black box" (Figure 6). The idea is that a third party may supply an executable deep learning model to an orchestrator as a black box that computes both a loss function and its gradient, and then the model is sent to data providers for local evaluation. Special care is taken to execute third-party software in an isolated environment and to monitor its output. Before communicating intermediate results for training back to the third party, results of several instances of isolated black-boxes are averaged by the orchestrator, making it extremely hard for the third party to gain any knowledge about individual data samples.

To know more

- Vizitiu, A., Ioan Niță, C., Puiu, A., Suciu, C., Itu, L.M. *"Towards privacy-preserving deep learning based medical imaging applications"*. IEEE International Symposium on Medical Measurements and Applications (MeMeA).
- Vizitiu, A., Nita, C.I., Puiu, A., Suciu, C. and Itu, L.M. *"Privacy-Preserving Artificial Intelligence: Application to Precision Medicine"*. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

SYNTHETIC DATA GENERATION: AN ALTERNATIVE SCENARIO FOR PRIVACY PRESERVATION IN MHMD

Anna Rizzo, Davide Zaccagnini, Edwin Morley-Fletcher
// LYNKEUS

Minos Garofalakis
// ATHENA RESEARCH CENTER

Aaron Lee, Michael Jennings, Steffen Petersen
// QUEEN MARY UNIVERSITY OF LONDON

Rudolf Mayer
// SBA RESEARCH

The rise of AI as a disruptive innovation force in the global economy has created an unprecedented demand for large, heterogenous and integrated data sets, increasing the friction between privacy considerations and the information need of modern machine learning. Despite the GDPR mandates for the use of privacy protection measures being specified in an intentionally broad way to allow local innovation, privacy preservation remains a daunting task in medical data-driven research. In the previous articles we have seen that, under a *statistical disclosure control* framework, methods exist to limit the risk of a person being re-identified from published data. In MHMD we have applied a variety of counter-measures, from an initial level of permissioned data publishing (i.e., guaranteeing k-anonymisation of published data), to further levels of privacy-enhancing technologies, like enacting the so-called "visiting mode", where secure computation *"bringing the algorithms to the data"* is provided through homomorphic encryption, secure multi-

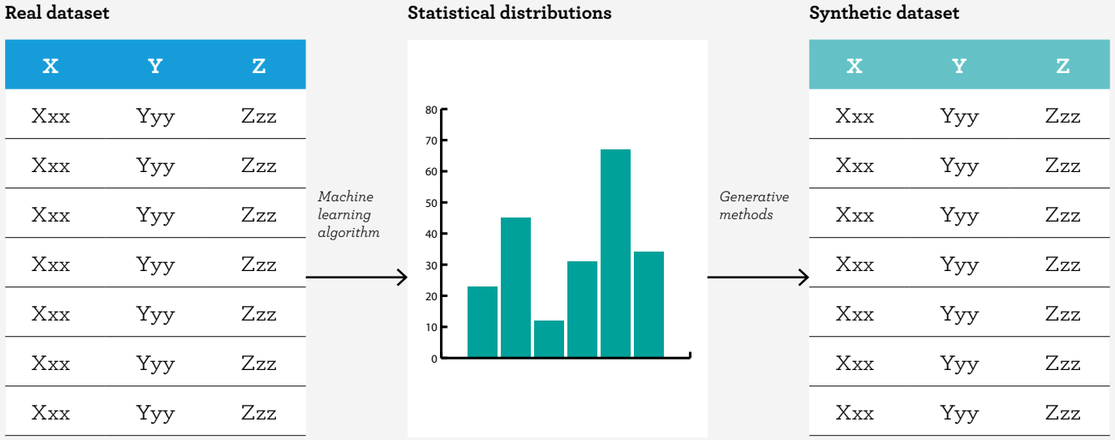


Figure 1. *Synthetic data is generated from a real dataset by "learning" statistical features of a known population by the application of machine learning algorithms. Then, using the same statistical distributions, new data is created "from scratch" retaining global properties of the original datasets without directly using the individual data.*

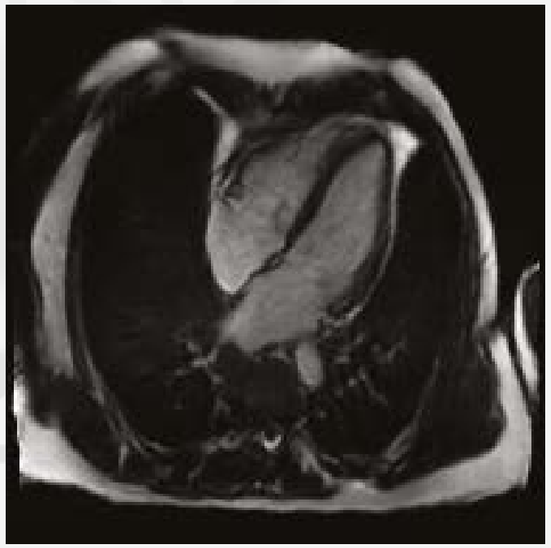


Figure 2. *Synthetic cardiovascular magnetic resonance image generated by the QMUL team from real UK Biobank data (also with the support of the "SmartHeart" EPSRC programme grant, EP/P001009/1).*

party computation or federated learning. While robust in its own merit, and operating under well constrained scenarios, though, even secure computation suffers from a relative lack of scalability for the data volumes needed in modern R&D ecosystems. In this landscape, another interesting approach to safe data sharing was identified during MHMD in the generation of synthetic data. What is this about?

Synthetic data: "machine learning enabling machine learning"

By definition, synthetic data is artificially produced data that replicates – through, for example, machine-learning algorithms – certain predetermined statistical characteristics of a real population by using features "learned" from the original data. In this process, global properties of the real dataset are retained without directly using the individual data, preventing the identification of the original data subjects. This data meets the GDPR specification of anonymous, namely of *"personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable"* (Recital 26 GDPR). This anonymous data can then be used for research and machine learning, such as the training of AI algorithms: in this sense, we can say that we have *"machine learning enabling machine learning"*. The breakthrough is that synthetic data uncouples sensitive information from the data information content, attaining anonymity while still preserving sufficient information richness. Initial evidence in MHMD suggests that, in applications such as clinical decision support tools and in-silico clinical trials, synthetic data might yield useful results when compared to those generated using the original data. The value of synthetic data resides in a series of key characteristics:

1. these datasets can be constructed to be used much like the original data sets and therefore use the same processing infrastructure;
2. they can maintain certain statistical characteristics of the original data and may be tailored to adjust for biased or incomplete original data sets;
3. while, in general, due to the effectively unlimited nature and type of synthetic data, the actual risk or re-identification cannot always be effectively quantified, in certain cases, differential privacy techniques can provide specific mathematical qualifications.

The US National Institute of Standards and Technology in fact launched a Differential Privacy Synthetic Data Challenge in 2019, specifying that *"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"*.

By virtue of its scalability and anonymity, artificially-generated data show the ability to jump-start AI development in areas where data is scarce or too expensive to obtain in volume, such as the biomedical sector which, despite the explosion in data collection devices, suffers from both economic and legal limitations when it comes to sharing that information.

Synthetic data in MHMD: health records, medical images and beyond

In MHMD, Queen Mary University of London (QMUL) proposed the use of synthetic data as a proxy for clinical data, to support prototyping of the infrastructure, training of clinical machine learning algorithms and penetration testing. As their value became more generally apparent, the Consortium decided to make synthetic data a key element of our data privacy-protection research and innovations. To this aim, the synthetic medical records and images generated by QMUL were utilised to test clinical decision support applications (see Section 3). In particular, deep learning techniques (i.e., generative adversarial networks, GANs) were applied to generate synthetic cardiac magnetic resonance images of the heart and surrounding anatomical structures, starting with a large sample population from the UK Biobank (under access application 2964). This demonstrated that it is possible to generate realistic images which convincingly reproduce key anatomic structures. Notwithstanding the need to improve these algorithms to remove artefacts, the potential use of synthetic data in AI remains considerable. As AI is particularly effective in imaging processing, databases of clinical images can be used to generate synthetic ones which can then be applied in AI training and knowledge discovery. As part of MHMD, SBA Research has, in addition, performed and published a thorough assessment of the utility and robustness of synthetic data balanced against the residual risk of attribute disclosure (i.e., the leakage of sensitive attributes predicted on the basis of prior knowledge), assessing optimal uses of this type of data.

> To know more

- Hittmeir, M., Mayer, R., Ekelhart, A., *"A Baseline for Attribute Disclosure Risk in Synthetic Data"*. Proceedings of the 10th ACM Conference on Data and Application Security and Privacy (CODASPY 2020)
- Hittmeir, M., Ekelhart, A., Mayer, R., "Utility and Privacy Assessments of Synthetic Data for Regression Tasks". Proceedings of the IEEE International Conference on Big Data (IEEE BigData 2019).
- Hittmeir, M., Ekelhart, A. & Mayer, R., *"On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks"*. Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019).

ASSESSING SYSTEM SECURITY: THE INTERNAL PENETRATION TESTING AND PUBLIC HACKING CHALLENGE

Enrico Cambiaso, Ivan Vaccari, Maurizio Aiello
// INSTITUTE OF ELECTRONICS, COMPUTER AND TELECOMMUNICATION ENGINEERING, NATIONAL RESEARCH COUNCIL (CNR-IEIIT)

Rudolf Mayer
// SBA RESEARCH

System security has represented a key priority for the development of the MHMD platform, given the high level of sensitivity of healthcare and wellbeing-related data from healthcare institutions and individual patients poised to be shared and exchanged within the network. For this reason, besides well described approaches for privacy-preserving data publishing and secure computation, the Consortium has elaborated a strategy for assessing the security of the MHMD architecture, considering both the platform as a whole and its different components individually.

To perform such assessment, the Consortium first tested system security and privacy internally, hence organised a dedicated penetration challenge. Such challenges are adopted to validate the ability of a connected system to counter cyber-attacks, exploiting digital mediums and communications (e.g., network attacks) executed to steal sensitive information.

The penetration testing effort has been entrusted to the *Institute of Electronics, Computer and Telecommunication Engineering* of the *Italian National Research Council* (CNR-IEIIT) and the Austrian research centre *SBA Research*, given their consolidated academic record in information and network security.

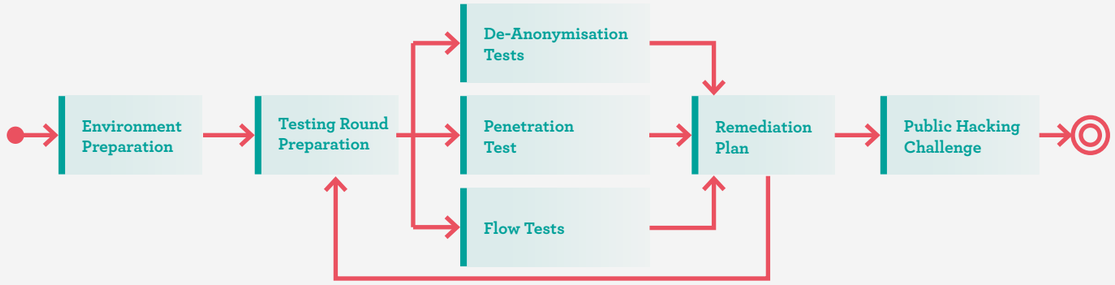
System security testing: the MHMD protection plan

The chosen approach for the system security assessment has been based on two pillars: (1) *repeated cycles of internal penetration tests* perpetuated by entrusted partners (CNR-IEIIT, SBA), accomplished through a "test and fix" cycle, followed by reporting and debugging, and (2) *a public hacking challenge* open to ethical hackers all over the EU and beyond, invited to try to break the MHMD system and its components, identify and exploit vulnerabilities, leak potentially sensitive data and report their results, followed by another remediation phase.

Preliminarily to this, some preparatory activities has been foreseen. A test environment, separated from the actual system, has been prepared and populated by synthetic medical datasets of virtual patients, generated *ad hoc* from real patients' data to avoid privacy loss in case of system breaches. Attacks to such data, after proper anonymisation, were conceptualized, in order to address privacy breaking threats.

Moreover, a series of internal security tests, reporting and remediation were performed to the different components (*individual testing*) as the system as a whole (*integration testing*), including vulnerability and penetration testing operations, aimed to identify system weaknesses and exploit them to access or break the platform, that provided a set of security improvements before the public exposure.

Figure 1. *The MHMD protection plan. The scheme describes the overall procedure adopted to test the system security, starting from synthetic dataset generation and creation of the host environment, to internal penetration test and remediation phases, to the public hackathon and the second reporting and remediation phase.*



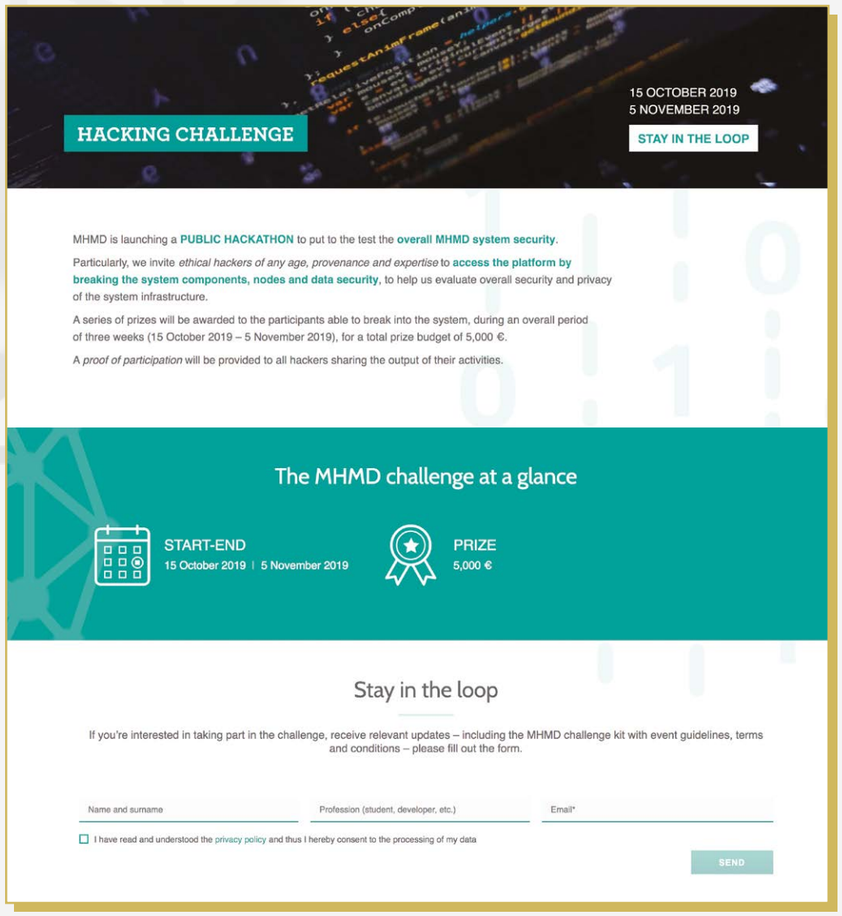


Figure 2. *Preview of the MHMD website page dedicated to the public hackathon, with the possibility for interested IT security specialists to subscribed at kept up to date with materials and deadlines.*

The public hackathon: MHMD put to the test

The public hacking challenge was organized to make the overall MHMD system and its components evaluated and tested against cyberattacks coming from external subjects. The challenge was organised between 15 October and 5 November 2019. Showcased through a dedicated page on the MHMD website and social media, the MHMD hacking challenge invited ethical hackers to break the system components and nodes (e.g., to escalate privileges, exploit vulnerabilities, identify software bugs, etc.) and report results to the dedicated team at CNR-SBA, with a series of prizes with an overall budget of 5,000 €.

The results of internal testing and hacking challenge

The internal testing allowed to perform iterative cycles of system debugging, removing potentially harmful bugs including obsolete libraries and certificates, but also more substantial vulnerabilities such as authentication issues, username management, not sufficiently strong crypto-algorithms or password policies. Through the debugging phases, such issues have been sorted before the public hacking challenge. As a result, the hacking challenge didn't really result in substantial findings in terms of vulnerabilities, such as the possibility of finding sensitive information through the UI server and mapping the server itself for finding possible vulnerabilities. The issues have been immediately resolved, thus allowing the release of the final MHMD platform for the activation of the nodes at the participating hospitals, and the recruitment of individual users for testing the app and the whole individual data sharing and consent management pipeline.

SECTION 03 >

LEVERAGING THE VALUE OF BIG DATA IN HEALTHCARE

HARMONISING DATA SOURCES AND DEVELOPING ADVANCED ANALYTICS FOR DE-IDENTIFIED MEDICAL DATA

ENHANCING DATA DISCOVERABILITY WHILE PRESERVING PRIVACY: THE MHMD (META)DATA CATALOGUE

Romain Tanzer, Douglas Teodoro, Emilie Pasche, Patrick Ruch
// UNIVERSITY OF APPLIED SCIENCES AND ARTS OF WESTERN SWITZERLAND (HES-SO)

The MHMD platform recognizes four main stakeholders in the data value chain - hospitals, citizens, research centres and industry - with different interests. While citizens and hospitals will share very heterogeneous and privacy-sensitive datasets in the network, research centres and industry need streamlined and homogeneous ways to search, discover and access these datasets.
In this context, our research team at *University of Applied Sciences and Arts of Western Switzerland* (HES-SO) developed a data catalogue, fed with metadata relevant to the datasets registered on the blockchain. The (meta)data catalogue main goal is to give the user a view of the data available on the MHMD platform as well as enable searching for records when a set of keywords are given. Once one or more datasets of interest are identified, the catalogue allows users to request access to data through the blockchain, which by means of automated smart contracts matches the data request with available datasets and relevant consent preferences. If there is correspondence between the defined consent options and the intended usage, the blockchain provides an off-chain download link to the data user.

Securing patient metadata: access control and data masking

Metadata describing the various datasets shared within the MHMD network contain critical information on patients, which needs to be adequately handled to avoid compromising privacy. To face this problem, we developed additional security measures to reduce the risk of patient identification.
One of our first measures was to introduce, from the API side, an extra verification step of the access rights to the metadata called "token authentication". For instance, each user would have first to authenticate through the web service by passing its credentials (username and password), which go to the authentication server. The server verifies the credentials and if it is a valid user, then it returns a signed token to the client system. This signed token is then used to authenticate the user request with the catalogue, enabling the request to be processed and the response to return to the user.
Residual privacy risks are related to the level of detail of the metadata within the general data structure, called 'granularity'. For instance, when breaking down information on a patient using multiple queries (e.g., gender, age, type of disease), some information could be revealed, and the patient potentially identified. In the case of rare diseases, for instance, it would be straightforward to identify a patient with queries combining rare diagnosis and some demographic filters. In order to mitigate this type of issue, we reduced the search granularity by hiding datasets corresponding to query results of less than 10 cases.
Despite the implementation of this "minimum results by query" concept, there could still be ways to identify individual patients.

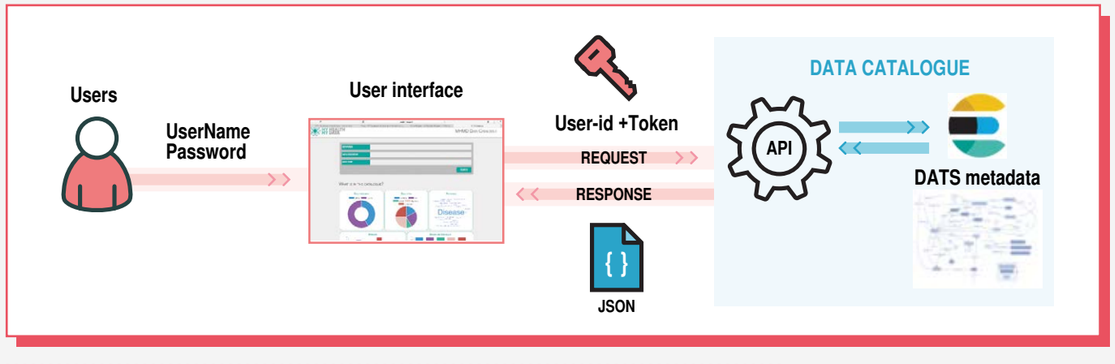


Figure 1. *Web service communication with the catalogue through the API using a token authentication.*

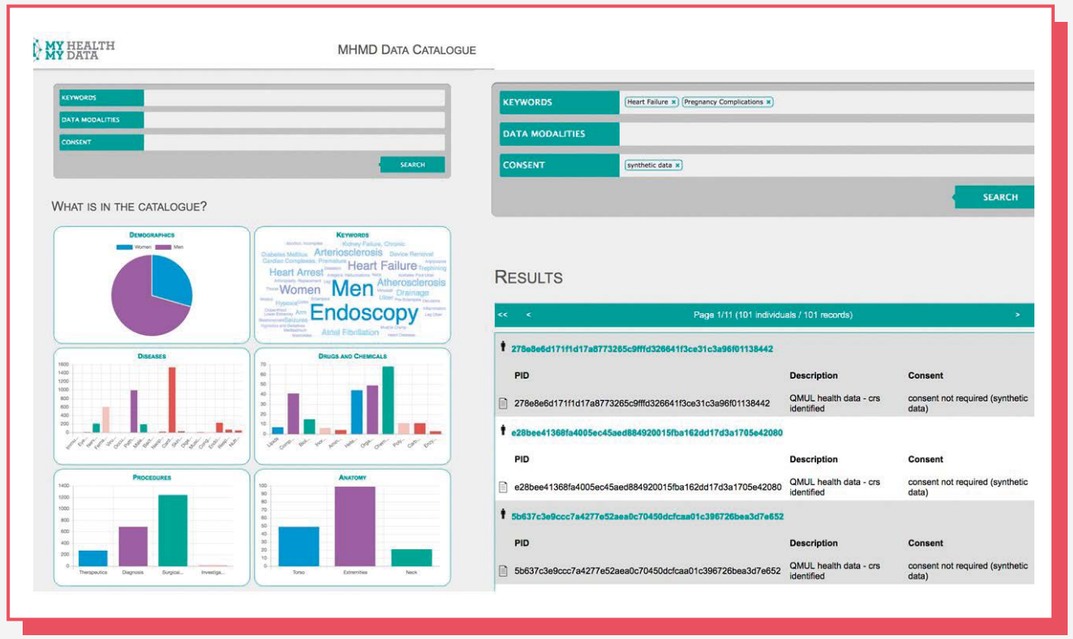


Figure 2. *Preview of the user interface to query the catalogue.*

For instance, by combining multiple queries together (e.g., gender, age, type of disease), it is possible to isolate a patient as the common result of all queries, and deduct the information corresponding to this patient.
However, this cross-queries identification at large scale requires a large amount of queries and a programmatic way to query the catalogue. To prevent this possibility, we set the rate limit of requests to 10 requests per minute per user. This will avoid a scraping of the catalogue and will give us time to detect and block any suspicious behavior.

Enhancing data discoverability: semantic and data-driven query expansion

Due to the heterogeneity of datasets and the existence of various possible ways to search for the same information, the use of the so-called 'query expansion' is fundamental to enhance dataset discoverability. The process of semantic query expansion consists of selecting and adding terms to the user's query with the goal of minimizing query-document mismatch and thereby improving retrieval performance. Two query expansion methods have been implemented: the former makes use of medical subject headings (MeSH) terminology, a comprehensive and controlled vocabulary of label terms used to index journal articles and books in the life sciences; the latter uses word-embedding based on PubMed corpus; these allow to expand, and thus better specify, an initial request with terms that are semantically related to the query.
The query expansion using MeSH has some limitations. For instance, misspelling of search terms is not considered in the query, and synonyms stay limited to the MeSH terminology, which does not represent the full spectrum. To address these limitations, the use of data-driven expansion methods, such as those based on word embeddings, which use GPU accelerators for training (due to their intensive computational needs), is complementary to the complexity of biomedical terminologies. Our approach includes a query expansion module that computes the word embedding for query terms and performs their expansion using the k-nearest embedded word vectors. The word embeddings represent each word as a dense vector of real numbers, where words that are semantically related to one another map to similar vectors.
Overall, the combination of these two types of query expansion methods improves significantly the number of relevant datasets returned by the API, providing users with additional options to include datasets in their study.

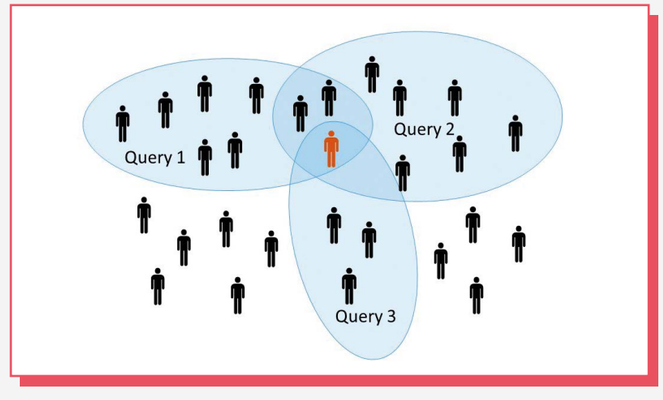


Figure 3. *Cross-queries identification, the patient can be identify as the common result of all queries.*

CLOSING THE VALUE CYCLE: IMPROVING CLINICAL CARE WITH THE MHMD PRIVACY-PRESERVING, PATIENT DATA SYSTEM

Martin Kraus and Andre Aichert
// SIEMENS HEALTHINEERS

The MHMD platform is designed to provide access to vast amounts of patient data for a variety of research and development uses, including the training of *artificial intelligence* (AI) solutions. The clinical value of such tools, though, is only generated once they are used in the hospital on actual clinical cases. The goal of Siemens Healthineers has been to demonstrate the added value of MHMD in the development of advanced analytics technologies, such as deep learning, by providing a secured, efficient data environment allowing both development of AI tools and their use in the clinical setting, thus closing the cycle between clinicians and concrete solutions. To this end, we implemented two prototypical tools for building and deploying AI models, namely *DeepExplorer* and *DeepReasoner*.

The *DeepExplorer*, on the one hand, represents the machine learning process from the point-of-view of the researcher, who needs to define a model, obtain data, train the model and finally select the best-performing solution (i.e., hyper-parameter search). The *DeepReasoner*, on the other hand, represents the point-of-view of the clinician, who can be supported in the medical decision-making process by AI models running on patient data acquired in the clinical routine.

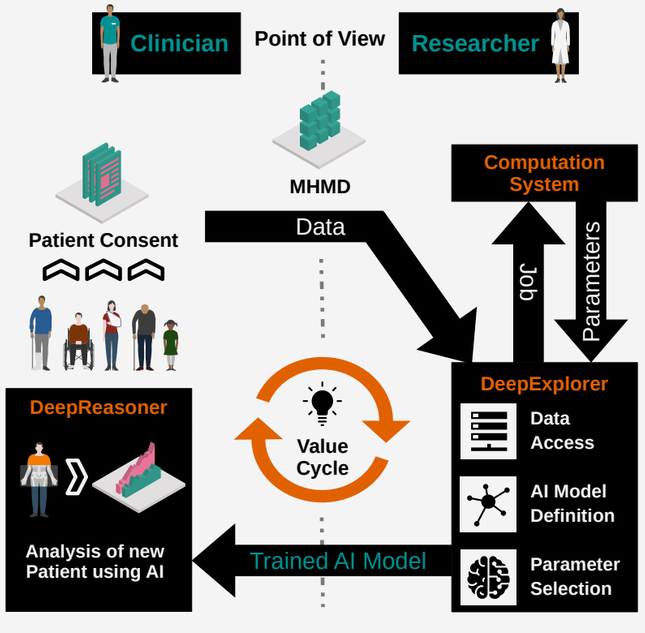


Figure 1. *Schematic overview of the DeepExplorer/DeepReasoner workflow, illustrating how value is created at the interface of clinician and researcher. DeepExplorer provides advanced tools for building AI models to researchers. DeepReasoner provides easy access to tools based on these models in a clinical setting.*

A perfect pair: DeepExplorer and DeepReasoner

DeepReasoner was originally a web-based interface with a cloud-based backend, developed in the context of the FP7 MD-Paedigree project (2013-2017). The tool was aimed at supporting physicians in the clinical decision process by finding cases similar to the one they are working on. Within MHMD, the scope of the tool was expanded into a platform to run any AI model, while a second pillar was added to support researchers in defining and training new AI models. The resulting design complements two web-based tools (Figure 1). *DeepExplorer* can be used by researchers to obtain data, build a model and deploy it into *DeepReasoner*. The new and improved *DeepReasoner*, in turn, enables clinicians to use the deployed models to make inferences on local data. This general concept may have various concrete embodiments. One instance may simply be to summarize patient data with an insightful visualization, while another may be to use clinical images to detect tumors or segment structures of interest.

In any case, it is here that the true potential of improving clinical care through aggregate knowledge is unlocked. The availability and continuous improvement of such solutions acts also as an incentive for hospitals to extend the MHMD network, by providing more data and encouraging patients to consent to research-related use of their data. About 50 patients have so far been engaged to share their clinical data from their personal data platform (Medicus, Digi.me) further enriching the pool of data available to third parties.

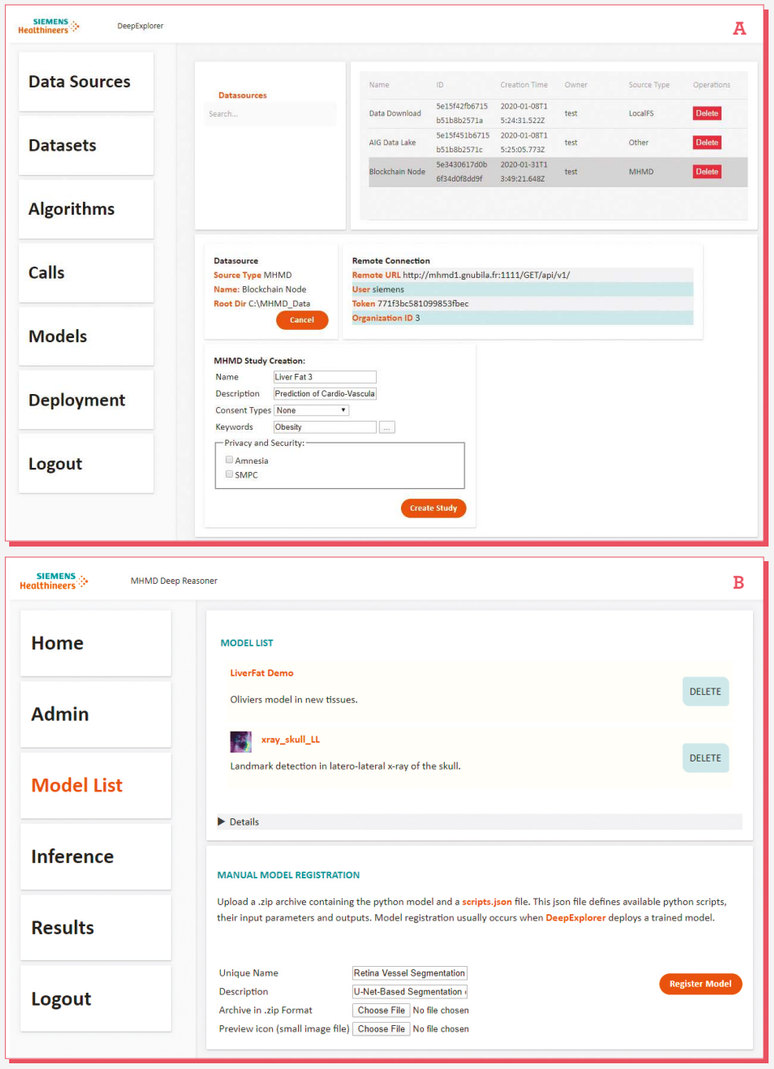


Figure 2.
a) Preview of the DeepExplorer interface.
b) Preview of the DeepReasoner interface.

The point-of-view of a researcher

The MHMD system provides access to patient data with appropriate consent, controlled and enforced through the blockchain. A researcher can use the central catalogue to find and select an appropriate population of interest. *DeepExplorer* allows to define a study through a local MHMD blockchain node, which grants secure, authenticated access to the dataset of interest, if the consent registered by the hospital, or the individual, matches with the intended access rights and intended usage. Data to train, test and clinically validate AI tools can now be securely and efficiently accessed, at much lower transaction costs and in full compliance with GDPR and national regulations, addressing one of the main obstacles to scaling up medical-AI solutions implementation. AI experts can define their own model for a specific task, which then becomes available for non-expert users. *DeepExplorer* performs a highly automated model training process and hyper-parameter search. Special care was taken to ensure that algorithm hyper parameters are exposed in an intuitive way for configuration even by non-expert users, and that the security processes and technologies do not substantially lessen the informative power of the data, as well as the ability to perform effective machine learning on them.

For training and hyper-parameter search, a distributed computation system automatically assigns tasks to available worker nodes, and each is provided a compute job along with data from the MHMD system. A trained model is created for deployment in one or multiple *DeepReasoner* instances. At this point, an appropriate presentation of the model input and output must be defined to enable clinicians to use the tool and interpret the results. While experts can define their own template, non-experts can use a general upload/download mask or use a pre-defined template, e.g., for classification or detection tasks (Figure 2a).

The point-of-view of a clinician

DeepReasoner is a web server associated with *DeepExplorer* and exposes model deployment functionality: the trained model is copied to the *DeepReasoner* and then served locally. The front end supports selection and interactive manipulation of local input data, such as patient scans or clinical reports. Once the patient input is defined, the trained model can be executed to provide an inference result. This could be anything of value in clinical decision making, e.g., a diagnosis, the identification of a detected lesion in a scan, the visual segmentation of such lesion against healthy tissue or a set of patients with similar clinical characteristics. The template defined by the researcher determines how this inference result is presented to the clinician, e.g., a diagnostic code or a graph of diagnostic probabilities, a cropped image around a tumor location and so on. This mechanism streamlines the technical workflow for the AI researcher, while supporting the flexibility to exchange any application-specific input data selection and output data visualization (Figure 2b).

PERSONALIZED PHYSIOLOGICAL MODELING FOR CLINICAL DECISION SUPPORT

Anamaria Vizitiu, Andrei Puiu, Cosmin Ioan Nita, Lucian Itu
// TRANSILVANIA UNIVERSITY OF BRAŞOV (UTBV)

As cardiovascular disease remains the major health burden worldwide, increasing efforts are being made for developing personalized, analytics-based approaches for early diagnosis, surgical planning and risk stratification. Among those, blood-flow computations are employed, in conjunction with patient-specific anatomical models extracted from medical images, to draw an accurate profile of a patient's cardiovascular system structure and functionality. In the context of MHMD, the team of *Transilvania University of Braşov* (UTBV) has been committed to the development of a customizable cardiovascular circulation blood flow model, built on a set of initial clinical measurements and a set of continuous measurements derived from wearable applications.

Modelling the human blood flow: the whole-body circulation model

The first version of the so-called *whole-body circulation* (WBC) model was originally developed in the context of the MD-Paedigree project (2013-2017), aimed at the design and clinical validation of patient-specific predictive models in pediatric cardiology. The model allowed for the personalized computation of patient-specific hemodynamic indicators (Table 1), such as systemic circulation properties. The WBC model integrates various components: (1) a heart model including ventricles, atriums and valves, (2) the systemic and (3) the pulmonary circulation (arteries, capillaries, veins).

In MHMD, the model was expanded in several directions to allow the incorporation of individual data from mobile apps and IoT devices through the MHMD secure, privacy-preserving data exchange functionalities, including the development of an enhanced personalization framework, an AI based real-time approach for computing the output measures of interest and the usage of encrypted input data.

The WBC model can be run therefore under patient-specific conditions, simulating different physical states (e.g., rest, exercise) to compute relevant measures of interest. However, model parameters need to be personalized according to the individual patient's condition. To do so, the personalization process consists of two sequential steps. First, a series of parameters are computed directly. Next, an automatic optimization-based calibration method estimates the values of the remaining parameters, ensuring that the personalized computations match the measurements.

NAME	DESCRIPTION
Arterial resistance	The resistance that must be overcome to push blood through the circulatory system and create flow.
Arterial compliance	The tendency of the arteries and veins to stretch in response to pressure. Blood vessels with a higher compliance deform easier than lower compliance blood vessels under the same pressure and volume conditions.
Dead volume of the left/right ventricle	Intersect of the end-systolic pressure volume ratio with the volume axis in the PV loop.
Stroke work	The work done by the ventricle to eject a volume of blood.
Ventricular/atrial/arterial elastance	A measure of the contractility.
Arterial ventricular coupling	The interaction of the LV with the arterial system, providing important mechanistic insights into the complex cardiovascular system and its changes with aging in the absence and presence of pathologies.
Pressure-volume loop	A plot of the ventricular pressure versus ventricular volume which has long been used to evaluate the work done by the heart and its efficiency.

Table 1. *Blood flow properties that can be computed through the WBC model developed by UTBV.*

Training deep neural networks for real-time hemodynamic analysis

The initial model was computationally very efficient (i.e., with a single forward runtime of about 10^{-3} seconds) but its personalization required hundreds of forward runs, leading to an overall computation time of 30 – 60 seconds for determining the patient-specific measures of interest. To address this limitation, we employed a deep neural network, which led to a significant acceleration of computation time, enhancing its clinical applicability. Such systems, at the same time, require large training data repositories to be assembled and curated to allow efficient calibration of the network parameters. This substantial drawback, which is one of the main systemic issues in medical-AI R&D operations, was addressed with the development of a database of 10,000 synthetically generated data samples reflecting anatomical and functional variations of healthy and pathological cases. The sample generation logic relies on pre-defined normal ranges of hemodynamic parameters, covering a wide range of anatomical variations. The sampling procedure is further constrained by a set of well-defined consistency rules to achieve physiological plausibility, e.g., left and right ventricular similar stroke volume. The sampling procedure was constrained by a set of consistency rules to achieve physiological plausibility (e.g., left and right ventricular similar stroke volume). Then, the lumped parameter model has been employed for computing the personalized output measures of interest, representing the ground truth information to be predicted by the deep learning model. Following standard approaches, 8,000 randomly selected data samples were used for model training and the remaining 2,000 for subsequent testing.

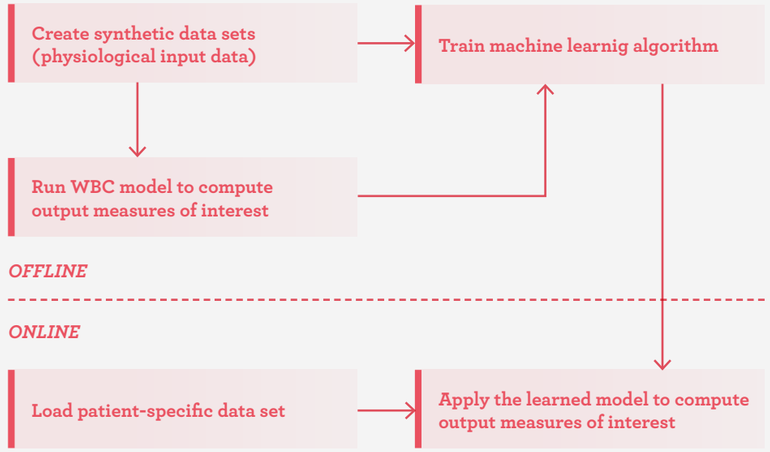


Figure 1. *Overall workflow of the proposed deep learning-based model. A deep neural network trained offline on the synthetic data created ad hoc for the purpose.*

Evaluating the prediction of time-dependent and -independent quantities

As the personalized measures of interest fall into two categories, time-independent and time-dependent quantities, we defined two specialized predictive neural networks. To assess the fidelity of the AI model in predicting the time-independent quantities, we computed on the test set the *mean absolute percentage error* (MAPE) and Pearson correlation coefficient, measuring consistency between predictions and real data. The resulting MAPE was low (2.7 % on average), and Pearson correlation values were all above 0.99, indicating an excellent model performance (Figure 2). In one example (Figure 3) we compared 10 time-dependent quantities of interest as computed by the WBC model and predicted by the AI model, obtaining low prediction errors (mean absolute error of: 0.83mmHg for pressure related measures, 1.6ml for volume related quantities and 4.7ml/s for flow related quantities).

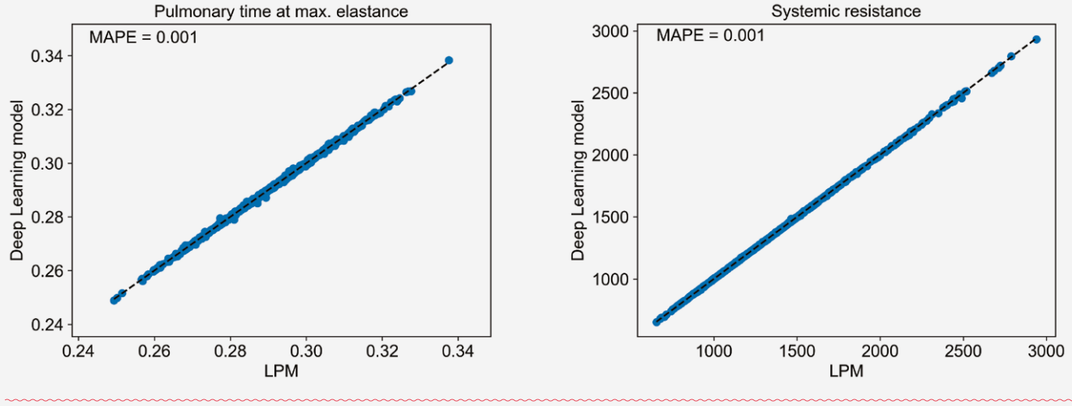


Figure 2. *Scatter plots of model predictions versus ground-truth for two parameters. Left: time at max. elastance in the pulmonary circulation. Right: systemic resistance (i.e., total resistance of the arterial system).*

Clinical applications

The developed solution is able to assess the evolution in time of diverse quantitative heart cycle indicators, useful to evaluate a patient's cardiac function. One illustrative application is the non-invasive, real-time computation of pressure-volume (PV) loops, such as the ventricular PV loop (Figure 3l), that illustrates important measures of the heart and systemic circulation (e.g., stroke volume, cardiac output, ejection fraction, myocardial contractility or cardiac oxygen consumption). As pathologies such as left ventricular hypertrophy, dilated cardiomyopathy, aortic and mitral valve stenosis and regurgitation are manifested in the PV-loop, its efficient estimation would represent a powerful diagnostic tool for clinicians complementing echocardiography exams. The work demonstrates, in addition, value and reliability of synthetic medical datasets to efficiently train machine learning modules in absence of large, curated original data sets. Such training, in our case, was proven to be as accurate as that on actual patient records and substantially less expensive and more secure from privacy and security stand points.

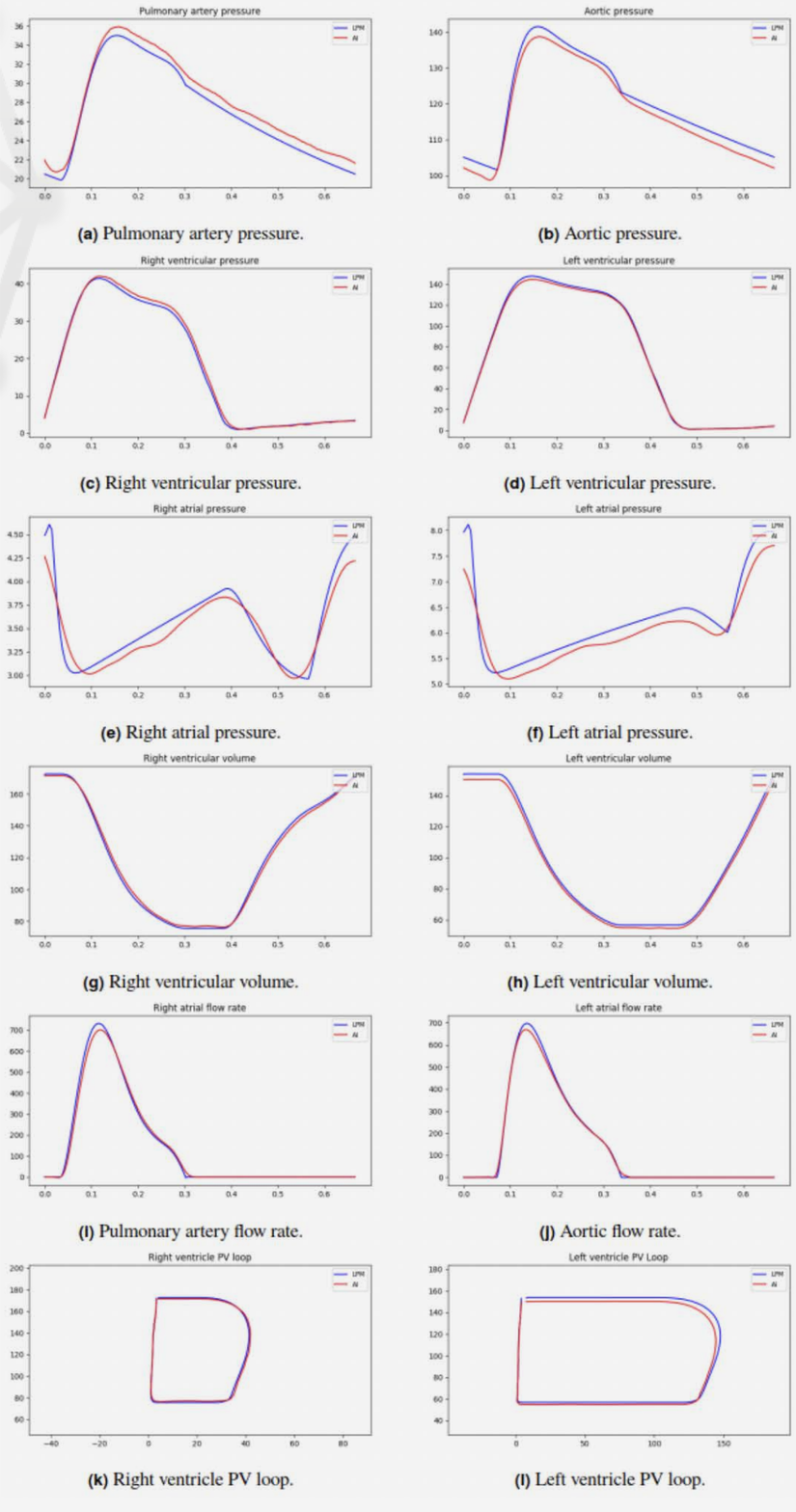


Figure 3. *Time-varying quantities of interest for the whole-body circulation model. Comparisonof deep learning-based predictions versus ground truth (x represents the time in (i) –(j) plots and pressure in (k) –(l) plots, y axis represents volume in (k)–(l) plots).*

ESTIMATING THE INFORMATION HIDDEN IN DATASETS

Matt Jeffryes, Valentine Rech de Laval, Douglas Teodoro, Patrick Ruch
// UNIVERSITY OF APPLIED SCIENCES AND ARTS OF WESTERN SWITZERLAND (HES-SO)

Data can carry more information than it looks, especially to layman users sharing content from their medical records. Similarly, healthcare institutions willing to engage into a research cooperation may not be able to assess the relevance of a given subset of data. According to the GDPR, informed consent is only possible when data subjects can understand the consequences of sharing personal data, but it is often difficult to turn such an understanding into actionable knowledge.

Many patients have little concern about sharing their medical diagnoses, but in some cases (e.g., HIV+ patients) they might prefer to keep their medical history confidential. In the MHMD data catalogue such patients can prevent access to their diagnosis, however other information could allow to infer their status. Most HIV patients, for instance, are treated with anti-retroviral therapy or drugs targeting HIV-related comorbidities (e.g., Kaposi's sarcoma) and if they share their prescription history, they may inadvertently disclose their HIV status. For this reason, data subjects should be informed about how two subsets of the data they are about to share are associated. An individual might think twice before sharing a certain set of drug prescriptions if he knows that, by doing so, e.g., about 60% of his diagnosis information can be automatically recovered.

Test dataset and experimental design

To demonstrate the possibility of hidden information being disclosed, we employed a test set of published clinical cases retrieved from the MEDLINE digital library, which didn't bear any legal or ethical constraints. As an analogue to EHR-associated medical encodings (e.g., diagnosis, drug prescription, surgical or diagnostic procedures) we used *medical subject headings* (MeSH) assigned to biomedical literature by the US National Library of Medicine. MeSH is a hierarchical vocabulary of terms attached to entries in the PubMed science literature database, based on the major themes of the publication. For

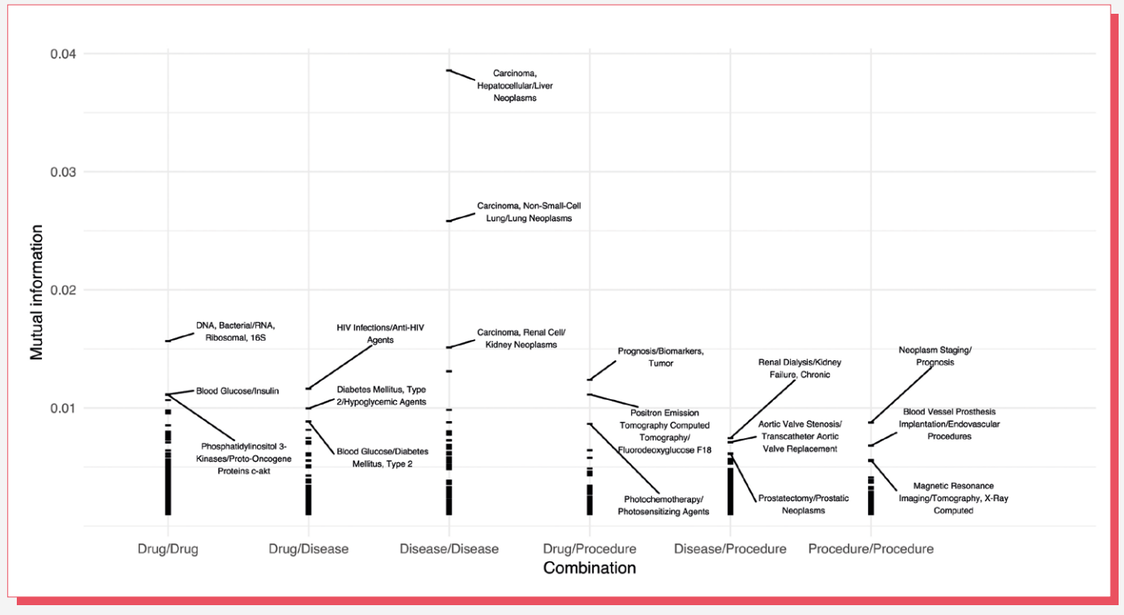


Figure 1. *The distribution of the mutual information of every pair of terms in the MeSH dataset falling into the drug, disease or procedure categories. The description of the three pairs with the highest mutual information in every combination of category is shown.*

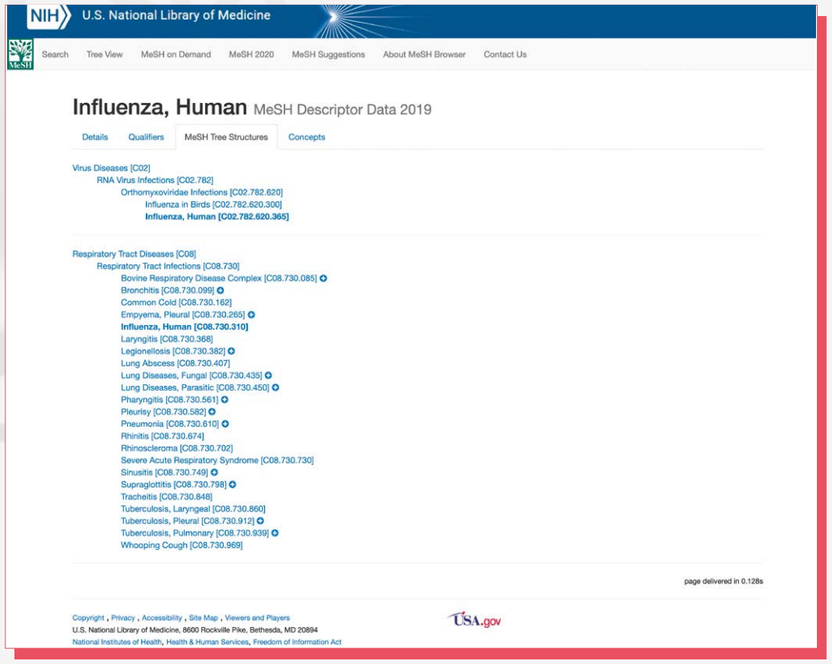


Figure 2. *Medical subject headings (MeSH) descriptors from the US National Library of Medicine.*

example, an epidemiology study on the influenza vaccine might be assigned the MeSH terms "Human Influenza", "Health Policy", "Vaccination". MeSH headings are part of a tree-like structure. For example, "Human Influenza" is part of the disease tree and is categorised under "Respiratory Tract Infections". Just as a drug and a disease may have a strong association, we assumed that their relevant MeSH terms would maintain the same association; for example, we expected the MeSH terms "Human Influenza" and "Oseltamivir" (Tamiflu) to appear together quite frequently. Therefore, we used the MeSH terms covering diseases, drugs and medical procedures as a testbed for investigating how hidden information can be identified and revealed.

Calculating mutual information between variables

A fundamental way of quantifying the "hidden" information contained within data is *mutual information* (MI). Formally, this is a measure of the amount of information that a random variable contains about another random variable. If two events are completely independent (e.g., the results of flipping two different coins) they have zero mutual information. Mutual information rises as much as the two variables are associated and thus provide information about each other.

In the case of the example dataset, we can treat the presence/absence of a given MeSH term in a publication as a binary random variable. The MI for a pair of MeSH terms will increase if they appear very often together, but infrequently without each other. MI is useful for identifying meaningful co-occurrences (as opposed to those which occur by chance) because it can be calculated for a wide range of data, requiring only counts of the entities appearing separately and together. It does not require any deep analysis of text, making it language independent.

We calculated the mutual information between pairs of MeSH terms which have appeared together in PubMed since 2018. This data contains 12,102 distinct terms describing a (1) drug or chemical, (2) a disease or diagnosis, or (3) a treatment or diagnostic procedure assigned to publications. The mutual information can be calculated for every pair of terms. To speed up the calculation, we only calculate it for terms that have been assigned to the same publication at least three times, of which there are 1,600,525 distinct pairs. Shown in Figure 1 is the distribution of the higher mutual information values for each of the possible combinations of categories of drug, disease or procedure. The pairs with the highest mutual information appear very naturally related; for example, the disease/procedure pair of *Chronic Kidney Failure* and *Renal Dialysis* and the disease/drug pair of *Type 2 Diabetes Mellitus* and *Hypoglycemic Agents*.

A web service for customised MI quantification

In MHMD we developed a web service for the estimation of hidden information encoded in a set of disclosed data for a given set of categories. The service is applied to the procedures, diagnoses and prescribed drugs found in patient health records, but the system could be easily applied to other kinds of data with minimal changes. As an example, a user can submit via web page or REST API a list of terms in a MeSH dataset, and the service returns the possible hidden information these terms encode, broken out by category. In the case of the MeSH term for Kaposi's sarcoma, the system returns as highest scoring diseases *Skin Neoplasms* and *HIV Infections*, and as highest scoring procedure *Highly Active Antiretroviral Therapy*. The service would be available to patients who are asked to disclose some or all of their health record to another stakeholder, e.g., a researcher. If patients are asked the drugs they have been prescribed, they could be shown the degree to which any disease they have been diagnosed with is a hidden information carried by their list of prescriptions; this will empower them to make an informed decision about sharing their personal data.

DISSEMINATION
EVENTS

MHMD has been extensively disseminated throughout its development by featuring at relevant conferences in the field of blockchain, AI and ICT technologies through presentations, panels and keynote session, for a total of about 80 events. A selection of the most significant ones in the last year is reported below

16th Meeting of the eHealth Network

Brussels, Belgium **28-29 NOVEMBER 2019**

This EC-organised meeting was aimed at discussing EU's priorities and programmes in digital health, including common exchange formats for electronic health records, best practices in the usage of health data, patient empowerment and access to data, digital service infrastructures, continuity of care, digital identification, cybersecurity, integration in national policies and sustainability. The PC Edwin Morley-Fletcher took part in the event within the *"Open eHealth Network"* session, presenting MHMD among other relevant initiatives (BigMedilytics, Smart4Health, Trillium II).

Beyond Privacy: Learning Data Ethics – European Big Data Community Forum 2019

Brussels, Belgium **14 NOVEMBER 2019**

The event, supported by the Big Data Value PPP and organized by five EU-funded research projects focused on privacy protection technologies, transparency and legal compliance (e-SIDES, SODA, SPECIAL, WeNet and MHMD) explored the most recent discussions about emerging ethical issues and provided practical guidelines in Big Data and Artificial Intelligence research. The PC Edwin Morley-Fletcher took part in the *Projects Panel* with two presentations: *"A GDPR-compliant blockchain-based system with advanced privacy-preserving solutions"* and *"Why have we preferred to opt for sharing synthetic data and for computation 'bringing the algorithms to the data'"*.

Convergence – The Global Blockchain Congress

Málaga, Spain **11-13 NOVEMBER 2019**

This global blockchain gathering explored the future of blockchain in AI, IoT, Finance, Mobility, Energy and many other fields, convergent trends in blockchain technology, regulation and research as well as blockchain impact on the economy, business and society. Representatives of Lynkeus (Edwin Morley-Fletcher), P&A (Lorenzo Cristofaro) and Athena RC (Minos Garofalakis) held the panel *"Blockchain and Healthcare: How is Blockchain Facilitating a Secure, Scalable, Data-Sharing Infrastructure in the Healthcare Industry?"* moderated by our former Project Officer Saila Rinne. A project exhibition booth was also set up to give attendees opportunities to get in contact and directly discuss MHMD challenges and innovations.

European Big Data Value Forum (EBDVF)

Helsinki, Finland **14-16 OCTOBER 2019**

EBDVF represents the main event of the European big data and data-driven AI research and innovation community, as the fusion of the former European Data Forum of the EC's Directorate-General for Communications Networks, Content and Technology and the Big Data Value Association (BDVA) Summit. The event explored AI, big data and robotics, data-driven bioeconomy, big data-driven smart connected factories, AI in public sector and smart cities, with the participation of the PC Edwin Morley-Fletcher (Lynkeus) as a full member of the BDVA.

European Research and Innovation Days

Brussels, Belgium **24-26 SEPTEMBER 2019**

The event brought together world leaders from industry, finance, academia and business to debate the future research and innovation landscape, mobilise EU citizens and increase awareness of its role in addressing societal challenges. The PC Edwin Morley-Fletcher (Lynkeus) took part in the session *"Data for Health"* reporting MHMD solutions for data privacy and security, with special focus on secure computing on encrypted data and synthetic dataset generation, namely *"A GDPR-compliant blockchain-based system for computation 'bringing the algorithms to the data' and for sharing synthetic data"*.

RDA 14th Plenary Meeting

Helsinki, Finland **23-25 OCTOBER 2019**

RDA Plenaries aims to bring together data experts in research, industry and policy-making from all around the world and from all disciplines. This year, under the theme *"Data Makes the Difference"*, the event explored the extensive ways data can address the extensive potential of research data in improving decision making, tackling grand societal challenges, engaging citizens in the creation of knowledge and betterment of society. Representatives of Lynkeus (Mirko De Maldè, Ludovica Durst and Edwin Morley-Fletcher), P&A (Lorenzo Cristofaro) and Athena RC (Yannis Ioannidis) played an active role within the Health Data Interest Group and the Blockchain Applications in Health Working Group.

Big Data: Fuelling the transformation of Europe's Healthcare Sector

Valencia, Spain **4-5 SEPTEMBER 2019**

The event, organised by the EU-funded BigMedilytics project, discussed advancements and latest innovations for the valorisation of Big Data in healthcare, involving key players such as healthcare providers, technology companies, payers, research institutes and academia from all across Europe. Representatives of the MHMD consortium, including Lynkeus (Edwin Morley-Fletcher), *P&A* (Lorenzo Cristofaro), Siemens Healthineers (Andre Aichert) and Athena RC (Minos Garofalakis) discussed the role of blockchain and advanced analytics applications in healthcare, with various presentations on GDPR compliance, secure computation, synthetic dataset generation, differential privacy, privacy by design, data anonymisation.

MyData 2019

Helsinki, Finland **25-27 SEPTEMBER 2019**

MyData is the flagship event of the MyData movement, a non-profit association and global network with the mission to empower individuals by improving their right to self-determination regarding their personal data. The conference provided interactive sessions, networking opportunities and inspirations to contribute to rebuilding trust and creating a more transparent and prosperous digital society. Among various panels, representatives of Lynkeus (Anna Rizzo) and UTBV (Anamaria Vizitiu) took part in the session *"Keeping control and minimizing risk in secondary usage of health data"* by reporting the general MHMD approach on data security and data subjects' privacy along with ad hoc innovations, homomorphic encryption above all.

BDV PPP Summit 2019

Riga, Latvia **26-28 JUNE 2019**

The BDV PPP Summit is among the most relevant events in big data and artificial intelligence innovation in Europe, gathering key industry leaders, academic representatives and policy-makers to foster cross-sector collaboration and shape strategies for European leadership in the field. The PC Edwin Morley Fletcher attended the event with a key session about presentation on the basic features of MHMD.

CONSORTIUM

LYNKEUS

ATHENA

IEIT

digi.me

gnúbila

Hes·SO GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

PANETTA & ASSOCIATI
STUDIO LEGALE

SBA Research

SIEMENS Healthineers

TRANSILVANIA University of Brasov

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Bambino Gesù
OSPEDALE PEDIATRICO

Queen Mary University of London

UCL

info@myhealthmydata.eu

myhealthmydata.eu

myhealthmydata

@myhealthmydata

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732907

M Y H E A L T H M Y D A T A . E U