



**Call identifier:** H2020-ICT-2016 - **Grant agreement no:** 732907  
**Topic:** ICT-18-2016 - Big data PPP: privacy-preserving big data technologies

## Deliverable 2.6

# *Privacy-by-design and compliance assessment*

Completion date: 31<sup>st</sup> October 2019

**Start of the project:** 1<sup>st</sup> November 2016  
**Ending Date:** 31<sup>st</sup> December 2019

Partner responsible for this analysis: P&A | Panetta & Associati



**List of Contributors**

<b>Name</b>	<b>Affiliation</b>
Lorenzo Cristofaro	P&A
Rocco Panetta	P&A

**List of reviewers**

<b>Name</b>	<b>Affiliation</b>
Edwin Morley Fletcher	Lynkeus
Ludovica Durst	Lynkeus
Aaron Mark Lee	QMUL
Minos Garofalakis	Athena RC
Andre Aichert	Siemens Healtineers
Aurelie Bayle	Gnubila

<b>1.</b>	<b>INTRODUCTION .....</b>	<b>5</b>
<b>2.</b>	<b>MHMD PILLARS.....</b>	<b>7</b>
<b>2.1</b>	<b>THE RATIONALE OF THIS ASSESSMENT .....</b>	<b>9</b>
<b>2.2</b>	<b>LEGISLATIVE FRAMEWORK.....</b>	<b>12</b>
<b>3.</b>	<b>DATA PROCESSING OPERATIONS IN MHMD.....</b>	<b>18</b>
<b>3.1.</b>	<b>THE ACTORS OF MHMD .....</b>	<b>20</b>
<b>3.1.1</b>	<b>PLATFORM OPERATOR.....</b>	<b>23</b>
<b>3.2.</b>	<b>DESCRIPTION OF DATA PROCESSING OPERATIONS .....</b>	<b>24</b>
<b>3.3.</b>	<b>CATEGORIES OF PERSONAL DATA INVOLVED .....</b>	<b>27</b>
<b>3.4.</b>	<b>MHMD CATALOGUE: DATA HARMONIZATION MODULE .....</b>	<b>33</b>
<b>4.</b>	<b>GOVERNANCE OF DATA FLOWS: ROLES AND RESPONSIBILITIES.....</b>	<b>38</b>
<b>4.1</b>	<b>CLINICAL DATASETS.....</b>	<b>38</b>
<b>4.1.1</b>	<b>HYPOTHESIS 1.....</b>	<b>40</b>
<b>4.1.2</b>	<b>HYPOTHESIS 2.....</b>	<b>42</b>
<b>4.1.3</b>	<b>HYPOTHESIS 3.....</b>	<b>43</b>
<b>4.2</b>	<b>INDIVIDUAL DATASETS.....</b>	<b>45</b>
<b>4.2.1</b>	<b>HYPOTHESIS 1.....</b>	<b>46</b>
<b>4.2.2</b>	<b>HYPOTHESIS 2.....</b>	<b>47</b>
<b>5.</b>	<b>WEB COMPONENTS .....</b>	<b>48</b>
<b>5.1.</b>	<b>MHMD USER INTERFACES .....</b>	<b>48</b>
<b>5.2.</b>	<b>PRIVATE WEB INTERFACE FOR HOSPITALS .....</b>	<b>53</b>
<b>5.3.</b>	<b>DATA CATALOGUE .....</b>	<b>60</b>
<b>6.</b>	<b>LEGAL FRAMEWORK TO PROTECT DATA SUBJECTS' RIGHTS.....</b>	<b>63</b>
<b>6.1</b>	<b>SEGREGATED COMPUTATION MODEL FOR CLINICAL DATA.....</b>	<b>63</b>
<b>6.1.1</b>	<b>NECESSITY AND PROPORTIONALITY IN THE SEGREGATED COMPUTATION MODEL .....</b>	<b>64</b>
<b>6.2</b>	<b>SECURE SHARING MODEL FOR CLINICAL AND INDIVIDUAL DATA.....</b>	<b>68</b>
<b>6.2.1</b>	<b>NECESSITY AND PROPORTIONALITY IN THE SECURE SHARING MODEL.....</b>	<b>70</b>
<b>7.</b>	<b>MHMD BLOCKCHAIN .....</b>	<b>78</b>
<b>7.1.</b>	<b>DESCRIPTION OF MHMD BLOCKCHAIN .....</b>	<b>79</b>
<b>7.2.</b>	<b>BLOCKCHAIN AS A SECURITY MEASURE .....</b>	<b>83</b>
<b>7.3.</b>	<b>DATA SUBJECTS' RIGHTS IN CONNECTION WITH MHMD BLOCKCHAIN .....</b>	<b>85</b>

- 7.4. PARTICIPANTS IN THE CONTEXT OF MHMD BLOCKCHAIN.....89**
- 8. MHMD SECURITY FRAMEWORK ..... 91**
- 8.1 DE-IDENTIFICATION MEASURES..... 92**
- 8.1.1 PSEUDONYMIZED AND ANONYMIZED DATA .....92
- 8.1.2 SYNTHETIC DATA .....94
- SYNTHETIC DATA.....94
- 8.1.3 SOLUTIONS IN MHMD .....96
- 8.1.4 MULTI-LAYERED PRIVACY-PRESERVING TECHNIQUES..... 102
- 8.2 INFRASTRUCTURE SECURITY ..... 104**
- A. Network and communication security ..... 105
- B. Blockchain and transactions security ..... 106
- C. Web services and applications security ..... 107
- D. Database security ..... 108
- E. Other security aspects..... 108
- 8.2.1 MHMD INFRASTRUCTURE SECURITY..... 108**
- 8.2.2 MHMD DISTRIBUTED INTRUSION DETECTION SYSTEM ..... 112**

## 1. INTRODUCTION

MyHealthMyData ('MHMD' or the 'Project') comes at the height of the eHealth era, where information technology and big data analytics have become the keys to personalized medicine and actual redesign of healthcare systems. While the rise of certain disease traits or differentiated responses to drugs have indeed emerged to be strictly dependent upon individual features, patients themselves are stepping forward to have a more active role in the clinical process, by staying informed, comparing symptoms and clinical histories, but also claiming the rights to access their own medical records and controlling their use by various stakeholders.

At the same time, the amount of biomedical data produced during clinical care, daily life and research is exploding, with the expectation to reach an amount of 2 to 40 exabyte per year in 10 years only in the field of genetic research.

As a result, personal data are threatened more than ever (27.8 to 67.7 million of medical records have been breached since 2009, according to the U.S. Department of Health and Human Services, and black-market prices for medical records are 10 times higher than other personal data).

Hospitals, as the main data gathering and storing facilities in this context, are taking on all risks and liabilities and are being exposed to threats while generally lacking the skills, experience and capital to establish appropriate defenses.

Consequently, researchers both in the public and private sector lack efficient ways to get sufficient amounts of data for their research and have to endure time consuming, expensive and often complicated procedures, which slow down the pace of new discoveries and prevent value-creation.

In this context, MHMD has been conceived as a way to protect personal data and ensure privacy, to help both hospitals and individuals to make the most out of medical data and, at the same time, making them available for scientific research lawfully and securely, while giving back to the citizens full power and control over their own data.

MHMD is developing the first open biomedical information network centered on the connection between individuals, healthcare organizations, research centres and industries, where pseudonymized clinical datasets and individual data can be shared through a blockchain-based and smart contracts-mediated transaction system for the benefit of medical care, research and innovation (also in connection with the achieving of a European Research Area).<sup>1</sup>

---

<sup>1</sup> «The Union shall have the objective of strengthening its scientific and technological bases by achieving a European research area in which researchers, scientific knowledge and technology circulate freely, and encouraging it to become more

The ultimate goal of the Project is to extract valuable and accurate information from clinical data, targeting specific similarity analysis and knowledge discovery uses cases related to precision medicine and biomedical research. Medical data residing in hospitals' repositories are used in conjunction with those coming from individual users and contribute to the overall data pool, supporting cross-domain knowledge discovery analyses.

The overall system implements trust and value-based relationships and strict protection of data subjects' identity, privacy and preferences. Strong, multitier de-identification and encryption solutions are in place to secure and dissociate data from individual identities, while private blockchain ledger and smart contracts-controlled data transactions manage consent from individual users to support direct data access requests.

Meanwhile personal data accounts ('PDA'), *i.e.* individual interfaces and clouds managed by mobile device, allow setting and managing dynamic consent according to personal preferences. In this way, patients are allowed to take control over the use of their data and are put in condition to fully leverage the value of their clinical information for personal use.

Researchers in public or private centres, on the other side, can be granted a new wealth of biomedical records available for their work. Through a dedicated data catalogue ('Catalogue') featuring high-level descriptive statistics on encrypted meta-datasets, researchers can browse and evaluate all available sources, pick the one they are most interested in, make a request and finally downloading it in de-identifiable form, or alternatively activating a segregated computation routine.

In the background, registered data are classified based on their sensitivity, informational value, while data curation and harmonization tools, encryption and de-identification technologies are applied to ensure privacy-by-design.

Advanced AI and knowledge discovery applications such as deep learning, medical annotation retrieval engines and patient-specific models for physiological prediction can now also be applied to the discovery of new drugs and devices and to the personalization of treatments. The ultimate frontier of the project is the creation of a true information marketplace governed by peer-to-peer relationships, where a constant flux of lawful data exchanges will be fueling European economy, giving a new boost to scientific research, technological advancement and clinical innovation.

---

*competitive, including in its industry, while promoting all the research activities deemed necessary by virtue of other Chapters of the Treaties. For this purpose the Union shall, throughout the Union, encourage undertakings, including small and medium-sized undertakings, research centres and universities in their research and technological development activities of high quality; it shall support their efforts to cooperate with one another, aiming, notably, at permitting researchers to cooperate freely across borders and at enabling undertakings to exploit the internal market potential to the full, in particular through the opening-up of national public contracts, the definition of common standards and the removal of legal and fiscal obstacles to that cooperation» (Art. 179 of the Treaty on the Functioning of the European Union).*

## 2. MHMD PILLARS

The Project grounds on and leverages the following main innovations:

- ✓ **Dynamic Consent:** a dynamic consent interface allows users to grant, deny and revoke data access for different uses according to their preferences through personal data accounts, *i.e.* storage clouds enabling individual access from any personal device. Dynamic consent is implemented through smart contracts running on a blockchain which makes consent management process (from its provision to the *ex-post* verification of existence of such consent) transparent, semi-automatic and tamper proof. In this way, patients are enabled to fully benefit from the value of their clinical information, turning to different healthcare professionals for second opinion, or searching for profiles of similar patients and contacting them upon their permission. Physicians, in turn, have the possibility to retrieve medical data or execute queries to identify patients with analogous features to find cues about a specific clinical case.
- ✓ **Multilevel de-identification and encryption technologies:** before the personal data are made available to the researchers, pseudonymisation or anonymisation procedures are applied dealing, according to the need on a case-by-case basis, not only with the removal of direct identifiers, but also with the possible removing of secondary information (quasi-identifiers) that might indirectly lead to backtracking an individual. MHMD-embedded tools allow to classify the datasets based on their nature and relevant informational value and, accordingly, to assesses the most suitable and robust de-identification and encryption technologies needed to secure different types of information, while still allowing advanced knowledge discovery through analytics and deep learning applications running on a growing amount of anonymised or pseudonymised data.
- ✓ **Web-based data Catalogue:** all datasets available in the MHMD infrastructure are indexed with persistent identifiers ('PIDs'). Such model is used to create non repudiable, persistent, unique and standard identifiers to selected data points. The resulting Catalogue is (i) populated by metadata, thus describing the data available in the network without revealing any identifiable information, and (ii) browsable by advanced semantic-enabled engines and interfaces, allowing to segment, group and thus create specific cohorts of data. PIDs are leveraged in transactions in lieu of the actual data and thus ensure that no sensitive data is compromised nor exposed at any time. Thence, researchers can browse datasets within MHMD network, checking what data are available and under what conditions, modality, sensitivity and privacy permissions.

- ✓ **Blockchain:** MHMD develops new mechanisms of trust and direct, value-based relationships between people, hospitals, research centres and businesses, by making use of a blockchain system, i.e. a digital ledger where information relating to the distributed storage of the health data is trimmed in hash-based language code, making it possible to describe exactly what types of data are available, referring to what cohorts of patients, and where data transactions are continuously validated to the entire network of stakeholders, avoiding any possibility of fraudulent usage.
- ✓ **Smart contracts:** self-executing contractual states in digital form regulate data transactions between MHMD users and stakeholders, granting the permission to access the data based on the consents expressed by individuals (unless their personal data are anonymized). Once they are embedded in a distributed ledger, such agreements become the only valid relationship between the parties, auto-executive and not requiring any intervention by a trusted third party.
- ✓ **Advanced big data analytics:** the Project explores the feasibility of (i) advanced data analytics for similarity search, data exploration and patient stratification, (ii) personalized physiological models for clinical decision support, (iii) machine learning algorithms for knowledge discovery and (iv) data value estimation models, on de-identified and encrypted data. This allows to respond to researchers' queries regarding specific datasets (corresponding to the requested cohorts) by implementing appropriate computation techniques (i.e. homomorphic encryption or secure multi-party computation) without any data is pulled out from the initial (hospitals or MHMD mobile App's) repositories where the first collection takes place.

Data management processes within MHMD addresses three crucial, and sometimes interrelated or competitive, goals: (i) maximize data usage and sharing, so unlocking the value of large volumes of biomedical data by allowing rapid merging of disparate, heterogeneous data sources and their lawful access by third party to support a proper privacy preserving Big Data analytical framework; (ii) assess and ensure the quality of the heterogeneous, multi-modal biomedical and personal data that feed the MHMD platform ('**Platform**'); (iii) ensure compliance with the GDPR and other applicable laws, implementing both privacy by design and privacy by default principles.

To deal with all these issues, MHMD's holistic and innovative data sharing architecture combines:

- a. a decentralized data management platform that enforces consent and peer-to-peer data transactions between healthcare stakeholders in a probative, secure and open manner, offering very strong privacy safeguards and security guarantees;



- b. a semi-automated data profiling and cleaning engine that ensures and assesses data quality, while at the same time guaranteeing the most appropriate de-identification or encryption mechanism, according to each type of data or modality;
- c. a well-designed privacy preserving and security layer that combines a multi-level anonymisation engine to support privacy preserving data publishing to external parties and segregated data mining and analytics within MHMD Platform.

## 2.1 THE RATIONALE OF THIS ASSESSMENT

According to Art. 35 of the Regulation (EU) 2016/679 ('GDPR' or the 'Regulation'), *«where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data»*.

A data protection impact assessment ('DPIA') is especially required when a processing on a large scale of special categories of data takes place.

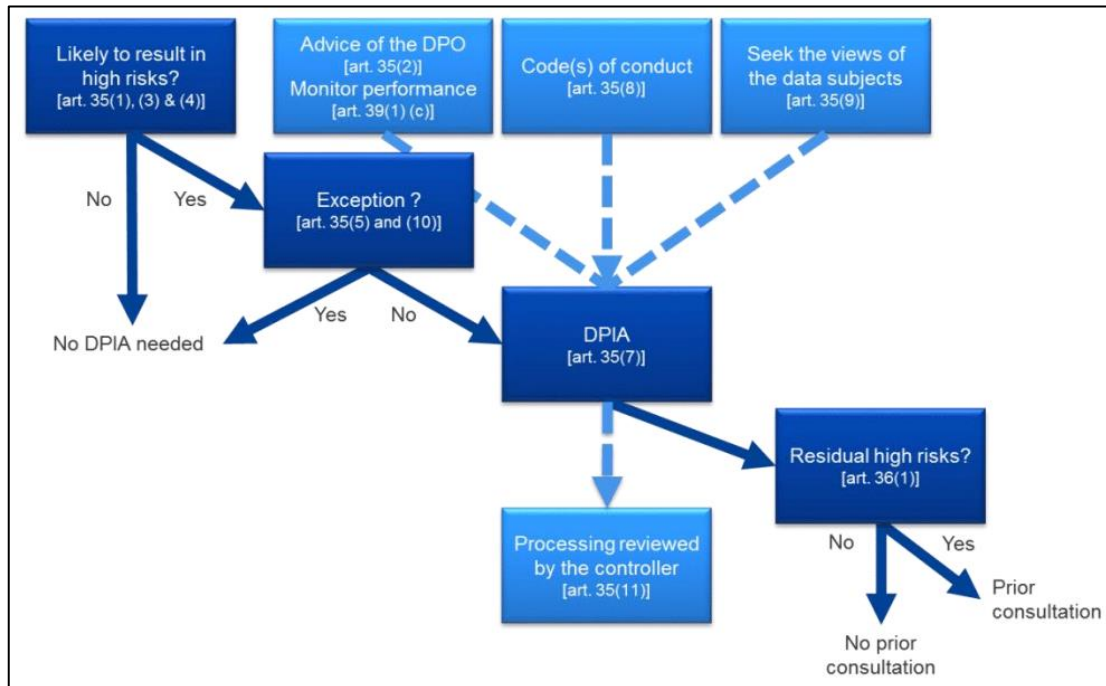
A DPIA is a tool designed to describe the processing, assess its necessity and proportionality and help manage the risks to the rights and freedoms of data subjects which may result from the envisaged operations involving personal data, in order to identify and then adopt the measures which allow the controller to best address such risks.

In other words, a DPIA is a process for building and demonstrating compliance.

The obligation for controllers to carry out a DPIA, under certain circumstances, should be understood against the background of their general obligation to appropriately manage the risks which may derive from the processing of personal data they have in place, considering that a 'risk' is a scenario describing an event and its consequences, estimated in terms of severity and likelihood.

In line with the risk-based approach underpinning by the GDPR, carrying out a DPIA is not mandatory for every processing operation: this is only required where a type of processing, on account of its nature, scope, context and purposes, is likely to result in a «high risk» to the rights and freedoms of natural persons (Art. 35.1).

The following figure illustrates the basic principles related to the DPIA in the GDPR



However, the mere fact that the conditions triggering such obligation are not met does not diminish the controllers' general obligation to implement measures to adequately address any risks for individuals' rights and freedoms.

When weighing the risks, two different elements must be taken into account (Art. 24, 25 and 32 of the GDPR):

- ✓ *severity*: meaning the significance of the risk, in terms of detrimental effects that it is capable of producing on the rights and freedoms of the individuals involved;
- ✓ *likelihood*: meaning the degree of possibility of the occurrence of one or more dreaded events.

As specified in the *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679 (WP248)*, adopted on 4 April 2017 (as last revised on 4 October 2017, the '**Guidelines**') by the Article 29 Working Party ('**WP29**'), now renamed European Data Protection Board ('**EDPB**'), a DPIA can be necessary – or useful – to assess the impact, from a data protection standpoint, of a given or novel technology, particularly where this is likely to be used by a number of data controllers to carry out different processing operations.

Of course, the data controllers deploying the product remain obliged to carry out their own DPIA with regard to the specific implementation, but this can be informed by a

preliminary assessment prepared by the product provider, if appropriate (an example could be the relationship between manufacturers of complex software and companies making use of them: each product provider should make available useful information without neither compromising secrets nor leading to security risks by disclosing vulnerabilities).

This is exactly the status of MHMD, at least in connection with a crucial part of the Project (as it will be better explained below).

In brief, the Projects' operator – see Par. 3.1.1. for more details – shall act, on the one hand, as data controller in respect of certain specific processing operations carried out through MHMD mobile APP and, on the other hand, as processor in relation to the datasets made available by the hospitals involved in the Project. However, as of today:

- a) there is not yet a legal entity created – or identified (e.g. among the partners of MHMD) to take over the management and responsibilities of the Project;
- b) due to the reason set out above, there is no national Data Protection Authority which is competent in connection to the Project, also pursuant to Art. 36 of the Regulation (EU) 2016/679 ('**GDPR**' or '**Regulation**');
- c) once implemented, the Project Operator (*i.e.* the legal entity managing MHMD web platform) will act, as said above, both as data processor on behalf of the controllers (namely hospitals and clinical centres) and as controller, depending on the source of the data entered in MHMD system. In the first case, therefore, the operator would have no obligation to carry out a DPIA under the GDPR (due to its role as processor).

Notwithstanding this, MHMD Consortium deemed opportune to conduct – in addition to the deliverables already entrusted to the legal partner, P&A – a privacy-by-design and compliance assessment, in order to evaluate if (i) all the fundamental principles of the GDPR are duly fulfilled, (ii) the risks to data subject's rights and liberties are appropriately addressed and minimized (or eliminated, where possible) and (iii) the entire range of processing operations underlying the Project is in line with applicable laws and regulations.

Moreover, this analysis can serve as a basis (and a reference) for all controllers which will have to carry out their own DPIAs before making their datasets available to third parties through MHMD Platform, and ensures that Art. 35 of the GDPR is put into practice in advance, as a precaution, for those processing activities which will be undertaken by the Project Operator – whatever legal entity should take on this role – as data controller, especially considering that in MHMD:

- new and (from the regulatory standpoint) unexplored technologies – such as blockchain and smart contracts – are leveraged to enhance security and individual control over each dataset;
- innovative schemes of data pseudonymisation and anonymization are applied to ensure data minimization and security;
- advanced techniques of distributed learning can be deployed on encrypted datasets, enabling secure and privacy-preserving computation on the data, without any personal information is pulled out from the initial (hospitals, or MHMD mobile APP) repositories where the first collection takes place;
- some data processing can be carried out in a fully automated way, leaving little or no space for human intervention and triggering legal effects for the individuals involved;
- data referred to vulnerable categories of data subjects (such as patients or children), requiring special protection, are processed;
- data of sensitive nature are collected and then elaborated on a large scale.

## 2.2 LEGISLATIVE FRAMEWORK

Even if the choice to adopt a Regulation, in lieu of another directive,<sup>2</sup> was mainly aimed at preventing the fragmentation of the rules on the circulation and the protection of personal data within the European Union – given that a regulation, according to EU treaties, does not need to and thus can't be transposed into national law (contrary to a directive, which must be implemented by local legislation) – all member States have adopted/are finalizing the adoption of their own laws to implement the GDPR or, in any case, to adapt the currently applicable national frameworks to the new European provisions.

For this reason, in its Communication to the European Parliament and the Council on 'Stronger protection, new opportunities - Commission guidance on the direct application of the General Data Protection Regulation as of 25 May 2018' (COM(2018) 43 final), the EU Commission deemed opportune to highlight that (Par. 3.1): «*when adapting their national legislation, **Member States have to take into account the fact that any national measures which would have the result of creating an obstacle to the direct applicability of the Regulation and of jeopardizing its simultaneous and uniform application in the whole of the EU are contrary to the Treaties.** Repeating the text of regulations in national law is also prohibited (...), unless such repetitions are strictly necessary for the sake of coherence and*

---

<sup>2</sup> Prior to the application of the GDPR, the EU legal framework applicable to the protection of personal data was set forth by Directive 95/46/EC of the European Parliament and of the Council.

*in order to make national laws comprehensible to those to whom they apply. Reproducing the text of the Regulation word for word in national specification law should be exceptional and justified, and cannot be used to add additional conditions or interpretations to the text of the Regulation. The interpretation of the Regulation is left to the European courts (...) and not to the Member States' legislators. The national legislator can therefore neither copy the text of the Regulation when it is not necessary in the light of the criteria provided by the case law, nor interpret it or add additional conditions to the rules directly applicable under the Regulation. If they did, operators throughout the Union would again be faced with fragmentation and would not know which rules they have to obey».*

Subject to these fundamental safeguards, member States have been granted the possibility to integrate the provisions of the GDPR, by establishing further and/or more specific rules, in a number of pre-determined areas including, *inter alia*, 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 the rights granted under the GDPR).<sup>3</sup> As of the end of May 2019, the GDPR implementation status in each member State was as follows:

MEMBER STATE	GDPR NATIONAL ADAPTATION LAWS
Austria	<a href="#">Datenschutz-Anpassungsgesetz 2018</a>
Belgium	<a href="#">Loi relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel / Wet betreffende de bescherming van natuurlijke personen met betrekking tot de verwerking van persoonsgegevens</a>
Bulgaria	<a href="#">Закон за защита на личните данни</a>
Croatia	<a href="#">Zakon o provedbi opće uredbe o zaštiti podataka</a>
Cyprus	<a href="#">Ο περί της Προστασίας των Φυσικών Προσώπων Έναντι της Επεξεργασίας των Δεδομένων Προσωπικού Χαρακτήρα και της Ελεύθερης Κυκλοφορίας των Δεδομένων αυτών Νόμος του 2018 (125(I)/2018)</a>

<sup>3</sup> Other areas of national 'delegation' include, for instance, the reconciliation of freedom of expression and data protection; employment and social security; public access to official documents; obligations of secrecy and so on.

<b>Czech Republic</b>	<a href="#">Návrh zákon o zpracování osobních údajů</a>
<b>Denmark</b>	<a href="#">Lov om supplerende bestemmelser til forordning om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger og om fri udveksling af sådanne oplysninger (databeskyttelsesloven)</a>
<b>Estonia</b>	<a href="#">Isikuandmete kaitse seadus</a>
<b>Finland</b>	<a href="#">Tietosuojalaki (1050/2018)</a>
<b>France</b>	<a href="#">LOI n°2018-493 du 20 juin 2018 relative à la protection des données personnelles</a>
<b>Germany</b>	<a href="#">Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Datenschutz-Anpassungs- und -Umsetzungsgesetz)</a>
<b>Greece</b>	<a href="#">Draft law -Νόμος για την Προστασία Δεδομένων των Προσωπικού Φακτόρα</a>
<b>Hungary</b>	<a href="#">Az információs önrendelkezési jogról és az információszabadságról szóló 2011. évi CXII. törvénynek az Európai Unió adatvédelmi reformjával összefüggő módosításáról, valamint más kapcsolódó törvények módosításáról szóló</a>
<b>Ireland</b>	<a href="#">Data Protection Act 2018</a>
<b>Italy</b>	<a href="#">D.lgs. 30 giugno 2003, n. 196 recante il "Codice in materia di protezione dei dati personali", come modificato dal D.lgs. 10 agosto 2018, n. 101</a>
<b>Latvia</b>	<a href="#">Fizisko personu datu apstrādes likums</a>
<b>Lithuania</b>	<a href="#">Lietuvos Respublikos asmens duomenų teisinė sapsaugos įstatymo</a>
<b>Luxembourg</b>	<a href="#">Loi du 1er août 2018 portant organisation de la Commission nationale pour la protection des données et mise en oeuvre du Règlement (UE) 2016/679</a> <a href="#">Loi du 1er août 2018 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel en matière pénale ainsi qu'en matière de sécurité nationale</a>

<b>Malta</b>	<a href="#">Data Protection Act, Cap. 586 (May 28, 2018)</a>
<b>The Netherlands</b>	<a href="#">Wet van 16 mei 2018, houdende regels ter uitvoering van Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (algemene verordening gegevensbescherming) (PbEU 2016, L 119) (Uitvoeringswet Algemene verordening gegevensbescherming)</a>
<b>Poland</b>	<a href="#">Ustawa z dnia 10 maja 2018 r. o ochronie danych osobowych</a>
<b>Portugal</b>	<a href="#">Draft law - Proposta de Lei nº 120/XIII</a>
<b>Romania</b>	<a href="#">LEGE nr. 190 din 18 iulie 2018 privind măsuri de punere în aplicare a Regulamentului (UE) 2016/679 al Parlamentului European și al Consiliului din 27 aprilie 2016 privind protecția persoanelor fizice în ceea ce privește prelucrarea datelor cu caracter personal și privind libera circulație a acestor date și de abrogare a Directivei 95/46/CE (Regulamentul general privind protecția datelor)</a>
<b>Slovakia</b>	<a href="#">Zákon o ochrane osobných údajov a o zmene a doplnení niektorých zákonov (18/2018)</a>
<b>Slovenia</b>	<a href="#">Draft law - Predlog Zakona o varstvu osebnih podatkov – predlog za obravnavo – nujni postopek – NOVO GRADIVO ŠT. 2</a>
<b>Spain</b>	<a href="#">Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales</a>
<b>Sweden</b>	<a href="#">Förordning (2018:219) med kompletterande bestämmelser till EU:s dataskyddsförordning</a>
<b>United Kingdom</b>	<a href="#">Data Protection Act 2018</a>

This obviously makes the task of identifying common rules governing the Project extremely more complex, because:

- a. to date, there are no best practices, case laws or binding guidance issued by competent authorities in the light of the Regulation that help understanding to what extent the openings outlined in the novel EU legislation may be lawfully levered in order to streamline and foster the development of scientific research through personal and sensitive data;
- b. the applicable obligations and derogations (if any), in particular, may deeply vary from a member State to another, so making unfeasible to implement the Project homogeneously throughout the EU and triggering some awkward operative inconsistencies between partners, stakeholders and users located in different jurisdictions.

In light of the above, this assessment will be based only on the legislation in force at the European level, so as to ensure a consistent overall approach in all member States, with particular reference to (besides the GDPR):

- ✓ Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (hereinafter '**Directive**');
- ✓ Regulation (EU) 2014/536 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC;
- ✓ the *Opinion 3/2007 on the concept of personal data* (WP136), adopted by the WP29 on 20 June 2007;
- ✓ the *Working Document on the processing of personal data relating to health in electronic health records* (WP131), adopted by the WP29 on 15 February 2007;
- ✓ the *Opinion 1/2010 on the concepts of 'controller' and 'processor'* (WP169), adopted by the WP29, on 16 February 2010;
- ✓ the *Opinion 3/2010 on the principle of accountability* (WP173), adopted by the WP29 on 13 July 2010;
- ✓ The *Opinion 15/2011 on the definition of consent* (WP187), adopted by the WP29 on 13 July 2011;
- ✓ the *Opinion 02/2013 on Apps on smart devices* (WP202) adopted by the WP29 on 27 February 2013;
- ✓ the *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*, (WP217), adopted by WP29 on 9 April 2014;



- ✓ The *Opinion 05/2014 on 'Anonymisation Techniques'* (WP216), adopted by the WP29 on 10 April 2014;
- ✓ The *Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679* (WP251), adopted by the WP29 on 6 February 2018;
- ✓ the *Guidelines on transparency under Regulation 2016/679* (WP260), adopted by the WP29 on 10 April 2018;
- ✓ the *Guidelines on consent under Regulation 2016/679* (WP259), adopted by the WP29 on 10 April 2018;
- ✓ the Guidelines mentioned above (*i.e. the Guidelines on data protection impact assessment and the criteria for establishing whether processing 'is likely to result in a high risk' pursuant to Regulation 2016/679*, adopted by the WP29 on 4 October 2017);
- ✓ the draft *Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects*, published by the EDPB on 12 April 2019 and now subject to public consultation;
- ✓ the *Data Protection Impact Assessments guidance* published by the UK Information Commissioner's Office;
- ✓ *Recommandation d'initiative concernant l'analyse d'impact relative à la protection des données et la consultation préalable*, adopted on 28 February 2018 by the Belgian Data Protection Authority;
- ✓ *Premiers éléments d'analyse de la CNIL sur la Blockchain*, adopted by the French Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés*) on September 2018;
- ✓ the report on *Blockchain and the GDPR: Solutions for a responsible use of the blockchain in the context of personal data*, published by the French *Commission Nationale de l'Informatique et des Libertés* on 6 November 2018;
- ✓ the report on *Blockchain innovation in Europe* adopted by the EU Blockchain Observatory & Forum on 7 July 2018;
- ✓ the report on *Blockchain and the GDPR* adopted by the EU Blockchain Observatory & Forum on 6 October 2018;
- ✓ the report on *Blockchain for Government and Public Services* adopted by the EU Blockchain Observatory & Forum on 7 December 2018.

### 3. DATA PROCESSING OPERATIONS IN MHMD

Final and desired output of the Project is, in brief, implementing an easy-to-use, GDPR compliant and privacy-preserving infrastructure and interface available to:

- i. research and clinical institutions seeking for greater amounts of longitudinal data to foster the development of biomedical sector (as well as to businesses for their own purposes, when permitted by law):
- ii. patients and users willing to share their personal data for scientific and medical research purposes (as well as for or other well-specified purposes, such as pharma industry commercial activities, if all conditions are satisfied to ensure the lawfulness of such processing according to the applicable laws).

To achieve this purpose, MHMD is fed by different datasets (all together, the **'Datasets'**):

- 1) patients' data that are routinely collected by hospitals and clinical centres in their own repositories (e.g. phenotype/demographic data, genomic data, medical images and signals, lab tests), already in accordance with the legal conditions and safeguards defined under MHMD (**'Routine Dataset'**). Such data are stored in a federated data storage platform where each hospital provides and controls access to its own local repository through a local MHMD driver which includes a blockchain node;
- 2) dataset collected in the past, prior to the deployment of the Project, under safeguards (mainly in terms of transparency *vis-a-vis* the data subjects and verification that consent and/or another appropriate legal ground is in place for the processing envisaged) which have not been audited in connection with this Project, to be divided in two distinct sub-categories (jointly, **'Legacy Datasets'**):
  - a. data that have already been collected by the hospitals in the context of their daily activities; and
  - b. data retrieved from previous EU-funded projects, such as MD-Paedigree ([link](#)) and Cardioproof ([link](#)), and kept in pseudonymized form.
- 3) data directly made available by individuals, *i.e.* patients and/or final users of the APP adhering to the Project spontaneously or upon request of their physician, via the digital interfaces which are being specifically designed for MHMD (**'Individual Dataset'**). MHMD aggregates personal data from disparate sources (e.g. clinical data repositories, personal drives) and data derived from commonly used wearables, or personal monitoring devices. Such data are then synchronized in a unified, user-owned account.

The requirements that must be complied with to ensure the lawfulness of the processing are very different depending on the type of Dataset concerned and may further vary on the basis of additional elements, including in particular: (i) the source of the data; (ii) whether the processing of the data in connection with MHMD amounts or not to a reuse of the specific Dataset (which is the case when the data were initially collected for another purpose); (iii) the nature and 'level of sensitivity' of the data (*i.e.* depending on whether any of the data fall or not into any special category pursuant to Art. 9 of the GDPR).

In this respect, while shaping the proper configuration of the Project's governance architecture, two separate approaches have been considered to define roles and responsibilities of all the parties involved (see Par. 4.1 and 4.2 below):

1. the first refers to the Individual Datasets (as defined above),<sup>4</sup> since they are made available to MHMD directly by the users (with the consequence that all conditions for lawfully processing their data must be satisfied directly *vis-a-vis* the data subjects by the operator of the App and the web interface);
2. the second relates to the Legacy and Routine Dataset, *i.e.* data retrieved (or that were retrieved in connection with the abovementioned EU-funded projects) by hospitals and clinical institutions during their daily activities, namely when providing health services to the patients (jointly, the 'Clinical Dataset'). In this case, the conditions for using (or sometimes re-using) the data for the purposes of the Project must have been preliminarily fulfilled by the hospitals.

As a general rule, each processing of personal data (such as «*collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction*» of data, Art. 4(2) of the Regulation) must be based on – and be carried out only after fulfillment by the data controller of – two different requirements (the '**Basic Conditions**'):

- ✓ provide the data subjects with a comprehensive notice setting out all information needed to give them a clear picture of the data processing operations intended to be carried out by means of their data. In this respect, the controller is required to take appropriate measures to provide the information in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for information specifically addressed to a child. This ensures that data are

---

<sup>4</sup> Individuals have digital datasets stored in many systems, such as social networks, wearables and clinical data repositories. They use MHMD platform to have their data integrated in a single local repository under their control, to visualize their own data in an engaging format and to participate in data sharing networks, which are of their own interest (e.g. clinical trials, primary care programs, etc.) or due to other incentives (e.g. access to private services, etc.).

collected for specified, explicit and legitimate purposes and processed fairly and in a transparent manner;

- ✓ ensure that a valid and sound legal basis exists, among those identified by the applicable law (depending on whether the data involved are of a sensitive nature or not), in relation to each processing, including the acquisition of the data subject's «*freely given, specific, informed and unambiguous*» consent (Art. 4(11) of the Regulation), when necessary.

It must be clearly pointed out that the Routine Datasets and the Legacy Datasets are put together and compared because the evaluations that will be made below with respect to the allocation of roles and responsibilities for these types of Datasets are identical. Conversely, the legal requirements (including the Basic Conditions) to be satisfied for lawfully registering on the MHMD Platform are different between the Routine Datasets and the Legacy Datasets.

In more detail, while the processing of Legacy Datasets in connection with the objectives of the Project is highly likely to amount to a 're-use' (almost by definition), this would be true in relation to Routine Datasets only for those hospitals which do not ensure that the Basic Conditions (as well as other additional requirements, if any) are duly complied with in order to carry out the activities of the Project. In brief, if hospitals (i) inform the patients, through adequate privacy notices, after the implementation of the Project, about the possible use of their data for medical and scientific research activities and (ii) acquire their specific consent for this purpose, then their processing of such data within MHMD would not amount to a re-use, because medical research would be one of the primary (and individually-permitted) purposes for which the data were initially collected by the hospitals.

The concept of 'reuse', in fact, implies that the relevant processing operations were not originally outlined to the data subject, emerging only at a later stage, so requiring a further investigation as to how ensure that the Basic Conditions are duly abided (see Art. 6.4 of the GDPR).<sup>5</sup>

### 3.1. THE ACTORS OF MHMD

The actors involved in the Project can be distinguished in five major categories:

1. the four clinical partners of MHMD<sup>6</sup> and any other hospital and clinical institution that will hopefully join the Project in the future (the '**Hospitals**');

---

<sup>5</sup> Regarding this profile, see in particular the *Opinion 03/2013 on purpose limitation* (WP 203) adopted by WP29 on the 2<sup>nd</sup> of April 2013.

<sup>6</sup> *Charité – Universitätsmedizin Berlin, Ospedale Pediatrico Bambin Gesù in Rome, Barts Heart Centre of the Queen Mary University London, Great Ormond Street Hospital for Children NHS Foundation Trust.*



- 
2. those who want to benefit from MHMD Datasets, in order to conduct scientific research, such as clinical research centres, hospitals, research groups or individual researchers (collectively '**Researchers**'), or pharma industries and other types of organisation when conducting research and development projects that serve population's needs pursuant to the applicable sectorial laws (generically '**Private Businesses**' and, together with the Researchers, the '**Stakeholders**').<sup>7</sup>



- 
3. those who have designed, set up, tested and implemented the technological, operative and legal processes underlying – and so offered by – the MHMD Platform, with the aim to enable a secure and privacy compliant exchange of personal data between healthcare facilities and Researchers and between Users (see below) and Researchers across the EU (the '**Platform Operator**', as better examined in par. 3.1.1 below);

---

<sup>7</sup> Private Businesses may be divided into two types of organizations: (i) industrial research enterprises, such as pharmaceuticals and CRO-like companies, that look for access to retrospective and prospective data of pertinent cohorts in the context of clinical studies or clinical trials, and (ii) commercial enterprises, such as Health Management Organization (HMO), Accountable Care Organizations (ACO) and health-tech companies, that seek for longitudinal retrospective and prospective data to develop primary care programs and health-tech professional solutions.



- 
4. those who spontaneously make their personal data available to third parties via MHMD web or mobile App (the '**Users**') and whose personal data are stored in freely chosen servers (such as cloud services, health apps, etc.);



- 
5. those whose personal data are collected by the hospitals within their daily routine activities (the '**Patients**') and then registered on the Platform.



### 3.1.1 PLATFORM OPERATOR

Any kind of liability that may derive from or be connected with the guarantee of compliance of the technical and legal processes underlying the Project with applicable law could not be attributed to the Hospitals that decide to embrace the Project and/or to the Researchers, given that none of them played any role in designing the architecture of MHMD (what would happen, for instance, if the protocol used by the Project to transmit the data to a Researcher did not work properly and some data were lost, or if the envisaged anonymity solutions were not considered appropriate by any Member State's Supervisory Authority?).

On the contrary, those who make use of the Platform for medical or scientific reasons (i.e. Researchers and/or Private Businesses) shall assume exclusive responsibility for carrying out only those processing operations which correspond to the purpose they have declared – and for which a specific Dataset has been made available – at the time their specific query was input on the Platform: the entire system architecture constitutes for them a mere standalone (i.e. take-it-as-it-is) service.

Accordingly, the developer of the Project must hold responsibility for any breach of the applicable law that should stem or result from the technical and operational features of MHMD, including the security measures implemented and the adequacy of the conditions that have been set to allow the processing and sharing of the data by and between Stakeholders.

Accordingly, a person or a legal entity must be identified who/which concretely and legally acts as the Platform Operator. In this respect, the most feasible options seem to be the following:

- a. one or more partners of the Consortium will become the key actors in charge of managing and assuring the proper functionality of the Platform, after its launch. Hence, such partners will replace the Consortium towards third parties – first and foremost Patients, Users and Researchers;
- b. establishing an *ad hoc* legal entity – in lieu of the Consortium – to guarantee the correct implementation of the Project. Such legal entity could be incorporated under one of the partners' legislation, considering that the practical management of the Platform (both via web or through the App) will greatly benefit from the designation of a single actor, which could be easily addressed, for any reason, both from the data subjects and the Researchers.

### 3.2. DESCRIPTION OF DATA PROCESSING OPERATIONS

Any processing of personal and health data in connection with MHMD and, in particular, the secondary use of Legacy Datasets for scientific purposes have been subject, since the very earliest stages of the Project, to an in-depth privacy-by-design evaluation aimed to ensure overall compliance of MHMD with the requirements of the GDPR and any other applicable data protection legislation.

Accordingly, each technical and operative process underlying the Project was defined – and so is envisioned – to fulfill all data protection obligations, with special reference to the principles of lawfulness, transparency, purpose limitation, minimizations, accuracy, storage limitation, integrity, security and confidentiality.

Consistently, in order to enable both Patients and Users to keep the highest level of control of their data, when either Individual or Clinical Datasets are involved, all the processing operations carried out under MHMD are based – unless the data are fully anonymous to anyone – on the data subject's consent, to foster true empowerment of the individuals in the health and scientific data environment.

In this respect, given the strengthened requirement of user-centricity, especially in relation to the processing of special categories of data pursuant to Art. 9.2 of the GDPR, an innovative dynamic consent tool has been devised, with the aim of meeting all legal requirements necessary to boost and streamline the circulation and exchange of medical data across Europe.

In more detail, it was decided to implement dynamic consent through smart contracts<sup>8</sup> and leverage the blockchain to automatically operationalise consent in the context of a blockchain architecture and to make relevant management process (from the provision of consent to the verification of its existence) transparent, semi-automatic and tamper proof, because:

- i. this approach is specifically useful when data subjects have to provide directly (*i.e.* without the intermediation of healthcare professionals) consent for third parties to access their data, as it allows patients to have a clear interface to understand the purpose of data usage and the consequences of the consent, while also selecting privacy and consent preferences in an intuitive and easy-to-understand way;
- ii. this allows Hospitals to clearly present relevant information to patients according to Art. 13 of the Regulation, in order to obtain consent and being sure that the

---

<sup>8</sup> The term 'smart contract' is referred to the incorporation into a software of self-executable contractual clauses which ensure full enforcement of the obligations agreed by the parties independently of the human intervention, so making any breach impossible.



patients have fully understood and agreed to the needed data collection and processing.

At the same time, dynamic consent will be implemented as smart contracts, thus triggering within the overall blockchain architecture the processes aimed to ensure the traceability of the given consent, the trustworthiness of the data sharing process and the operationalisation of the consent preferences, as collected by the Hospitals or directly defined by the Patients.

Thanks to the smart contracts enacting individual consents:

- Hospitals will be able to document the tamper-proof record of the consent obtained from each Patient, as to allow its easy traceability (also in case of an external auditing procedure), while automating data sharing under specific conditions, providing third parties with ready-to-use consented Datasets, without the need for contacting back the data subject, or the first data controller (*i.e.* the Hospital itself). 'Cleared' Datasets will enable easier sharing, thus laying the foundation for a proper health data sharing Platform;
- Patients will be able to activate directly their data sharing or expose to segregated computation under precise pre-defined conditions. Smart contracts will automatically execute the data exchange when the conditions defined by the data subject and then embedded in the smart contract will be met by the data access request made by the Stakeholder.

An interactive interface allows individuals to select and alter the consents in real time, while the system provides reliable storage and enforcement of these choices by cryptographically protecting sensitive personal data in a way that permits the access to such data solely for the purposes for which consent has been specifically given by the individual, tailoring consent on a wider variety of research initiatives, in a more open and flexible manner.

Data minimization standards set by Art. 89 of the GDPR have not only been implemented, but rather enhanced and reinforced by means of a 'multilayered security scheme' applied on the base of (i) an in-depth assessment of a number of intrinsic factors relevant to data sensitivity and consequent grade of risks (the use of «*pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations*»)<sup>9</sup> and (ii) whether the goal is not publishing the data on the Catalogue, but applying secure computational privacy on the output of specific queries regarding the datasets residing in the Hospitals' repositories.

---

<sup>9</sup> Recital 28 of the GDPR.

More specifically, when the intended operation consists in registering the data on the Catalogue (which makes it possible to understand what type of data have been entered into the Project's basin and to select the specific dataset cohort needed for the planned research), thus making them available to Researchers, encryption or anonymisation are applied by a specialized tool according to the need, dealing from time to time not only with the removal of direct identifiers, but also with the possible discarding of secondary information (quasi-identifiers) that might indirectly allow the reidentification of an individual.

Such security architecture implies that when a transaction – i.e. an exchange of or segregated computation on personal data – is launched and validated via smart contracts on the blockchain, only a set of anonymised information will be automatically provided to the Researcher who makes the request, if the Basic Conditions (information and consent requirements) have not been adequately abided by the Controller in relation to the selected cohort of data, thus enforcing privacy-by-design and data minimization in accordance with the GDPR.<sup>10</sup>

As Datasets are made available by the Hospitals or by individuals for the purposes of MHMD, a metadata description of the information registered on the Projects' blockchain appears safely in the Catalogue, which is freely open for browsing to all Researchers and other interested subjects (such as Private Businesses, for instance). Once the Researchers have decided to request a specific cohort of data, according to the non-re-identifiable description shown by the metadata Catalogue, they must previously register on the blockchain and get subsequently allowed to enter their request into the system.

Once these steps have been taken, the specific datasets corresponding to the cohort requested by the Researcher (or by Private Businesses, when permitted) are then dealt with in two possible alternative ways:

- i. **Applying analytics at local level:** as output of specific queries which can be responded through appropriate computation applied on the Clinical Data without any of them is pulled out from the Hospitals' repositories. This means that Researchers and Private Businesses would only see unidentifiable aggregated outputs generated on the data residing exclusively at the Hospitals' level (the '**Segregated Computation Model**'); or
- ii. **Registration on the Catalogue:** publishing the requested data on the Catalogue ([link](#)), which implies a direct exchange of data after the appropriate pseudonymization or anonymization techniques are applied. In this case, Researchers and Private Businesses receive and get access to some Datasets in the

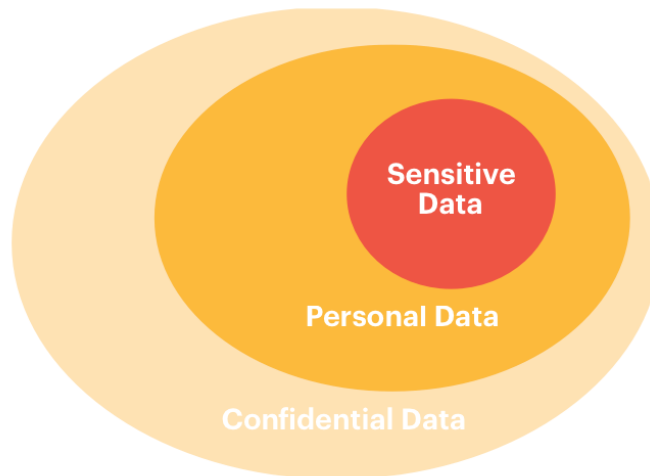
---

<sup>10</sup> WP29, *Opinion on the concept of 'controller' and 'processor'*, p. 14.

form permitted by applicable laws depending on whether – and which – specific consents have been given by Patients and Users (the ‘Secure Sharing Model’).

These two models merge together and are both made available thanks to the comprehensive MHMD Catalogue, where statistical representations can be generated and basic analytics can be run and which easily allows, in the first place, to understand what type of data have been registered into MHMD and then to select the specific Dataset cohort needed for the planned research.

### 3.3. CATEGORIES OF PERSONAL DATA INVOLVED









Due to the complexity of the Project, the categories of Patients’ and Users’ personal data involved are various, including:

<p><b>Personal data</b></p> <p><i>‘any information relating to a natural person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, ID number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person’ (Art. 4(1) of the GDPR)</i></p>	
<p><b>Identification data</b></p>	<ul style="list-style-type: none"> <li>Name and surname</li> <li>Date of birth</li> <li>Gender</li> <li>Weight and height</li> </ul>

	<ul style="list-style-type: none"> <li>Identity document</li> <li>Social security number (or equivalent)</li> </ul>
<b>Contact information</b>	<ul style="list-style-type: none"> <li>Address of residence</li> <li>Fixed telephone number</li> <li>Mobile number</li> <li>E-mail</li> </ul>

## WHAT IS PERSONAL DATA?

DEFINITION AND SCOPE UNDER THE GDPR

**ANY INFORMATION**  
Objective (earns 10k per year); Subjective (opinion); and, Sensitive data (gay woman).

**RELATING TO**  
An individual, about a particular person, impacts a specific person.

**IDENTIFIED OR IDENTIFIABLE**  
Direct or indirectly e.g. You know me by name, direct, you know me as "a Lawyer doing these graphics", indirect.

**NATURAL PERSON**  
applies ONLY to a living human being. National Law may give rules for deceased persons.

**ONLINE IDENTIFIER & LOCATION DATA**  
Include data provided by the electronic devices we use: mobiles, cookies identifiers, IP address, others.

**TO ONE OR MORE FACTORS**  
Include data that when combined with unique identifiers and other info create a profile and identify a person.

**Special categories of personal data**

*personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation (Art. 9.1 of the GDPR)*

<p style="text-align: center;"><b>Health data</b> <sup>11</sup></p> <p><i>'personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status' (Art. 4(15) of the GDPR)</i></p>	<ul style="list-style-type: none"> <li>• Blood type</li> <li>• Results of medical examinations</li> <li>• Electronic Health Record</li> <li>• Medical records and history</li> <li>• Medical images</li> <li>• Pathologies, dysfunctions and diseases</li> <li>• Clinical data</li> <li>• Wellbeing and lifestyle data<sup>12</sup></li> </ul>
<p style="text-align: center;"><b>Genetic data</b> <sup>13</sup></p> <p><i>'personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question' (art. 4(13) of the GDPR)</i></p>	<ul style="list-style-type: none"> <li>• Information regarding a specific gene or its product or function, or other parts of DNA or of a chromosome;</li> <li>• Biological samples from which genetic characteristic of an individual can be extracted:</li> <li>• Results of diagnostic, presymptomatic and predictive tests;</li> <li>• Outcomes of pharmacogenomic and pharmacogenetic tests.</li> </ul>

The processing of special categories of personal data requires, according to Art. 9 of the Regulation, which lays down a general prohibition to process this particular type of data, additional security measures and an extra-effort by the controller as to the identification of the legal basis on which to rely to ensure the lawfulness of the processing operations.

<sup>11</sup> «Personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject. This includes information about the natural person collected in the course of the registration for, or the provision of, health care services; a number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes; information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test» (Recital 35 of the GDPR).

<sup>12</sup> For instance, geolocation and physical activity data, can provide valuable indicators for the classification of medical risk profiles.

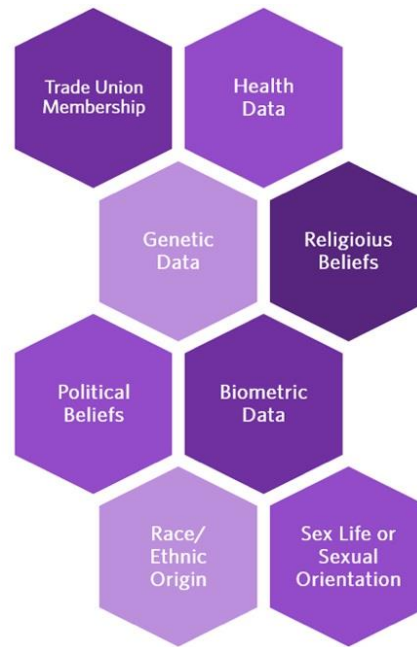
<sup>13</sup> «Genetic data should be defined as personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained» (Recital 34 of the GDPR).

More precisely, the said prohibition does not apply under certain circumstances, such as when the processing of personal data falling within any special category under the GDPR (Art. 9.2):

- i. is based on the data subject's explicit consent, except when the EU or Member State law provides otherwise;
- ii. is necessary to protect the vital interests of the data subject or of another natural person, where the data subject is physically or legally incapable of giving consent;
- iii. relates to personal data which are manifestly made public by the data subject;
- iv. is necessary for reasons of substantial public interest, on the basis of EU or Member State law, which must be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;
- v. is necessary for the purposes of preventive or occupational medicine, assessment of the working capacity of the employee, medical diagnosis, provision of health or social care or treatment or the management of health or social care systems and services, on the basis of EU or Member State law or pursuant to contract with (and under the responsibility of) a professional or any person as long as bound by the obligation of secrecy;
- vi. is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health, or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of EU or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, with particular regard to professional secrecy;
- vii. is necessary for scientific research purposes in accordance with Article 89(1) of the GDPR, based on EU or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

Nonetheless, as already said in Par. 1.2.1 above, Member States are allowed to introduce further conditions, including limitations, with regard to the processing of genetic

data or data concerning health, somewhat undermining the goal of having a single regulatory baseline at European level (Art. 9.4 of the GDPR).<sup>14</sup>

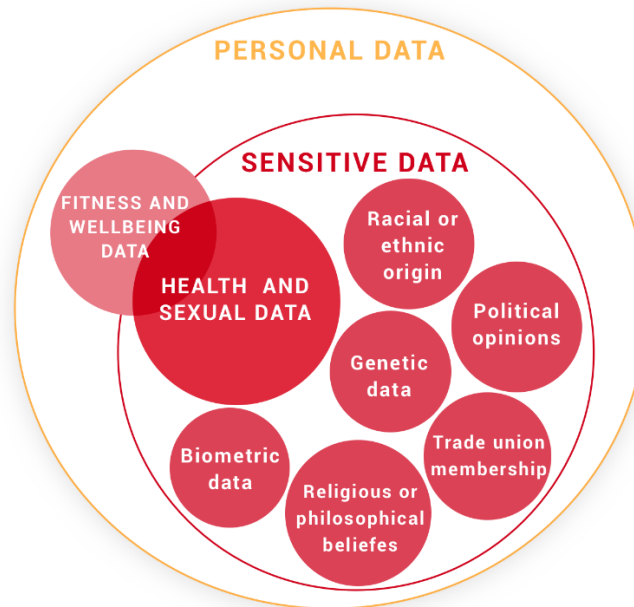


On the other side of the spectrum, there is a category of personal data generated by lifestyle Apps and devices that is, in general, not to be regarded as health data within the meaning of applicable legislation: e.g. data from which no conclusions can be reasonably drawn about the health status of a data subject. Not all raw data collected through an App qualify as information about the health of a person. For instance, if an App only counts the number of steps during a single walk, without being able to combine those data with other information from and about the same data subject, so long as the purpose of the processing is not connected with any medical context, the collected data would not be likely to have a significant impact on the privacy of the data subject.<sup>15</sup>

<sup>14</sup> 'Health data' is a much broader concept than 'medical data'. Based on the applicable laws, national legislators, judges and DPA's have concluded that information such as the fact that a woman has broken her leg, that a person is wearing glasses or contact lenses, data about a person's intellectual and emotional capacity, information about smoking and drinking habits, data on allergies disclosed to private entities (such as airlines) or to public bodies (such as schools); data on health conditions to be used in an emergency (for example information that a child taking part in a summer camp or similar event suffers from asthma); membership of an individual in a patient support group, Weight Watchers, Alcoholics Anonymous or other self-help and support groups with a health-related objective; and the mere mentioning of the fact that somebody is ill in an employment context are all data concerning the health of individual data subjects.

<sup>15</sup> See the Annex enclosed to the Opinion issued on 5 February 2015 by the WP29 in regard of 'Health data in apps and devices'.

There remain some types of processing, where it is not obvious at first sight whether or not any or all the information collected should qualify as health data. This is especially the case where the data are processed for additional purposes and/or combined with other information, or transferred to third parties.



Raw personal data can quickly change into health data when they can be used to determine the health status of a person. «To assess this, it does not suffice to look at the character of the data as is: their intended use must also be taken into account, by itself, or in combination with other information»<sup>16</sup>. For example, a single registration of a person's weight, blood pressure or pulse/heart rate, at least without any further information about age or sex, is very unlikely to allow for the inference of information about the actual or likely future health status of that individual. However, should that aspect be measured over time, particularly in combination with age and sex, then it would become suitable to reveal significant aspects of an individual's health status, such as the health risks related to obesity or an illness causing a significant loss of weight, high/low blood pressure, arrhythmia, etc.

Clearly, these types of processing operations deserve significant attention.

For this reason, all wellbeing and lifestyle data collected in connection with MHMD are processed as if they were health data, so to ensure high level security and privacy-by-design.

<sup>16</sup> *Ibidem*.

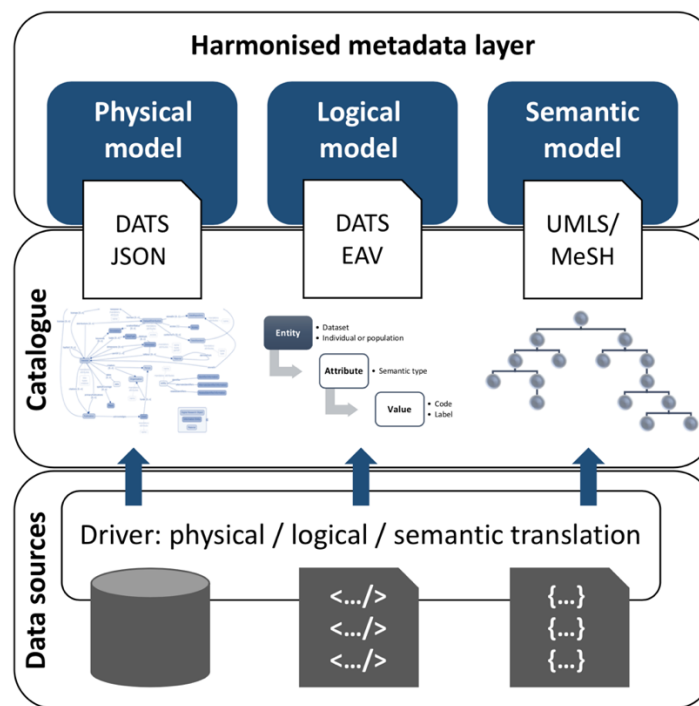


### 3.4. MHMD CATALOGUE: DATA HARMONIZATION MODULE

While the Datasets that Hospitals and Users make available to the Project – both under the Segregated Computation and the Secure Sharing Model – are very heterogeneous, Researchers and Private Businesses need streamlined and homogeneous ways to search and access the data in the MHMD Catalogue.

To address this issue, a complex work was made in order to integrate and normalize all the data coming from the various sources, by harmonizing, ingesting, cataloguing and discovering relevant metadata across the network.

Leveraging a minimal set of semantic descriptors and Dataset properties, MHMD harmonization layer is able to make the data findable, accessible, interoperable and reusable ('FAIR' principles) within the Platform.<sup>17</sup>



*MHMD metadata ingestion and cataloguing model*

Normalisation is achieved through a series of four steps, starting from the preparation and sourcing of data, necessary for the application of specialized harmonization services.

<sup>17</sup> Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. 'The FAIR Guiding Principles for scientific data management and stewardship', in *Scientific Data*, Volume 3, Article number: 160018 (2016). Abstract available [here](#).

Then, a minimum set of metadata is generated for each Dataset, to serve as a representation of the data available in the Platform. The online cataloguing of data, consisting in the publication and indexing of metadata and in the development of a cohort-search service, will provide the functionalities for finding existing Datasets and facilitate relevant consultation. As last step, the definition of flexible data sharing pipelines over data exposed through the Catalogue – where the data are organized into modalities, with the metadata granularity defined along with the blockchain infrastructure, so to accommodate possible performance limitations – will help speeding up data ‘transactions’ (*i.e.* exchanges or remote computations).

This ensures a homogeneous network of information for data mining and analytics.

Standard biomedical terminologies and ontologies – selected due to their relevance to the medical and life sciences community – were analysed in depth and, particularly, compared against four dimensions (comprehensiveness, generality, complexity and availability of the available Datasets),<sup>18</sup> to define the reference terminologies and create a coherent and comprehensive dictionary applicable to such data sources.

To allow the registration, cataloguing, search and discovery of MHMD datasets, all the existing metadata models that can be used to harmonize disparate formats and data models found in biomedical datasets were analysed, by identifying four main principles that the model should respect in order to achieve the intended objectives:

- *Generality*: the model shall be able to generalize to different types of Datasets, allowing representation of disparate data sources, such as EHR, sensors and social media;
- *Expressiveness*: the model shall allow comprehensive expression of Datasets so make them easily findable;
- *Complexity*: the model must not be complex so that they data sources can be readily integrated into the network architecture; and

---

<sup>18</sup> The first dimension, ‘comprehensiveness’, measures how complete the resource is describing its domain, *i.e.* the expressivity of the ontology language to enable representing the complexities of the domain as comprehensive as possible. The second one, ‘generality’, measures how the resource can generalize in terms of domain coverage, *i.e.* how broad is the coverage of the terminology (despite MHMD Datasets having a major focus on healthcare, they are also expected to originate from non-clinical/medical domain, thus, it is important that the resource is able to cover non-clinical/healthcare concepts, such as devices, sensors, etc). The third dimension, ‘complexity’, measures the complexity of the resource, *i.e.* how easy is for data sources to find concepts and map them into the ontology. Finally, the ‘availability’ of annotated resources dimension measures the amount of existing annotated resources using the ontology.

- *Flexibility*: the model must be flexible so that it can adapt to specific needs of Datasets, when new data sources join the Project's network.

Among the various models which were taken into account for representing biomedical data (e.g. i2b2 and Bioschemas), the DATS (Data Tag Suite) appeared as the most suitable and secure one.<sup>19</sup>

Another crucial goal which was needed to achieve in connection with the Catalogue is to ensure that the evolution of data is captured at the right level of granularity and detail, to be able to identify when such data was created, modified or deleted ('data stewardship'). This provides a basis for reproducing the Datasets as they were used at a specific moment in time, in order to verify and analyse what exact data records were retrieved and accessed by the users of the Platform, enabling a transparent account of any data exchange that takes place.

The data records utilized to investigate an ecosystem like MHMD are difficult to identify because, on the one hand, they are volatile in nature (they might change by addition, deletion or updating of records) and, on the other hand, they are composed of records from several distributed sources. At the same time, Stakeholders generally access only a specific subset of the data, namely those that best fit the needs related to their specific clinical study.

To properly address these issues, MHMD Platform relies on the implementation of the recommendations on data subset identification and citation published by the Research Data Alliance (RDA) *Data Citation Working Group*.<sup>20</sup>

The objective of the MHMD Catalogue is first to give Stakeholders a view of the data available in the Platform and then to enable them to efficiently search for the desired records based on a set of given keywords, before accessing the data, when permitted, or aggregated results, under the Segregated Computation Model.

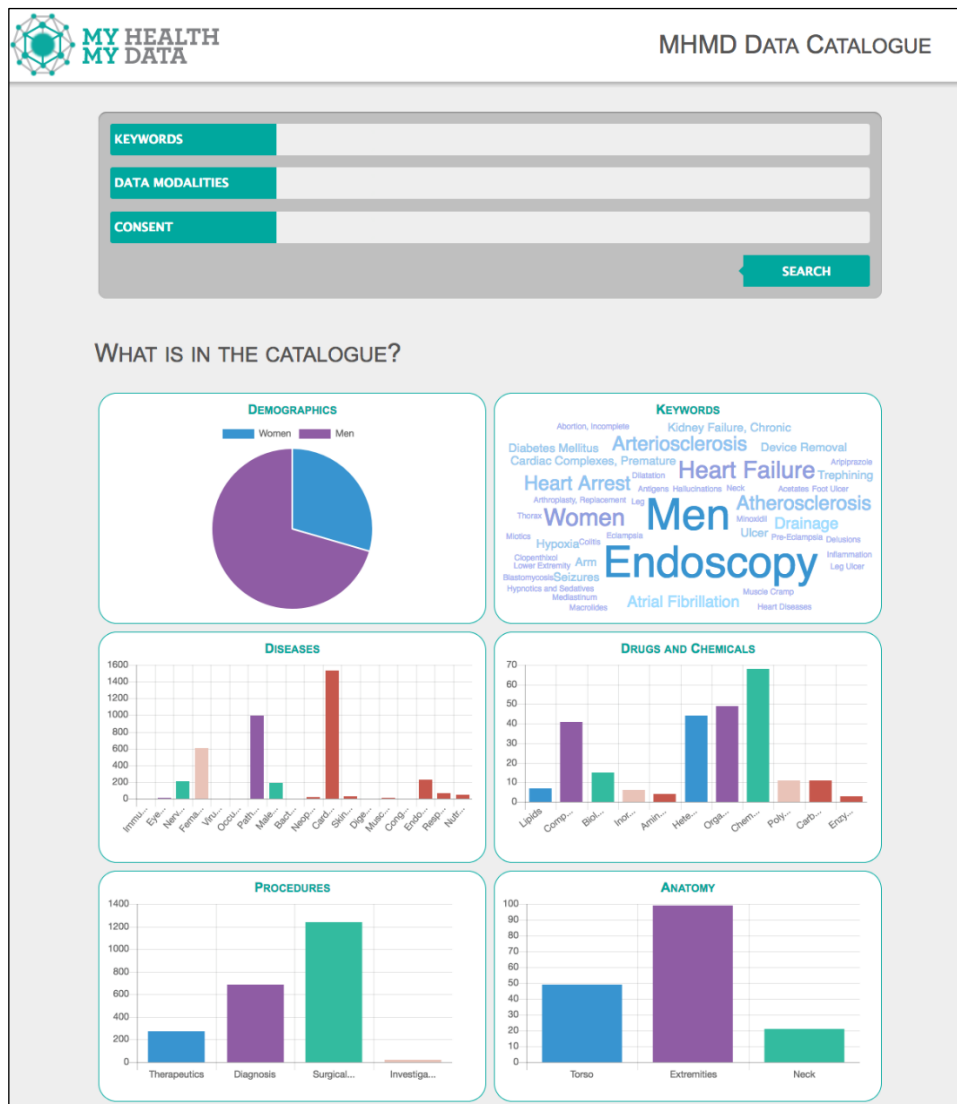
---

<sup>19</sup> DATS is the underlying model powering metadata ingestion, indexing and searches in DataMed, a NIH (National Institute of Health – US) funded project that aims to represent for biomedical datasets what PubMed ([link](#)) constitutes for the biomedical literature. Currently, DATS is used to index more than 70 repositories, including dbGaP (the database of Genotypes and Phenotypes), ClinicalTrials.gov and ClinVar and catalogues more than 2.3 million biomedical datasets. DATS is designed in a modular approach with a core model, containing the most essential metadata elements, and an extended module, with specific elements for life, environmental and biomedical science domains and can be further extended as needed. The model is able to represent more than 15 data types, including phenotype, gene expressions, imaging data and clinical trials. It has been designed with the FAIR principles for data management in mind, allowing the assignment of persistent identifiers, enrichment of formal metadata, provenance tracking and licensing.

<sup>20</sup> See [here](#) for more info.

In the first page of the Catalogue, a search box is available to explore the data available in the MHMD Platform on the basis of:

- *keywords*, to search for all records mentioning one or more terms (e.g. “heart diseases”);
- *data modalities*, to filter returned records to one or more data modalities (e.g. “prescription”);
- *consent*, to filter the records based on whether data in clear, pseudonymized or anonymized form are needed (e.g. ‘synthetic data’ – see below).



*View page of the MHMD Catalogue*

If several records are found (e.g. corresponding to different modalities or consent types), they are grouped together. For each record, a short description is displayed, as well as the consent-need type.

**RESULTS**

Page 1/66 (653 individuals / 653 records)

PID	Description	Consent
278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442	QMUL health data - crs identified	consent not required (synthetic data)
f377ba0c34fd82ad354deeb4310550d042823f8bf72422cd918aacb0	QMUL health data - crs identified	consent not required (synthetic data)
f71766949585031e89685cc021e63ae2b881d582a75e0555624becca	QMUL health data - crs identified	consent not required (synthetic data)

Finally, the Stakeholder can select and request access to, or segregated computation on, one or more Datasets of interest: the query will be anonymously entered into the blockchain system and then dispatched to the different data sources (nodes).

**DATA ACCESS REQUEST**

**QUERY**

Query: keywords:Heart Failure;  
 PID: -593102693

**SELECTED DATA**

You have requested access to the following records:

- 278e8e6d171f1d17a8773265c9fffd326641f3ce31c3a96f01138442
- 0f40f6c311bc7dbedfd919edb749f4e2efa7d9c849b7df48695861f8
- de9178147801c8f10afe173c65303754284c44a6d0dd02a8fa7cda9b
- a9c2114a3fc70e0b25ad6e8504ddccea84656e5a1f3f181c4792238
- a0b79490cd8703a6f3ac482748ab391383c4660f2459ff62b9d48679
- f9cf9804bc619972a5d35059459d2323e4d76f05acr30ffb3b297555

**ADHERENCE COMMITMENT**

Please fill the adherence commitment form and send a signed copy to sibtextmining@gmail.com

**REQUEST FINALIZATION**

Enter your email address

Click on the "Request data" button to finalize your request. Once your adherence commitment form will have been accepted, you will receive an email providing you access to the selected data.

[Back](#) **REQUEST DATA**

## 4. GOVERNANCE OF DATA FLOWS: ROLES AND RESPONSIBILITIES

In order to ensure overall compliance with data protection legislation, also in view of the accountability principle, a correct allocation of roles and responsibilities between all the parties involved in the Project is one of the main objectives to be achieved.

As different hypotheses were evaluated under the Project, pros and cons of each of them will be illustrated below, so as to better comprehend which is the most appropriate model to adopt for any type of Dataset, bearing in mind, however, that the choice regarding the allocation of roles and responsibilities can never be ‘fictitiously’ built up by the parties, but shall naturally arise from the peculiar processes and features underlying the data flows and relevant processing activities.

### 4.1 CLINICAL DATASETS

Once the Clinical Datasets have been registered on the Catalogue (under the Secure Sharing Model), so becoming accessible or remotely computable according to the requirements set forth by applicable laws, Researchers are allowed to request to receive or compute, as autonomous data controllers, specific cohorts of data.

As a consequence, from that moment on, each Researcher shall assume all liabilities that may stem from any breach of the applicable rules.

This means that any failure by the Researchers to comply with legal requirements after the Datasets have been made available, with particular reference to the communication of such data to unauthorized third parties or their processing for purposes other than those permitted based on the consents given by the data subject (in light of which the data registered on the Catalogue are ‘filtered’ via smart contracts before being made available or computable to the Researcher), may trigger the liabilities established both by the GDPR and by national legislations.

The parties involved in the processing of Clinical Datasets are the following:

- *Patients*, who provide their personal and sensitive data;
- *Hospitals*, which collect and hold the data;
- *Platform Operator*, which provides the platform and means necessary to make the data held by Hospitals available to Researchers;

- *Researchers*, who (i) access the requested datasets depending on the consents given by the Patient, or (ii) receive aggregated outputs regarding the desired cohorts of data based on the computation which is directly applied on the Hospitals' own repositories.

It is therefore necessary to define the relationships between (some of) them from a data protection compliance standpoint, as follows:

- Hospital / Platform Operator;
- Platform Operator / Researcher.

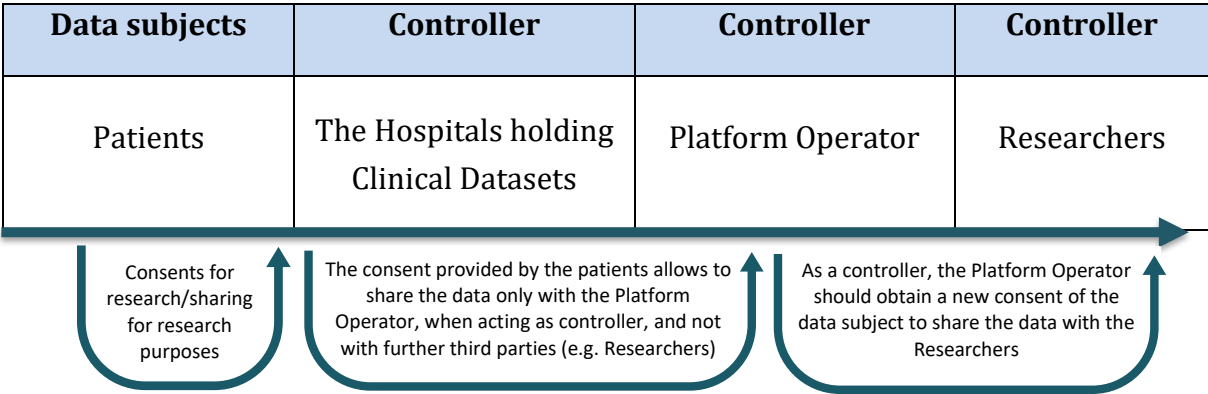
Indeed, the roles played by some of the parties identified above are somehow bound and conditioned:

- a) the Hospital which registers the data on the Catalogue and the Researcher who makes the query are each an autonomous data controller, being free to decide the purposes and the means of the processing;
- b) the Platform Operator should act as a processor on behalf of each Hospital sharing the Clinical Datasets (comprised of Legacy and Routine Datasets), for a number of reasons, such as in particular:
  - i. under a different scheme, a distinct and specific consent from the patients would be necessary to share the data with the Platform Operator (see 4.1.1 below for a more detailed explanation on this;
  - ii. the Platform Operator's intrinsic and more inherent function is that of a trusted technological service provider capable to apply both to the Legacy and Routine Dataset made available by the Hospitals all those measures which are needed to allow the Researchers to lawfully access and process such data. Accordingly, however wide the Platform Operator's room for maneuver may be in relation to making the data usable for research-related activities, it appears clear that the purposes of the processing are not independently decided by the Operator itself (which thus cannot be the data controller).

In light of this, attention must be drawn mainly on the role of the Platform Operator *vis-a-vis* the Researcher, bearing in mind that the relationship between the Hospital and the Platform Operator (controller/processor) shall not affect at all the distinct relationship

between the Researcher and the Platform Operator: the same party can – and is entitled to – play different roles towards separate counterparties and in connection with distinct processing operations.

**4.1.1 HYPOTHESIS 1**



From the Researchers’ standpoint, having the Platform Operator as a data controller would ensure stronger segregation of respective responsibilities.

Nonetheless, the Project is underpinned by the inputs and the requests coming from both the Hospitals and the Researchers, being conceived as a platform aimed to enhance secure and privacy-preserving sharing of health data. As such, MHMD needs to be:

- i. fed with Clinical Data by the Hospitals and queried by Researchers in order to access specific cohorts of data (under the *Secure Sharing Model*);
- ii. allowed to apply computation directly at the Hospitals’ level, without any Clinical Data is pulled out from their repositories, and queried by Researchers in order to access the outputs of such analytics (under the *Segregated Computation Model*).

The Platform Operator is not free, in either of the two cases, to decide the purposes of the processing (e.g. pseudonymizing a Clinical Dataset on behalf of the Hospital or helping a Researcher to find additional longitudinal data to foster a specific clinical study) and benefits from a highly marginal discretion as regards the methods to accomplish such purposes (as it mainly depends on whether the Hospital is in position or not to warrant that the Basic Conditions have been satisfied).

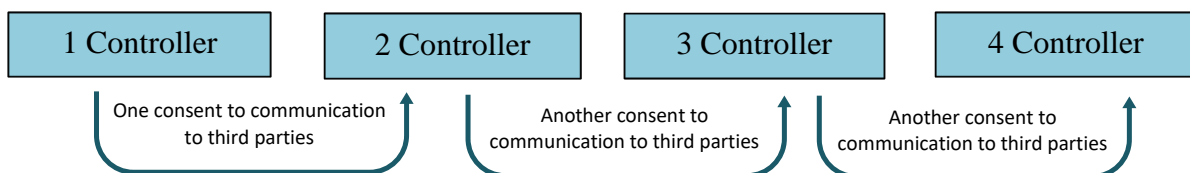


Due to this reason, the Platform Operator is not in the position to act as a data controller independently from the Hospitals and the Researchers.

In addition, another hindrance to the implementation of this model would be the necessity to identify an appropriate legal basis (namely, the Patient's consent) which legitimizes the transmission to the Researchers of the Clinical Datasets processed by the Platform Operator.

The Patient's consent eventually collected by the Hospitals (controller) allows the communication of the data, still for the purposes specified in the privacy notice, only to another controller.

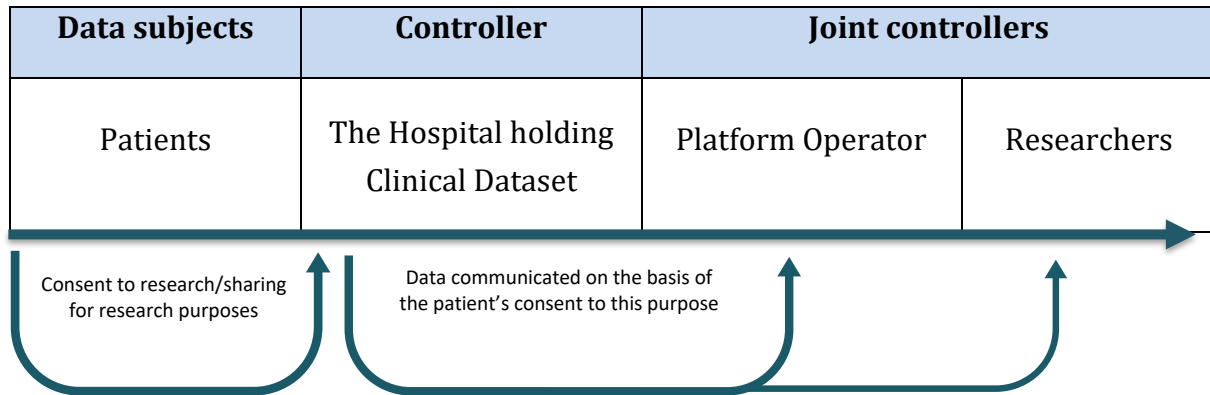
In brief: one consent for communication justifies one single communication (from a controller to another controller), as follows:



This means that if the controller (the Platform Operator, in case it should play this role) wishes, after receiving the data, to transmit them to another controller (the Researcher, in the present case), it would have to ensure that a proper legal basis is in place which ensures the lawfulness of this further communication. An indefinite chain of controllers is not acceptable, in accordance with the GDPR, on the basis of a unique consent given by the data subject.

Finally, in case the Platform Operator should act as a controller, it would have to reply and put autonomously into effect any request made by data subjects to exercise their rights under the GDPR. On the contrary, acting as a processor on behalf of the Hospital and the Researcher, respectively, all the obligations stemming from the exercise of individual rights – as set forth by Art. From 15 to 22 of the GDPR – should be fulfilled exclusively by said controllers.

4.1.2 HYPOTHESIS 2



This second model analyzes the appropriateness and, in case, the consequences that would derive from the adoption of a model of joint controllership between the Platform Operator and the Researchers (more precisely, the Platform Operator and each Researcher according to a 1:1 scheme – *i.e.* two joint controllers).<sup>21</sup>

As regards the first aspect (appropriateness), Art. 26 of the GDPR specifies that «*where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers*». In brief, in order for multiple parties to operate under this role, each of them shall be concretely in the position to exercise a significant decision-making power in respect to the objectives, the operational arrangements and the security measures of the processing.

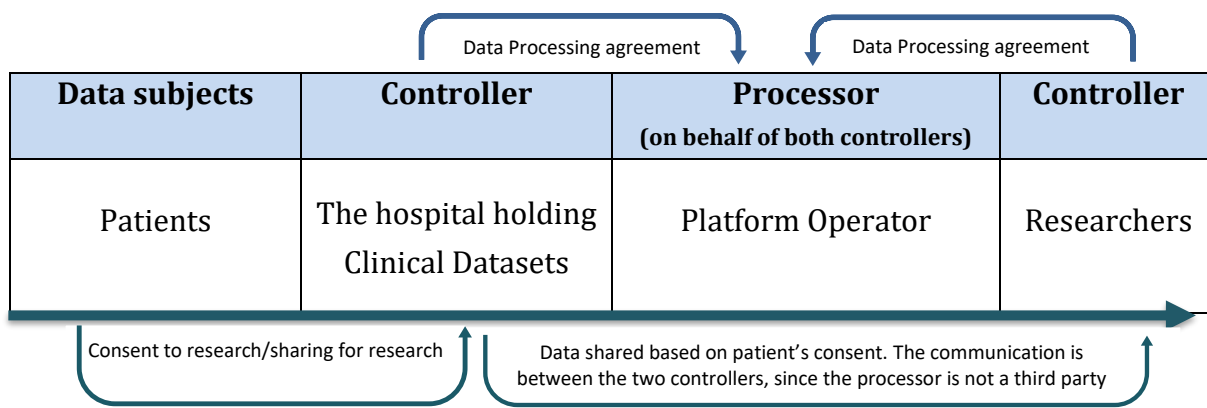
Clearly, this is not the case at issue, since the technical and legal processes underlying the Platform have been – and are being – designed and put in place without any of the Researchers (or more generally of the Stakeholders) that may benefit from the Project being involved in the set-up of its features.

Moreover, notwithstanding joint controllers are granted the possibility to determine by means of a written arrangement «*their respective responsibilities for compliance with the obligations*» of the GDPR, in particular as regards the exercising of the rights of the data

<sup>21</sup> The benefit of this mechanism is the ‘free circulation’ of data between joint-controllers, since the transmission of data between entities or individuals acting in this quality does not amount to a ‘communication’, hence not requiring that a legal basis exists to justify this processing. Conversely, the ‘weak point’ of joint controllership is that all parties involved share any liabilities arising from breaches of applicable law. However, joint controllers are entitled to determine in a transparent manner, by means of an agreement, their respective responsibilities for compliance with the obligations set forth by the legislation in force, unless such contractual regime is already determined by, or proves to be in conflict with, European or national laws.

subject, it would be reasonably unfeasible – and in any case not an appealing ‘commercial model’ for the Researchers – to accurately map and allocate the responsibilities that may derive from any failure to abide the rules applicable to the Project, especially in light of the novelty (and consequent unexplored nature, from a legal perspective) of the technological processes that underpin its operation.

**4.1.3 HYPOTHESIS 3**



As underlined by the WP29, «while determining the purpose of the processing would in any case trigger the qualification as controller, determining the means would imply control only when the determination concerns essential elements of the means».<sup>22</sup>

This means that the technical and organisational means to achieve the purposes identified by the controller (the lawfulness of such purposes shall fall under the sole liability of the controller itself) can be defined exclusively by the data processor.

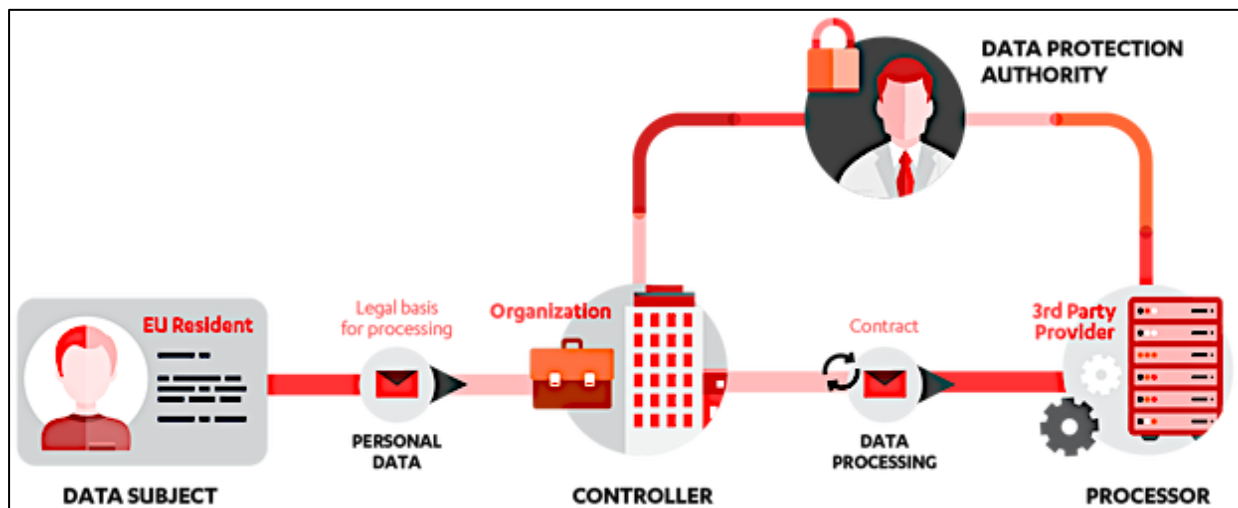
It is commonly accepted – as also indicated by the WP29 – that service providers specialized in certain peculiar processing of personal data can be in charge of setting up standard services and contracts to be signed by data controllers. More precisely, «the fact that the contract and its detailed terms of business are prepared by the service provider rather than by the controller is not in itself a sufficient basis to conclude that the service provider should be considered as a controller, in so far as the controller has freely accepted the contractual terms, thus accepting full responsibility for them».<sup>23</sup>

<sup>22</sup> WP29, *Opinion on the concept of ‘controller’ and ‘processor’*, p. 14.

<sup>23</sup> *Ibidem*, p. 26.

At the same time, the imbalance of contractual power between a (potentially) small data controller and a big digital company/operator cannot be considered a justification for the controller to accept – and the processor to impose – contractual clauses which are not in compliance with data protection law.

Under this third model, where the Platform Operator should act as data processor on behalf of the Researcher (as well as of the Hospital, by means of two separate agreements), all the relevant data flows would be based on a solid legal ground, because the transmission of data between a controller and a processor must not be considered a communication, thus not requiring the individual's consent.<sup>24</sup>



Researchers might sign up for and so access, if interested for any reason, a ready-to-use and functionally unmodifiable service allowing to readily leverage on duly verified and pseudonymised Datasets available for research. As occurs every day when consumers buy a technological product or download an App, they may use it as they prefer, to the extent permitted by the functionalities and features of such tools, but they cannot modify the

<sup>24</sup> Processors not only have additional duties under the GDPR, they also face enhanced liability for non-compliance, or for acting beyond the authority granted by the controller. Nonetheless, the major data protection obligations still rest primarily with controllers, with particular reference to the lawfulness of the instructions given to the processor: e.g. should a controller request a processor to pseudonymize a specific dataset, the duty of the service provider would be having that dataset appropriately dealt with in accordance with the state-of-the-art technology of pseudonymization or encryption. On the contrary, the adequacy of pseudonymization to achieve the intended purpose and the compliance of this technical measure with the applicable provisions shall still fall under the liability of the controller (e.g. where the pseudonymization is needed to share health data with third parties in absence of the data subject's specific consent, the violation of the applicable law – pursuant to which it would have been necessary to anonymise the data and not to encrypt them – should be attributed to the controller and not the processor that was entrusted with the application of such insufficient privacy measure).

underlying processes and, more important, would not assume any responsibility regarding compliance of these products or software with the applicable laws.

For these reasons, this architecture represents the best solution for all the parties of MHMD, both to streamline the implementation of the Project under a legal standpoint and to set a proper allocation of responsibilities between them, especially with a view to ensuring appropriate level of protection to data subject's fundamental rights and freedoms.

The Platform guarantees that all conditions relevant for publishing and making the Datasets lawfully available to third parties, under the Secure Sharing Model, are duly complied with, so that Researchers can expect to receive valid Datasets which will serve their research studies best. As an alternative, it is possible to leverage on the application of safe distributed computing capabilities on the Hospitals' repositories, so to draw reliable analytics without any need to access the Datasets.

## 4.2 INDIVIDUAL DATASETS

The Patients who wish to take part in the Project to foster the development of scientific research may make their personal data available on the MHMD Platform thus triggering the registration of relevant metadata on the Catalogue through specific web interfaces and rely on the features of the dynamic consent tool, which allows to keep control of their data flows every step of the way.

The MHMD User interfaces ("UIs") are being developed according to the most advanced standards in the field of privacy enhancing technologies, including security-by-design and strong encryption techniques, with the aim of assuring full respect of data protection principles. Nonetheless, without a correct and transparent allocation of responsibilities, the necessary levels of compliance are not achieved.

In this regard, the same considerations that led to exclude the opportunity to adopt a joint controllership model between the Platform Operator and each Researcher in relation to Clinical Datasets (see par. 4.1.2 above) shall apply, *mutatis mutandis*, also to Individual Datasets.

Please note that, for legal and technological consistency reasons, as well as to streamline the management of the Project, it is assumed that the MHMD UI will be provided and operated by the same entity which will act as Platform Operator.

#### 4.2.1 HYPOTHESIS 1



The allocation of roles illustrated above (recalling the scheme set out in par. 4.1.3 above with reference to Clinical Datasets) is much harder to be implemented – and be considered valid – in connection with Individual Datasets. There are indeed some crucial differences:

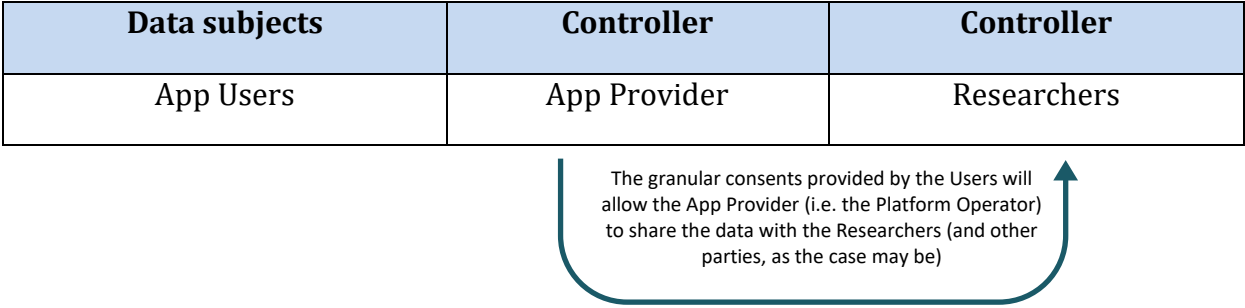
- a) while Clinical Datasets were first collected by Hospitals during their daily ordinary activities and then registered on the Catalogue (under the Secure Sharing Model) or however open for advanced analytics (in connection with the Segregated Computation Model), the Individual Datasets are directly made available by the Users through the Project's interfaces. Therefore, the Platform Operator directly engages with the data subjects;
- b) there is no Hospital involved in the processing of Individual Datasets, at least not in quality as controller which first collects such data (unlike Clinical Datasets which are by definition gathered in connection with the healthcare services rendered to the Patients).

Being the entity which freely determines (i) the purposes for which the data are collected and then processed, (ii) the conditions to be met in order for the data can be shared with the Researchers, (iii) whether the User's consent is the most adequate legal basis to rely upon, pursuant to Art. 6 and/or 9 of the GDPR, to ensure the lawfulness of the envisaged processing operations, the Platform Operator should not act as data processor also in relation to the Project's UI.

Also from the Researchers' standpoint, the option to have the provider of the interfaces operating as a data processor on their behalf is quite unfeasible (and for sure not advisable), because it would mean, in concrete, that any processing connected to such tools would substantially appear as carried out by them (albeit thanks to the support of the Platform

Operator) and, accordingly, that the relevant privacy notice should mention the Researchers themselves as the controllers competent for all the data collected from the Users, which is not true.

**4.2.2 HYPOTHESIS 2**



As it clearly emerges from the preceding paragraph, the Platform Operator should act as data controller.

According to the WP29, given that App developers design and/or create the software, they «decide the extent to which the App will access and process the different categories of personal data in the device and/or through remote computing resources. To the extent the App developer determines the purposes and the means of the processing of personal data on smart devices, he is the data controller».<sup>25</sup>

In this respect, given that the Platform Operator, as provider of the MHMD UI, is responsible of the configuration of their features and overall architecture both from a technological and legal perspective, it must be regarded as the entity in charge of compliance requirements.

This would also reflect and somehow substantiate the impossibility for both Users and Researchers to influence and change in any manner the legal and operational processes underlying the operating system, in addition to ensuring a clear and precise separation of respective responsibilities.

---

<sup>25</sup> WP29, Opinion 02/2013 on apps on smart devices (WP202), p. 9.

## 5. WEB COMPONENTS

### 5.1. MHMD USER INTERFACES

The MHMD UIs comprise three components intended for different type of 'users':

- ✓ a public web-based interface aimed to inform visitors about MHMD and encouraging the public to participate.

This UI includes access-controlled areas for:

- ✓ any individual who wants to join MHMD, by presenting the benefits of participation and to funnel the visitor to download the smartphone App;
- ✓ any Researcher who wishes to leverage the opportunities offered by MHMD.
- ✓ a smartphone application to enable individuals to manage their data and to provide access to Researchers under secure conditions.
- ✓ a private interface for the Hospitals willing to participate in the Project. Hospitals will be recruited using the public interface, but will then use a private interface running inside their own firewall. They will receive support to set up a node inside their IT department and their existing data repository structure will be mapped to the MHMD structure ready for data upload (Hospitals will thus operate as trusted nodes in MHMD blockchain network). Once this fairly manual process has been completed, username and password will be given to the Hospital's Data Protection Officer to access an internal MHMD web interface (not publicly available on the Internet) which allows to upload bulk data into MHMD. This interface also allows the DPO to curate the data that have been initially indexed, e.g. by revoking/altering permissions as appropriate from time to time based on the data subjects' consents. The process described above was already applied to all the Hospitals which are already onboard (see note 7 at page 17 above).

Following are attached the images which show how the main pages of the MHMD UIs ([link](#)) have been designed to satisfy the requirements of applicable law.

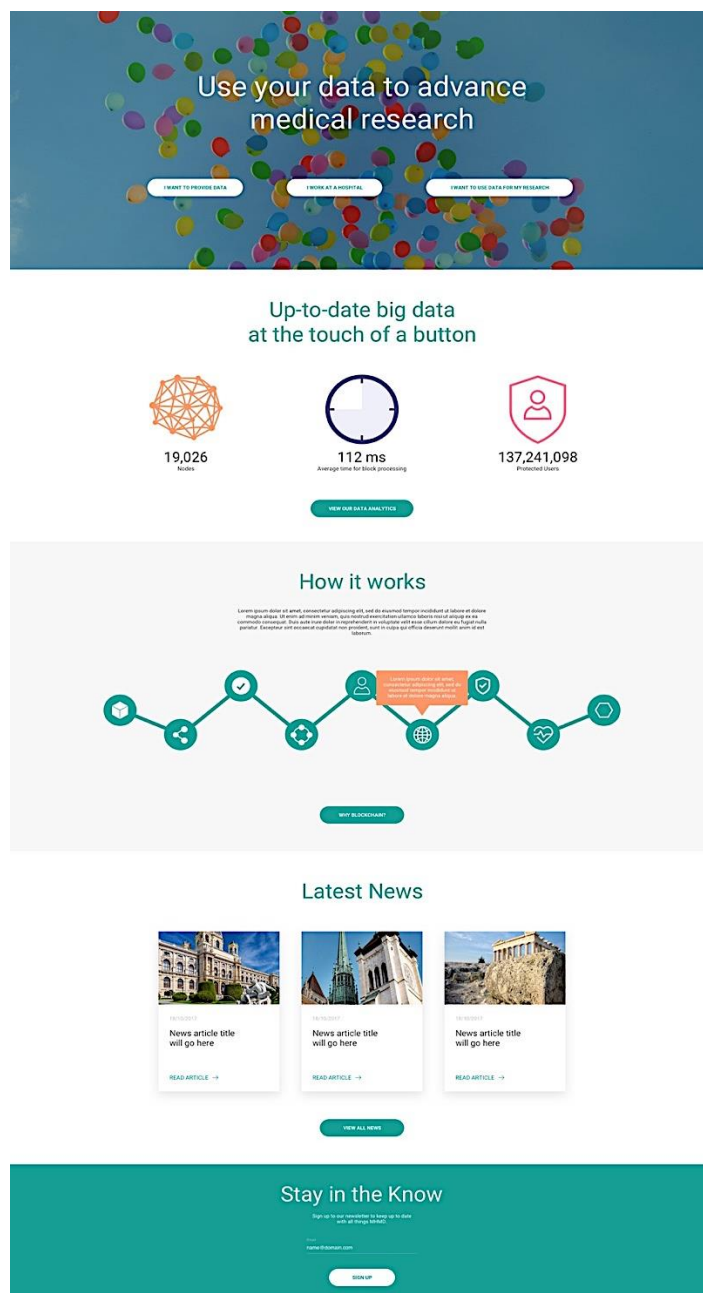
#### A. Landing page

Key elements for this page are:

- a generic call to action for any type of users;

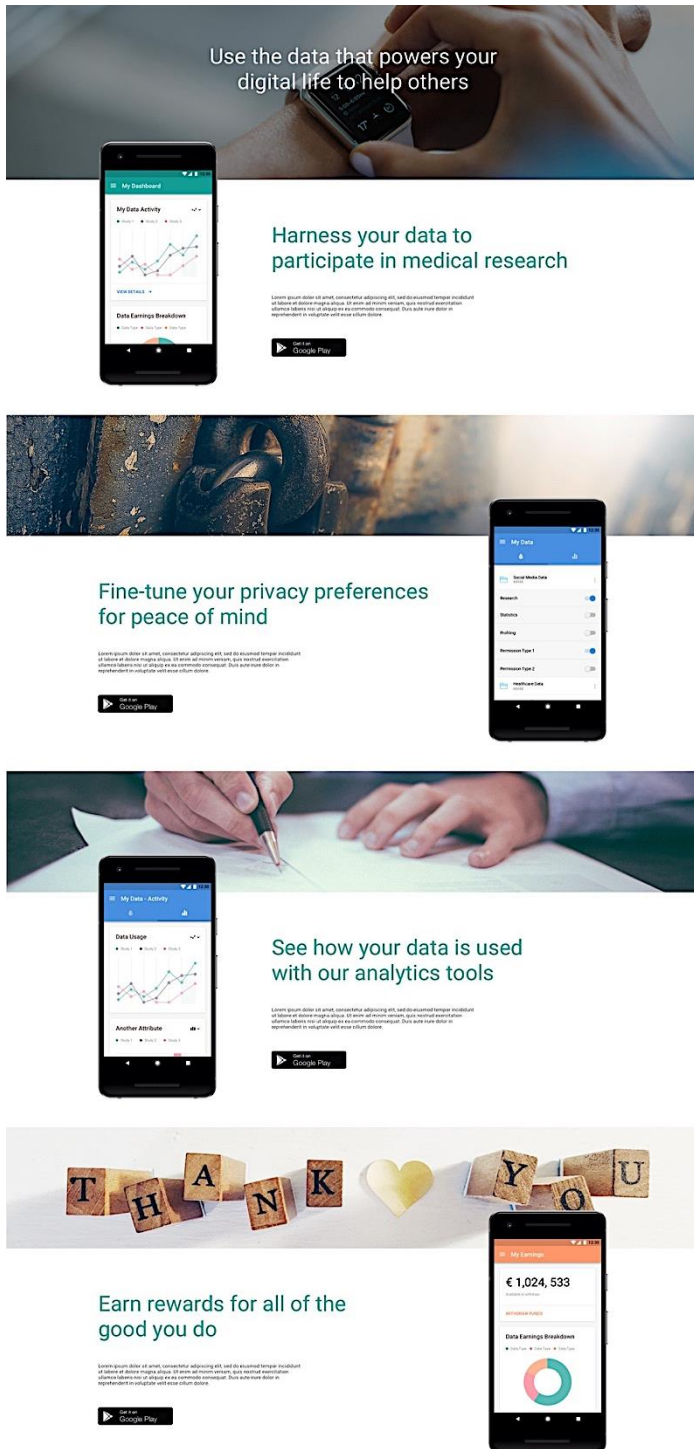


- funnelling buttons to separate each type of user so to deliver messages appropriate to their needs;
- basic aggregate blockchain statistics highlight the use of this technology in connection with MHMD. The data are accessible via an underlying API call.
- the 'How it works' section and the dynamic blockchain diagram neatly show what the blockchain is at a high level, while the use of mouse-over popups lets the visitor select to know more rather than be presented with large blocks of text.



**B. Individual Dataset page**

This page is intended to persuade a new visitor to install the App on his/her smartphone. It is organised into a banner message and 5 subsequent sections, each with a call to action to download the MHMD App.



Key elements for this page are:

- a first section setting out a ‘participate in medical research’ message and call to action to download the smartphone App. This message is targeted at those individuals who would like to participate purely on a selfless basis;
- a second section addresses key privacy concerns and reassures data subjects that their data will remain fully and ceaselessly in their control;
- a third section reaffirms the privacy protection message by highlighting that an individual can provide and revoke data access selectively and dynamically on a study-by-study basis, if they wish to;
- a fourth section informs the individuals about the benefits they may receive if they make their data available for medical research;
- a fifth and last section is an attempt to retain visitors who have not yet clicked to download the App to ‘stay in the know’.

**C. Clinical Data page**

This page is intended to allow Hospitals to participate in MHMD. It comprises four sections (each with a call to action button) and a final ‘contact us’ form:

- the first two sections highlight the key public-interest objective for Hospitals: improve medical science by providing easier data access in a secure and privacy-preserving way to a wider research base;
- two following sections focus on the organisational benefits that a hospital can acquire from MHMD, as well as on IT overheads reduction.



Find out how we can help you

0845 217 4123

First Name	Last Name
Work/Phone Number	Cell/Alt. Phone Number
Residence Address	Working Hospital
Country/Region	

Submit



**D. Page for Researchers**

This page is intended to persuade a Researcher to take a look at MHMD by browsing the Catalogue.

Key elements for this page are:

- a title message focused on what Researchers can get from MHMD (it is mentioned that data comes both from traditional sources, such as Hospitals’ repositories, and directly from patients);
- the next section encourages the researcher to browse the Catalogue;

Up to date data, powered by people and hospitals to help your research

Access a catalogue of verified and accurate data

A constant stream of fresh data from real people

Our data is easily integrated into your research workflow

Big data at a fraction of the usual cost

Browse our data catalogue now - for free!

Sign up for instant access to data

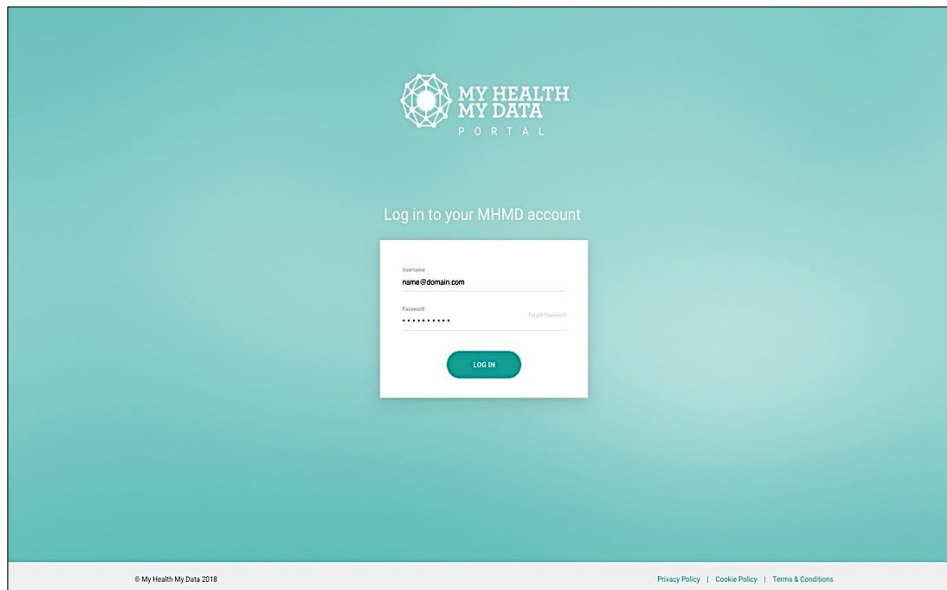
- the third section highlights that Individual Datasets are always accurate and up-to-date;
- the fourth section aims to point out that, because data are easily and quickly available, Researchers can devote all their energies on data analysis and interpretation, fostering medical development, without having to spend time to gather sufficient longitudinal data;
- the final message emphasizes the economic aspect of the value proposition: access more data for less cost.



## 5.2. PRIVATE WEB INTERFACE FOR HOSPITALS

This website ([link](#)) does not contain any marketing messages as it is specifically and exclusively addressed to the Hospitals (and thus to healthcare professionals).

### A. Login page



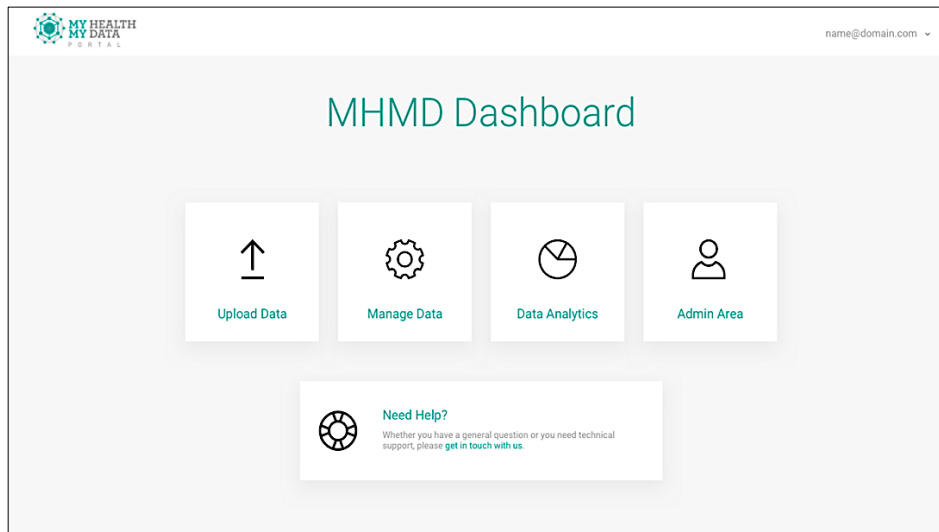
This page ensures that only duly authorized and well-identifiable clinical institutions may access to the Datasets. The user session is time limited to increase security.

Key elements for this page are:

- standard two factor authentication. It has long been debated whether to use or not 'multi factor authentication', but it was then decided to not complicate, slow down and so discourage the registration process. A user authentication facility has been embedded in the API to support this functionality.

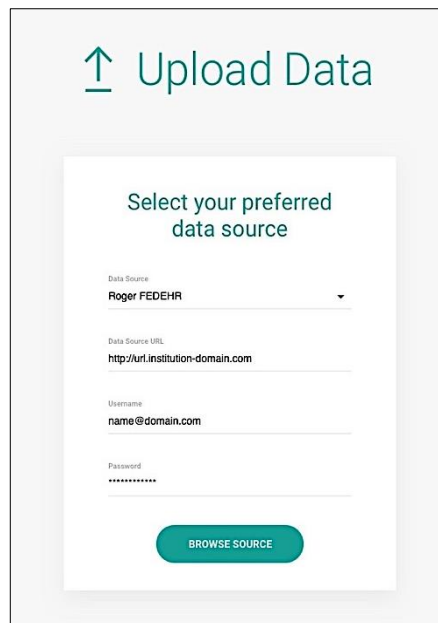
### B. Dashboard page

This page provides a dashboard overview of what type of processes can be activated on this website.



**C. Upload data – Select data source page**

This page enables the user to select the datasets to upload to MHMD for indexing from the ‘pre-configured’ and ‘pre-mapped’ Hospital’s own repository.



**D. Upload data – Data browser page**

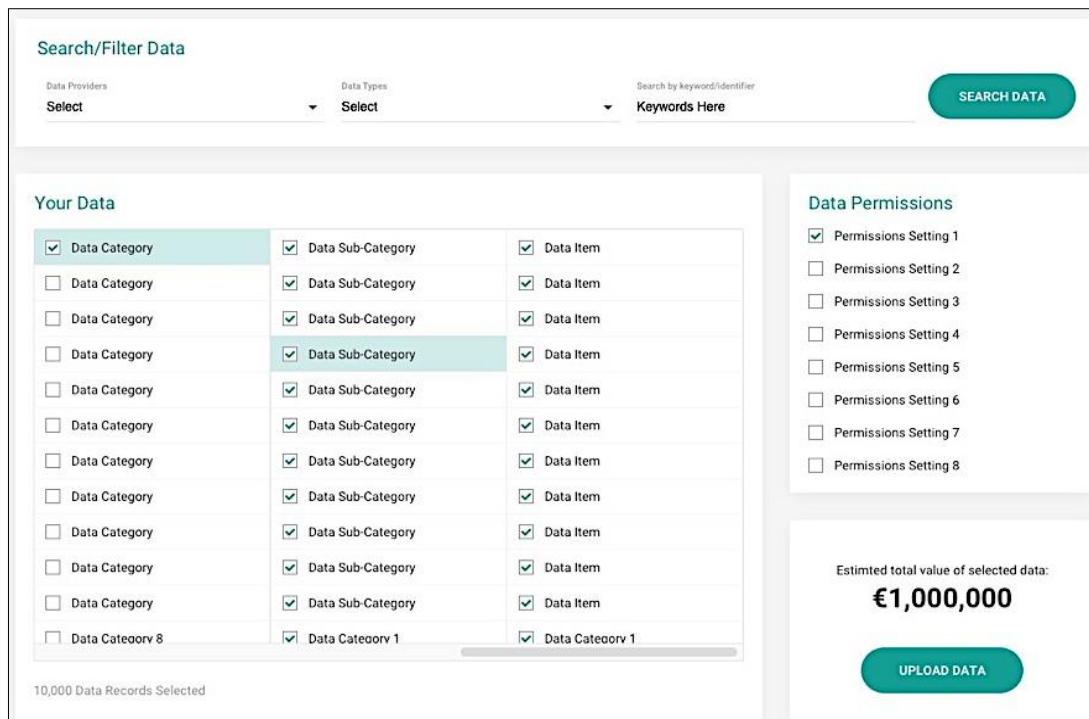
This page allows the users to set the processing permissions associated with a specific dataset.

Notwithstanding the possibility to select a single data record for upload, the focus is on adding a lot of data in one go. For this reason, users can select first a set of permissions and then the dataset which fulfil such restrictions, using an approach inspired by standard file explorers.

Permissions are associated with each data record, before uploading, based on the consent(s) provided by the data subject and, more generally, on whether the Basic Conditions have been appropriately complied with or not. In brief, the permissions set by the Hospital as data controller serve as a trigger for allowing or denying access to the Datasets by any Stakeholder, as if they were ‘close or open ‘barriers’.

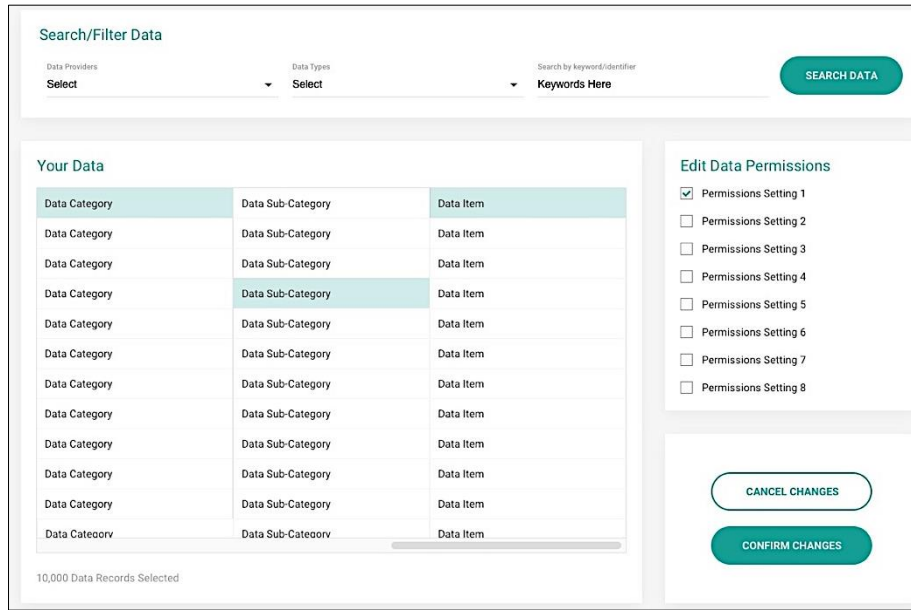
Key elements for this page are:

- *Search/Filter bar*: the user can search the data using generic or specific terms and view a large data set in the specific window (‘Your Data’);
- *Data Permissions window*: whereby the user sets the permissions associated to the data he/she intends to upload to MHMD;
- *Your Data window*: using a file browser style interface, the user can select the data he/she wants to assign the selected permissions to;
- *Value indicator*: there is a window which indicates the potential value of the data that are being uploaded and a button to send the data to MHMD for indexing.



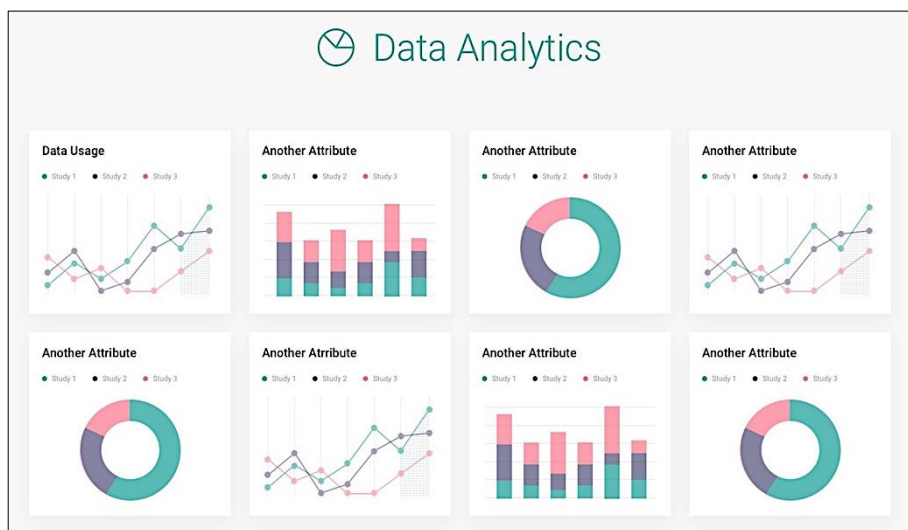
**E. Data management page**

This page allows the user to find the data and modify the associated permissions once added to MHMD. To facilitate cognitive ease, a similar browse mechanism is used for managing as for uploading.



**F. Data analytics page**

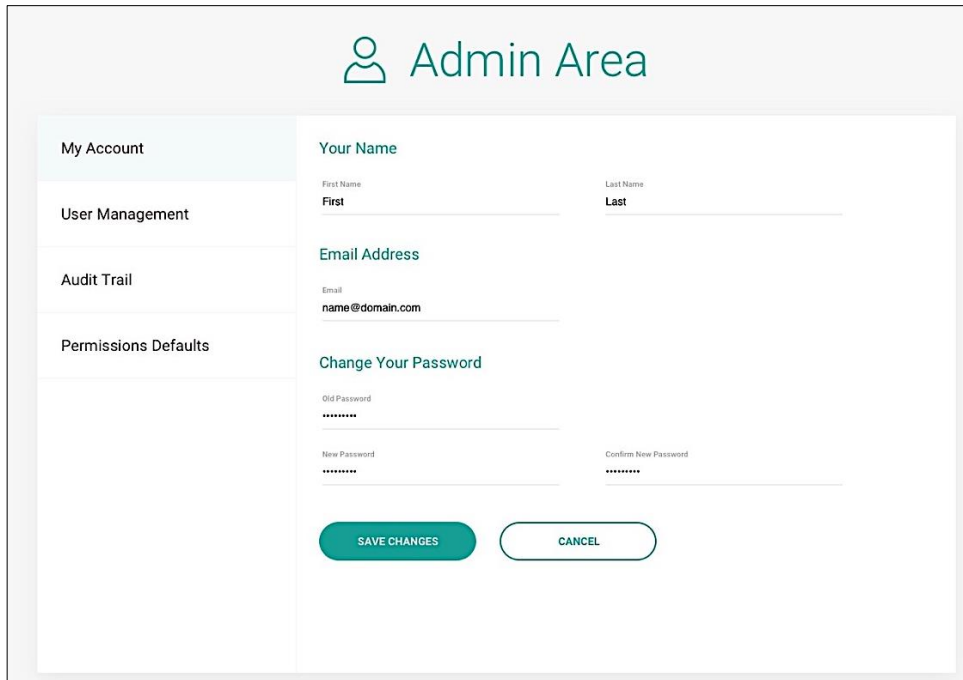
This page allows the user to know at a glance what happened to the Clinical Datasets uploaded to MHMD, thanks to a set of chart-type visualizations of aggregate blockchain data pertaining to the specific user.



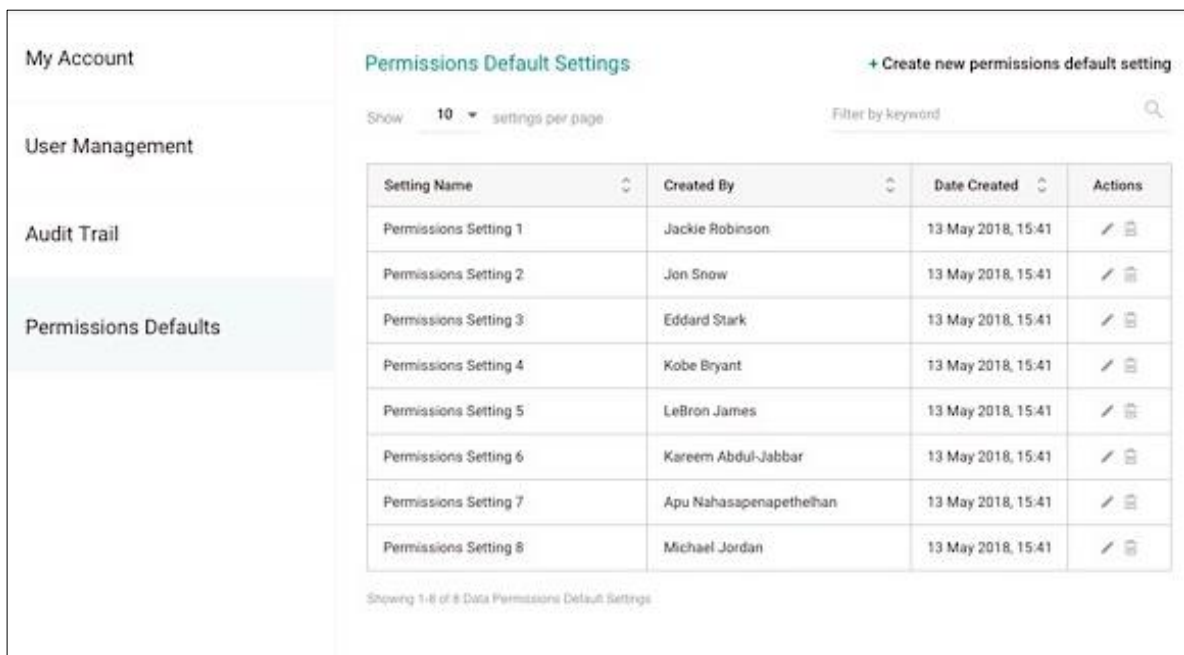


### G. Admin area

This page enables the Hospital’s system administrator and/or DPO to complete standard admin tasks such as changing the applicable password.



### H. Permissions settings



The permissions settings workflow has been the subject of the most intense focus of the UI designs, as easily selecting and assigning permissions to large datasets is a key element for ensuring that the processing of Clinical Data comply with applicable law.

After several iterations, a customisable ‘permissions settings objects’ approach proved to be the most suitable solution. Each object can represent a complex set of conditional permissions/consent-based restrictions. Users can then select from their set of permission settings objects when selecting the data to be uploaded for indexing (see also par. I below).

### I. Create permissions setting

**Create a new Permissions Default Setting**

Permissions Setting Name  
Name: **Permissions Setting 1**

**Research**

For a Specific Disease: **None** | Not For: **None**

Available Until: **12/06/2019** | Secondary Use Consent:

Specific Research Type: **Private Sector**  | **Public Sector**

**Clinical Trials**

Virtual Cohort Composition:  | Consent to Contact Patient?:

**Industrial Usage**

For a Specific Disease: **None** | Not For: **None**

For a Specific Category of Tools: **Drugs** | Available Until: **12/06/2019**

**Profiling**

Allow Profiling?:

**Statistical Analysis**

Allow Statistical Analysis?:

**SAVE DEFAULT SETTING** | **CANCEL**

This is the key page where the user can configure the permissions to enable access to the data made available to MHMD.

In brief, in their quality as data controllers, Hospital are required to map each data record collected and stored in their repository and, based on the information provided to each patient to whom the data are referred and, particularly, the consents given by the data subject, set the purposes which can lawfully be carried out by then-authorized Stakeholders.

This process is set up as a system of automatic doors – controlled via smart contracts running on the blockchain to avoid any misuse or alteration of the permissions – which allow or deny access to each data record depending on (i) the consent given by the patient and, as a consequence, (ii) the intended use declared by the Stakeholder at the time when applying for the data. For instance, should a research institution ask to receive data regarding patients aged between 40 and 50 years suffering from arrhythmia, to be used in connection with a clinical study, then only data referred to individuals who gave their specific consent for medical research purposes will be made available, because smart contracts activate the transmission only of those data which are supported by appropriate informed and specific consent.<sup>26</sup>

## J. User management

This page allows the Hospital's DPO and/or system administrator to manage other users within his/her organization.

The screenshot shows a web interface for user management. On the left is a sidebar with menu items: My Account, User Management (highlighted), Audit Trail, and Permissions Defaults. The main content area is titled 'Users' and includes a '+ Create new user' button, a 'Show 10 users per page' dropdown, and a search box. Below is a table with 10 rows of user data. The table columns are ID, Username, Full Name, User Type, and Actions. The 'User Type' column uses red boxes for 'Admin' and green boxes for 'User'. The 'Actions' column contains edit and delete icons for each user.

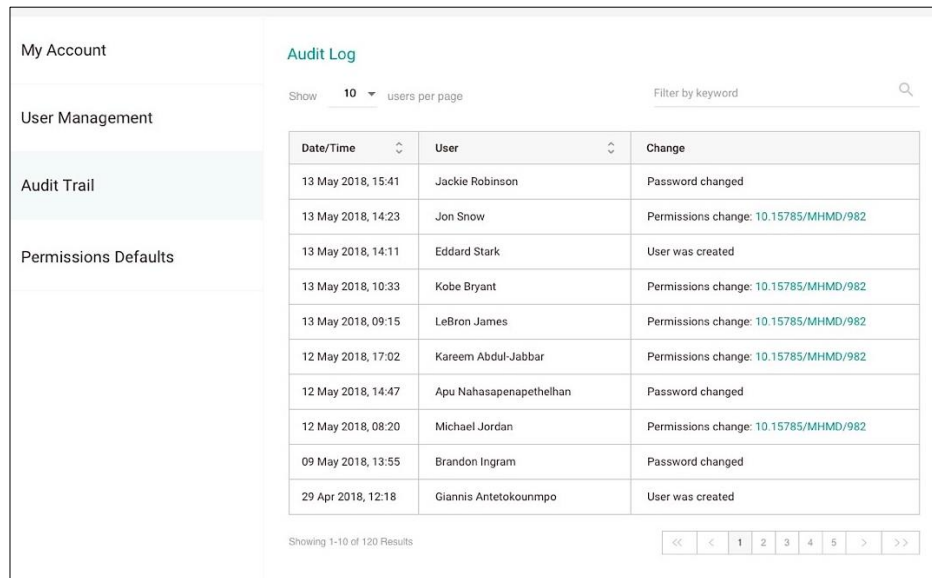
ID	Username	Full Name	User Type	Actions
42	name@domain.com	Jackie Robinson	Admin	[edit] [delete]
43	name@domain.com	Jon Snow	User	[edit] [delete]
18	name@domain.com	Eddard Stark	User	[edit] [delete]
24	name@domain.com	Kobe Bryant	Admin	[edit] [delete]
6	name@domain.com	LeBron James	User	[edit] [delete]
33	name@domain.com	Kareem Abdul-Jabbar	User	[edit] [delete]
66	name@domain.com	Apu Nahasapenapethelhan	User	[edit] [delete]
23	name@domain.com	Michael Jordan	Admin	[edit] [delete]
14	name@domain.com	Brandon Ingram	User	[edit] [delete]
34	name@domain.com	Giannis Antetokounmpo	User	[edit] [delete]

Showing 1-10 of 120 users

<sup>26</sup> The consents set out in the image above are purely illustrative.

## K. Audit trail

This page allows the user – generally the Hospitals’ DPO and/or system administrator – to see and trace a log of everything that has happened in the system under the Hospital’s credentials.



The screenshot shows a web interface with a sidebar on the left containing menu items: My Account, User Management, Audit Trail (highlighted), and Permissions Defaults. The main content area is titled 'Audit Log' and includes a 'Show 10 users per page' dropdown and a 'Filter by keyword' search bar. Below this is a table with three columns: Date/Time, User, and Change. The table contains 10 rows of log entries. At the bottom, it indicates 'Showing 1-10 of 120 Results' and a pagination control with buttons for first, previous, next, last, and search.

Date/Time	User	Change
13 May 2018, 15:41	Jackie Robinson	Password changed
13 May 2018, 14:23	Jon Snow	Permissions change: 10.15785/MHMD/982
13 May 2018, 14:11	Eddard Stark	User was created
13 May 2018, 10:33	Kobe Bryant	Permissions change: 10.15785/MHMD/982
13 May 2018, 09:15	LeBron James	Permissions change: 10.15785/MHMD/982
12 May 2018, 17:02	Kareem Abdul-Jabbar	Permissions change: 10.15785/MHMD/982
12 May 2018, 14:47	Apu Nahasapenathelhan	Password changed
12 May 2018, 08:20	Michael Jordan	Permissions change: 10.15785/MHMD/982
09 May 2018, 13:55	Brandon Ingram	Password changed
29 Apr 2018, 12:18	Giannis Antetokounmpo	User was created

## 5.3. DATA CATALOGUE

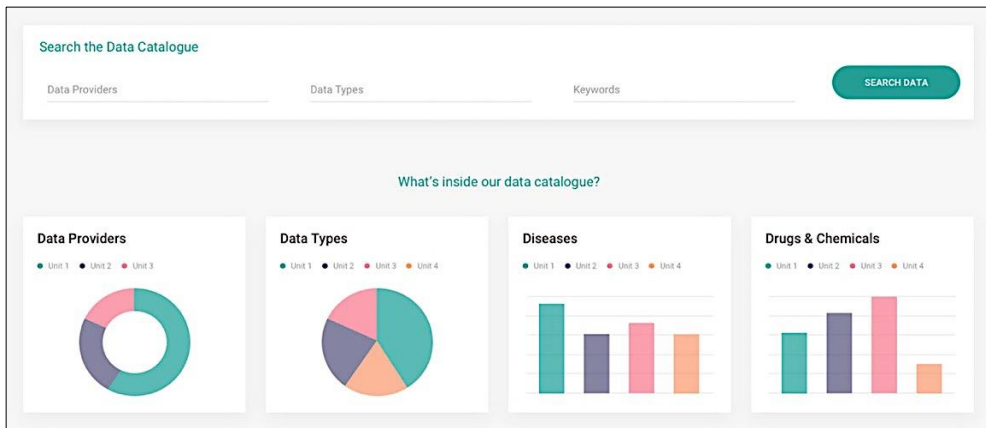
The data Catalogue ([link](#)) is included in the Web UI that is publicly accessible by Researchers (see par. 5.1 above).

### A. Landing page

This page presents the search mechanism to the user and shows some statistics about what type and how many data are indexed in the Catalogue (*i.e.* registered in the system).

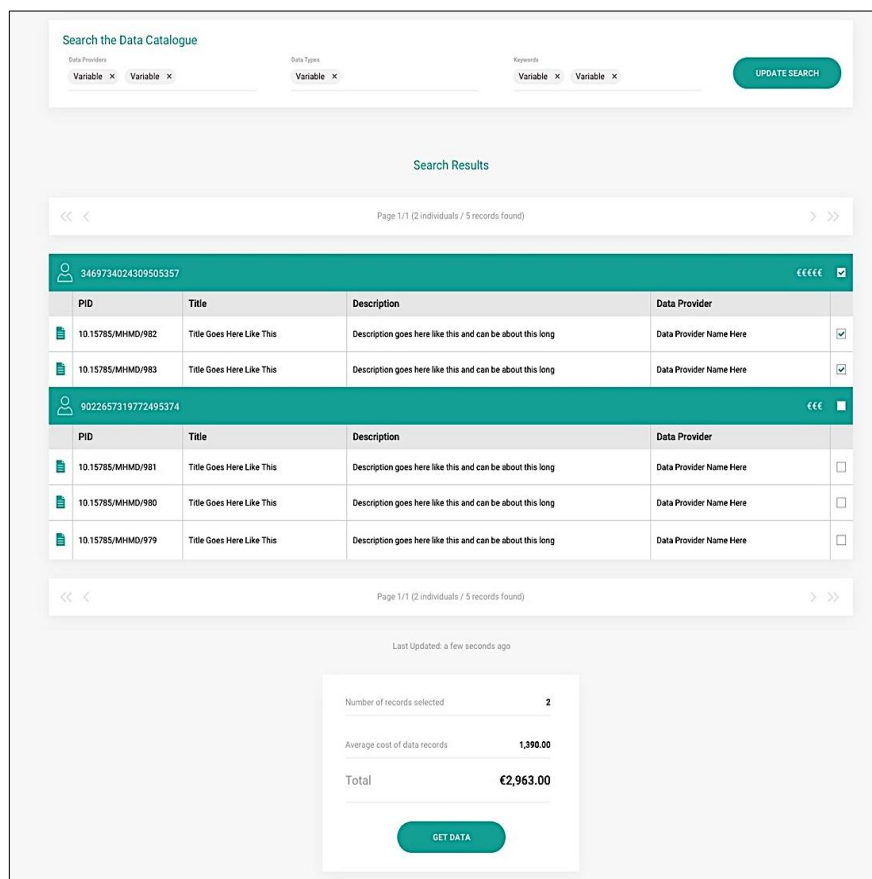
Key elements for this page are:

- three-element search interface: the search bars make use of dynamic autocomplete to assist the user;
- data analytics to represent the categories and amount of data available inside the Catalogue.



**B. Search results page**

This page provides the user with the Catalogue-search results.



Key elements for this page are:

- the already mentioned search interface, so to allow the users to amend their search, if they wish so;

- a search results table showing the type of data available and an indication of their value;
- a summary section indicating which data have been selected;
- a ‘get data’ call to action button at the bottom of the page.

### C. Login or sign-up page

If the users have not already been authenticated, they are prompted to log in at this stage. Accordingly, in case they have not signed up to MHMD yet, they are requested to create their accounts. Identity verification is a crucial element to ensure that each user can be audited and traceable. For this reason, all the data needed to ascertain without any margin of error the identity of the user acting in the name and on behalf of the Hospital must be requested (such as name, surname, date and place of birth, ID number, role within the Hospital’s organization), together with the corporate information regarding the Hospital itself. It is still under debate whether to ask or not the user to upload a scanned version of his/her personal ID document (very likely).<sup>27</sup>

### D. Payment details and user verification page

This is the section where the user is prompted to enter payment details and confirm the transaction before hitting a ‘pay now’ button.

Please complete payment for your selected data

**Payment Information**

Name visible on card\*  
Simon Harper

Company/Institution name\*  
name@domain.com

Card number\*  
XXXX-XXXX-XXXX-XXXX

Expiry date\*  
MM - YY

CVC/CVV\*  
\*\*\*

Postal Code\*  
X123BS

Country\*  
United Kingdom

Store my payment information for faster checkout next time

**PAY NOW**

Number of records selected	2
Average cost of data records	1,390.00
<b>Total</b>	<b>€2,963.00</b>

**PLEASE NOTE:**  
As this is your first purchase of data from MHMD, we will take a few extra steps once checkout has been completed in order to verify who you are. We wish to assure that you are a safe recipient of confidential data and once done, you will not have to complete this step again.  
If the verification process is unsuccessful, you will not be charged.

<sup>27</sup> It was decided not to rely on any mechanism adopted under the Regulation (EU) 910/2014 ‘on electronic identification and trust services for electronic transactions in the internal market’ (eIDAS Regulation) for a number of reasons such as (i) this Regulation has not yet been translated into practice by many member States; (ii) this kind of mechanism would complicate and slow down greatly the authentication by the users, discouraging them from joining the Project; (iii) it would make necessary to fulfill the different requirements imposed by member States’ national legislation, hindering the need of a unique approach.

## 6. LEGAL FRAMEWORK TO PROTECT DATA SUBJECTS' RIGHTS

Protecting the rights and liberties of any individual whose personal data are collected and used in connection with the Project is neither a formality which has been incidentally taken into consideration, nor a bureaucratic and burdensome (or bothersome, as many unaccountable industries would say) requirement to fulfill while designing the Platform and mapping out the operational and technical processes which govern the data flows.

Right to the contrary, the need to set out innovative measures to strengthen data security, better safeguard the rights granted by the GDPR and foster medical and scientific research sector by means of enhanced privacy-preserving processes was precisely the objective from which the idea of this Project started out.

As already described above, the Stakeholders may rely on two different solutions to nourish their research studies and more generally satisfy their needs.

### 6.1 SEGREGATED COMPUTATION MODEL FOR CLINICAL DATA

Under the Segregated Computation Model, analytics are run directly on encrypted data residing in the Hospitals' own repositories, which are part of a federated data storage platform where each Hospital has also installed its own blockchain node.

Therefore, no Clinical Data are pulled out from the controller's systems or in any manner transmitted or made available to third parties, while only metadata are registered on the Catalogue.

The Project relies on securely anonymized or encrypted Clinical Datasets for advanced data analytics and patient specific and model-based prediction applications directly within the organization of the Hospitals (data controllers) which collect and hold such data.

In more detail, specific applications have been developed and deployed to query, process and analyze the encrypted data through a well-defined secure API that implements multi-level privacy preservation techniques (including among others secure multi-party computation, differential privacy and homomorphic encryption) targeting interactive data mining and analytics.

It is worth stressing that the anonymization procedure elaborated and implemented within the Project is not limited to the removal of direct identifiers that might exist in the Datasets (e.g. name or Social Security Number), but also includes removing secondary

information (*quasi-identifiers*), such as age or zip code, that might indirectly allow to trace back the identity of an individual.

Although Researchers and Private Businesses can only receive unidentifiable aggregated outcomes coming from the big data analytics computed at local level in secure environments, each Hospital is in any case required to set the permissions for the processing of the Clinical Datasets, indicating in particular which kind of activities can be lawfully carried out based on the information provided to the patients and the consents which have been consequently given by them (if any).

This ensures data minimization and purpose limitation, as imposed by Art. 5 of the GDPR, given that the fulfilment of the Basic Conditions – including transparency (guaranteeing that the data subjects have received all needed information regarding the processing of their personal data) and lawfulness (ascertaining that a valid legal basis exists, among those established in Art. 6 and 9 of the GDPR, for each processing operation envisaged, with particular regard to medical research) – remains a key obligation for the Hospitals (as data controllers),<sup>28</sup> notwithstanding Clinical Datasets continue to be safely stored in their local repositories without any Stakeholder being allowed, when the Segregated Computation Model is applied, to access any personal data, as analytics run on encrypted data produce only statistics, graphs and aggregated unidentifiable data.

Data subjects' request to exercise any of the rights they are granted under the GDPR will – and actually can only – be addressed and properly put into effect by the Hospitals in their quality as controllers *vis-a-vis* the patients.

Accordingly, should the MHMD Platform or APP Operator receive any such request, it will be timely forwarded to the competent data controller (even if the probability that this could happen appears, given the characteristics of the processing described above, reasonably very limited).

### 6.1.1 NECESSITY AND PROPORTIONALITY IN THE SEGREGATED COMPUTATION MODEL

Because the Project is conceived to set a new benchmark for the security of health and medical data exchange for research purposes and for the definition of standardized processes to safeguard data subjects' rights (either Patients' or Users') in this field, all the



---



<sup>28</sup> For the same reason, the duty to carry out a Data Protection Impact Assessment pursuant to Art. 35 of the GDPR, before implementing the Platform (and so joining the Project), lies with each Hospital, acting as controller in respect of its Clinical Datasets.






steps have been taken, according to the principles of privacy-by-design and by-default, to ensure that the processing operations arising from MHMD comply with the requirements of applicable law.

Taking as reference the foundations of data protection legislation established by Art. 5 of the GDPR, in connection with the accountability principle, all the measures described in the preceding paragraphs have been implemented, as summarized here below:

<b>Principle</b>	<b>Description of the action</b>	<b>Risk status for individual rights</b>
<b>Transparency</b>	Hospitals which are, or ask to become, members of the Project must ensure – and hold responsibility regarding the fact – that all patients whose personal data are intended to be open for computation, albeit inside the Hospitals’ organization without being pulled out of their databases, have received a comprehensive privacy notice detailing all the necessary elements as per Art. 13 of the GDPR.	
<b>Lawfulness and fairness</b>	Although Stakeholders can receive, under the Segregated Computation Model, only aggregated statistics computed through analytics algorithms run directly inside each Hospital’s organization, both in order to substantiate individual control and due to the absence of an adequate legal basis pursuant to Art. 9 of the GDPR, controllers are still required to ensure that computation is run only on those Clinical data (including both Legacy and Routine Datasets) which are referred to patients who have given their specific consent for third parties’ medical and scientific research. Consent management process is made transparent and tamper proof thanks to smart contracts running on the Project’s blockchain. Any change	

	<p>(including withdrawal) made by the data subjects to the consent(s) they have initially given must be enacted and ‘mirrored’ in the Platform by the Hospitals by modifying the applicable specific access permissions.</p>	
<b>Purpose limitation</b>	<p>The permissions set by each Hospital based on the consents provided by the data subjects are matched, thanks to specific smart contracts, with the queries made by Stakeholders in order to enforce purpose limitation, by allowing to access the outputs of the analytics applied on the Clinical Data only to duly ‘authorized’ (indirectly, through the consent provided by the data subjects) third parties.</p> <p>In addition, before becoming members of the Project, all Hospitals are required to enter into specific ‘Platform Terms and Conditions’ by assuming <i>inter alia</i> – even if the computation applied on the Clinical Datasets can only generate unidentifiable information – the responsibility not to process the data for purposes which are incompatible with those (explicit and specific) for which they were initially collected (as declared while entering the query into the Platform).</p>	
<b>Data minimization</b>	<p>The data are stored in the Hospitals’ local repositories in pseudonymized form, in accordance with Art. 89 of the GDPR. To ensure minimization, innovative homomorphic encryption schemes have been (and still are being) developed so that advanced computation can be applied on the encrypted Clinical Data as if they were decrypted, without pulling any personal information out of the local federated local databases. For the same reason,</p>	

	Stakeholders can receive only aggregated data that do not allow to trace back any individual.	
<b>Accuracy</b>	In order to ensure that the Clinical Data are always accurate and kept up to date, the data subjects can at any moment ask – and Hospitals must take any reasonable step to ensure – that any incorrect data is erased or rectified without delay. For the reasons explained above, the Platform does not need to update any data, as no personal information ever comes out the Hospital’s own database.	
<b>Storage Limitations</b>	Data retention responsibilities lie exclusively with the Hospital, in their quality as data controllers, since the outputs of the analytics applied by the Platform result in statistics and highly-aggregated information which do not allow any Stakeholder or third party – taking account of all the means that are reasonably likely to be used to this purpose – to re-identify any individual.	
<b>Integrity and confidentiality</b>	As already specified in the preceding paragraphs – and as it will be better detailed below – a number of security measures have been implemented, in the light of the technological state of the art, and many others have been (and still are being) specifically developed from scratch, to guarantee the seamless integrity and confidentiality of the data collected and processed in connection with MHMD. Among such measures, with reference to the Segregated Computation Model, it is worth mentioning particularly the following, due to their innovative nature: (i) all personal data undergo ‘multi-level’ encryption schemes (ii) analytics are run on encrypted datasets held by the Hospitals, without having to decrypting them, thanks to the cutting-edge solutions developed within the Project in	

	<p>relation to Homomorphic Encryption and Secure Multiparty Computation; (iii) the consents given by the data subject, translated into usage permissions set through the dedicated Platform interface by the Hospitals, are securely enforced by means of specific smart contracts; (iv) encrypted metadata regarding each query made by Stakeholders under the Segregated Computation Model is registered and stored on the MHMD blockchain, in order to keep trace of any 'transaction' activated through the Platform.</p>	
--	---	--

**6.2 SECURE SHARING MODEL FOR CLINICAL AND INDIVIDUAL DATA**

The MHMD harmonized metadata Catalogue allows Stakeholders to browse and appraise the existing Datasets by performing descriptive statistics on the underlying sources, identified by PIDs (persistent identifiers) along with *Uniform Resource Identifier* (URI) and basic attributes describing each Dataset, such as creation time, provenance, sensitivity, type, semantics and version. The queries allow to search data by modality, standardized keywords and, especially, verification of existing consent-based access restrictions (see Par. 3.4 above).

Any request of access to a cohort of data transits through the blockchain, where the relevant query is distributed to all ledger nodes.

Under this Secure Sharing Model, differently from the Segregated Computation Model, Stakeholders do not receive unidentifiable outputs deriving from the analytics applied on the Clinical Datasets stored locally by the Hospitals, but are allowed to get material access to the Datasets, so long as the usage permissions set by the Hospitals and enforced via smart contracts – depending on whether at least the Basic Conditions are duly satisfied – duly match with the purpose-based requests made by the Stakeholder when entering the data query into the Platform.

To ensure security of this process and compliance with the requirements laid down by data protection law, with particular regard to the integrity and the enforcement of individual rights, various levels of de-identification are applied to overcome the failures which may be caused by lack of transparency towards the data subjects (Patients or Users), or by the

absence of suitable legal bases for processing their data (*i.e.* permitting to share the Datasets even when no – or not all the – requirements established by the GDPR are properly met).

Given that Researchers and Private Businesses that seek to get access to greater amounts of longitudinal data to foster clinical studies shall act as autonomous controllers pursuant to Art. 4(7) and 24 of the GDPR, the Datasets can be published on the Catalogue – so being made available to Stakeholders – under the following stringent conditions:

a) with reference to Clinical Datasets:

In pseudonymized form		In anonymized form
The information notice given to the patients by the Hospital clearly identifies the purpose of sharing their data with external researchers for a specific clinical study, or for certain areas of scientific research <sup>29</sup>	The patients gave (for Legacy Dataset) or give (for Routine Dataset) their consent for a specific research project, or to certain areas of scientific research	No information notice was/is given and no consent was/is acquired as described in the green left column, or when such information notice or consent do not fulfill the requirements of the GDPR

b) With reference to Individual Datasets:

In pseudonymized form		In anonymized form
The information notice provided to the Users by the APP Operator clearly identifies the purpose of making their data available to third parties for a specific research, or for certain areas of scientific research	The Users have given their consent for a specific research project, or to certain areas of scientific research	No information notice is given and no consent is acquired as described in the green left column, or when such information notice or

---



<sup>29</sup> Recital 33 of the GDPR states that “It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose”.

		consent do not fulfill the requirements of the GDPR <sup>30</sup>
--	--	---


### 6.2.1 NECESSITY AND PROPORTIONALITY IN THE SECURE SHARING MODEL

Accountability was taken in the utmost consideration, putting in place the most suitable technical and organizational measures to guarantee full protection of the rights vested in the data subjects involved, as shown here below:

- a) with reference to Clinical Data:


Principle	Description of the action	Risk status for individual rights
<b>Transparency</b>	Hospitals which are, or ask to become, members of the Project must ensure and be responsible that all patients whose personal data, including special categories of data pursuant to Art. 9 of the GDPR, are shared with third parties ( <i>i.e.</i> the Stakeholders) through the Platform, have received a comprehensive privacy notice detailing all the necessary elements as per Art. 13 of the GDPR.	
<b>Lawfulness and fairness</b>	Hospitals are required to declare and warrant, through the Platform settings ( <i>i.e.</i> selecting appropriate access and usage permissions), that the Clinical Dataset made available to the Project are exclusively referred to patients who have given their specific consent to share their data with third parties in relation to a specific clinical study, or for certain areas of scientific research. Such data are always pseudonymized, to ensure security and data minimization according to Art.	

<sup>30</sup> This option may be substantially excluded, given that all steps are being taken to ensure full compliance of the novel MHMD App with the applicable legislation.



	<p>89 of the GDPR and, when there is no valid or sufficient legal basis for sharing them, or when the patients did not receive clear information about this processing, they are made accessible only after adequate anonymization is applied – adding various level of encryption – to avoid any singling-out. The consents provided by the Patients are enacted in the Platform thanks to dedicated smart contracts which automatically deny the access to the data to those Stakeholders whose query does not fulfill the usage requirements set by the Hospitals.</p>	
<b>Purpose limitation</b>	<p>The permissions set by each Hospital based on the consents provided by the data subjects are matched, thanks to specific smart contracts, with the usage purposes identified by the Stakeholders when entering queries into the Platform. In brief, the data access ‘door’ will be open only to those who declare, under their own responsibility, that they want to use the Clinical Data for one (or more) of the purposes that were originally and expressly consented by the patients.</p> <p>In addition, before becoming members of the Project, all Hospitals are required to enter into specific ‘Platform Terms and Conditions’ by assuming, <i>inter alia</i>, the responsibility of not processing the data for purposes which are incompatible with those (explicit and specific) for which they were initially collected (as declared by each Stakeholder when making a data query).</p>	
	<p>The MHMD Platform is designed to apply strong encryption algorithms and end-to-end encryption by design and to ensure that, whenever the purposes described by the Stakeholders querying the Platform can be fulfilled by a “<i>processing which does not permit, or</i></p>	



<b>Data minimization</b>	<p><i>no longer permits, the identification of the data subjects” (Art. 89.2 of the GDPR), only anonymized Clinical Datasets are made available to achieve such purposes. To strengthen security, all Clinical Datasets are pseudonymized by default, to prevent direct identification of individuals except by means of further separate information. A number of additional security measures have been put in place, as already outlined above (and better described below), also in connection with distributed ledger mechanisms to avoid misinterpretation, abuse, fraudulent usage, unauthorized access to the data and similar circumstances.</i></p>	●
<b>Accuracy</b>	<p>In order to ensure that the Clinical Data are always accurate and kept up-to-date, the data subjects can at any moment ask – and Hospitals must take any reasonable step to ensure – that any incorrect data is erased or rectified without undue delay. Similarly, it remains up to the Hospitals to enter the correct data into the Platform, once they have been modified.</p>	●
<b>Storage Limitations</b>	<p>Although the ‘storage limitation’ principle stipulates that the data must be kept in a form which permits identification of data subjects for no longer than is necessary to achieve the purposes for which such data have been collected, Art. 5.1(e) of the GDPR establishes that personal data may be stored for longer periods insofar as they are processed solely for, <i>inter alia</i>, scientific research purposes in accordance with Article 89(1), to safeguard the rights and freedoms of the data subjects. On account of this and considering that a number of innovative security measures are applied in addition to those laid down by said Art. 89 (mainly consisting in the pseudonymization of data), Clinical</p>	●






	<p>Datasets will be stored by the MHMD Platform, at least under appropriate pseudonymization, until Hospitals change the data access/usage limitations through the dedicated settings interface (<i>i.e.</i> until specific instructions are provided by the controllers).</p>	
<b>Integrity and confidentiality</b>	<p>Many state-of-the-art security measures have been implemented and many others have been developed – or taken to a higher technological level – specifically for the Project, with a view to guaranteeing seamless integrity and confidentiality of the data collected. Among such measures: (i) ‘multi-level’ encryption schemes will be applied to the data by an innovative tool able to assess a number of intrinsic factors relevant to data sensitivity and consequent grade of risks, then automatically selecting the de-identification technique which best fit the purpose to secure the data; (ii) the consents given by the data subject, translated into access permissions set by the Hospitals through the dedicated Platform interface, are securely enacted by means of specific smart contracts which prevent any unauthorized use by the Stakeholder, by accurately matching their usage requests with the correspondent consent-based restrictions; (iii) encrypted metadata regarding each query by the Stakeholders is safely registered and stored on the MHMD blockchain, in order to keep trace of any ‘transaction’ activated through the Platform.</p>	

b) With reference to Individual Data:

Principle	Description of the action	Risk status for individual rights
<p><b>Transparency</b></p>	<p>Data subjects who decide to register to the MHMD APP, with the main objective of generously making their personal, lifestyle, health and medical data available to scientific research, are provided with a clear and comprehensive privacy notice – written specifically for the Project – detailing all the necessary elements as per Art. 13 of the GDPR.</p>	
<p><b>Lawfulness and fairness</b></p>	<p>At the moment of the registration to the APP, Users are requested to give their optional and distinct consents to a number of specific activities, by setting their own preferences in relation to a wide range of aspects that enable them to exercise full control over each processing of their data.</p> <p>E.g. (this wording is not used, but merely illustrative) <i>‘Based on the information received, I do consent to the processing of my personal data, including clinical data:</i></p> <ul style="list-style-type: none"> <li>○ <i>Only for a specific disease: _____ (e.g. diabetes);</i></li> <li>○ <i>Not for certain clinical areas or diseases: _____ (e.g. cardiac pathologies and cancer);</i></li> <li>○ <i>For a given period of time: (e.g. available until 31 December 2030);</i></li> <li>○ <i>For secondary research usage:</i> <ul style="list-style-type: none"> <li>▪ <i>only for research carried out by the Hospital (YES/NO);</i></li> </ul> </li> </ul>	

	<ul style="list-style-type: none"> <li>▪ <i>for research carried out by a third party (YES/NO)</i>.</li> </ul> <p>Granularity is the main requirement of the consents that are requested to the APP Users, so as to allow them to freely and easily establish everything that may or may not be done with the datasets they decide to make available to the Project.</p>	
<b>Purpose limitation</b>	<p>The ‘itemized’ consents provided by the data subjects are operationalized in the Platform thanks to smart contracts which automatically audit the Stakeholders’ data access queries to verify that the usage conditions outlined by the User are properly met. In other words, the ‘doors’ will be open only to those Stakeholders who declare, under their own responsibility, that they want to use the Individual Data for one (or more) of the purposes that were originally and expressly consented by the patients.</p>	
<b>Data minimization</b>	<p>The MHMD Platform is designed to apply strong encryption algorithms and end-to-end encryption by design and to ensure that, whenever the purposes described by the Stakeholders querying the Platform can be fulfilled by a “<i>processing which does not permit, or no longer permits, the identification of the data subjects</i>” (Art. 89.2 of the GDPR), only anonymized Clinical Datasets are made available to achieve such purposes. To strengthen security, Individual Datasets are pseudonymized by default, to prevent re-identification of individuals except by means of further separate information. A number of additional security measures have been put in place, as already outlined above (and better described below), also in connection with distributed ledger mechanisms to avoid</p>	

	<p>misinterpretation, abuse, fraudulent usage, unauthorized access to the data and similar circumstances.</p>	
<b>Accuracy</b>	<p>In order to ensure that Individual Data are always accurate and kept up-to-date, the data subjects can at any moment ask the APP Operator to erase or rectify any incorrect data without undue delay.</p>	
<b>Storage Limitations</b>	<p>According to 'storage limitation' principle, personal data must be kept in a form which permits identification of data subjects for no longer than is necessary to achieve the purposes for which such data have been collected. Nonetheless, Art. 5.1(e) of the GDPR establishes that personal data may be stored for longer periods insofar as they are processed solely for, <i>inter alia</i>, scientific research purposes in accordance with Article 89(1), to safeguard the rights and freedoms of the data subjects. On account of this and considering that a number of innovative security measures are applied in addition to those laid down by said Art. 89 (mainly consisting in the pseudonymization of data), Individual Datasets will be stored, at least under appropriate pseudonymization, until Users change the data access/usage limitations through the dedicated settings interface.</p>	
<b>Integrity and confidentiality</b>	<p>Many state-of-the-art security measures have been implemented and many others have been (and are being) developed – or taken to a higher technological level – specifically for the Project, with a view to guaranteeing seamless integrity and confidentiality of the data collected. Among such measures: (i) 'multi-level' encryption schemes will be applied also to the Individual Datasets; (ii) the consents given by the data subject, translated into access permissions they can set thanks to a MHMD-native interface, are</p>	

	securely enacted by means of specific smart contracts which prevent any unauthorized use by the Stakeholder, by accurately matching their usage requests with the correspondent consent-based restrictions; (iii) encrypted metadata regarding each query by the Stakeholders is safely registered and stored on the MHMD blockchain, in order to keep trace of any 'transaction' activated through the Platform.	
--	---	--

## 7. MHMD BLOCKCHAIN

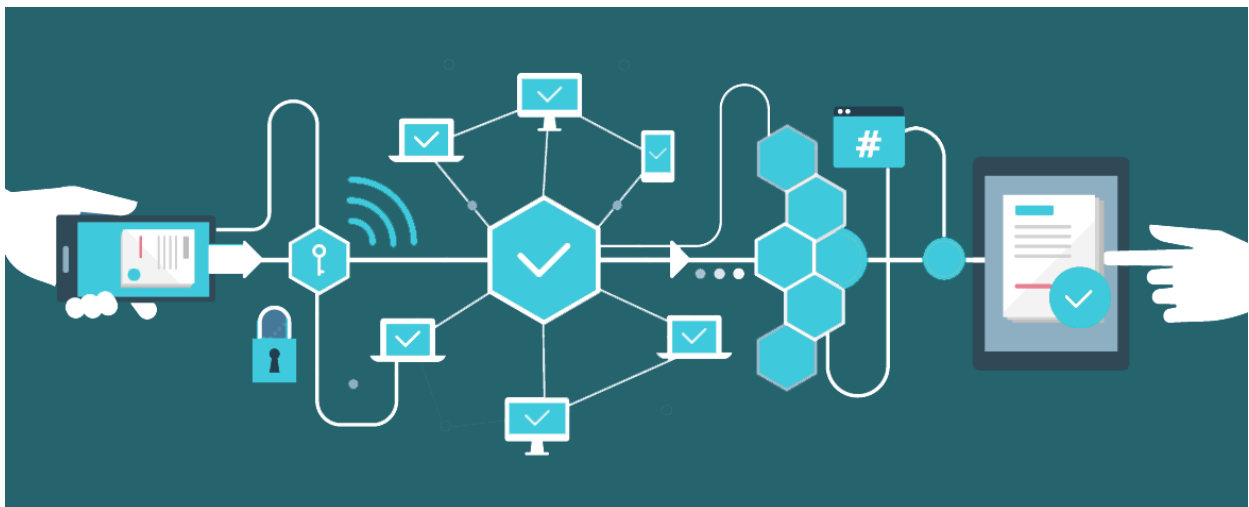
MHMD relies on a decentralized, blockchain-based infrastructure that provides a new mechanism of trust and direct, value-based relationships between Hospitals, data subjects, research centres and businesses and monitors and securely orchestrates any processing of the Datasets, be it under the Segregated Computation or the Secure Sharing Model.

It is worth mentioning that this technology amounts to an append-only ledger organized as a chain of blocks that relies on a peer-to-peer network to perform its management, updates and operations.

Roughly speaking, blocks are merely containers for transactions and they can be linked to an existing chain of blocks, allowing it to grow. As a data structure, a blockchain has two distinctive features which are block timestamps and hash pointers that link the last block of the chain to the previous one, in such a way that any modification made on a block compels the regeneration of the following blocks in the chain.

Given the current state-of-the-art in the distributed ledger technologies field, the prototype developed under MHMD entails enhanced and privacy-preserving peculiarities.

Firstly, in order to meet the highest standard from both a data protection and security standpoint, a private and permissioned blockchain was designed and implemented. Secondly, this distributed ledger infrastructure has been deployed to enhance security and to make consent-based data exchanges tamper proof, while personal data are stored exclusively off chain.

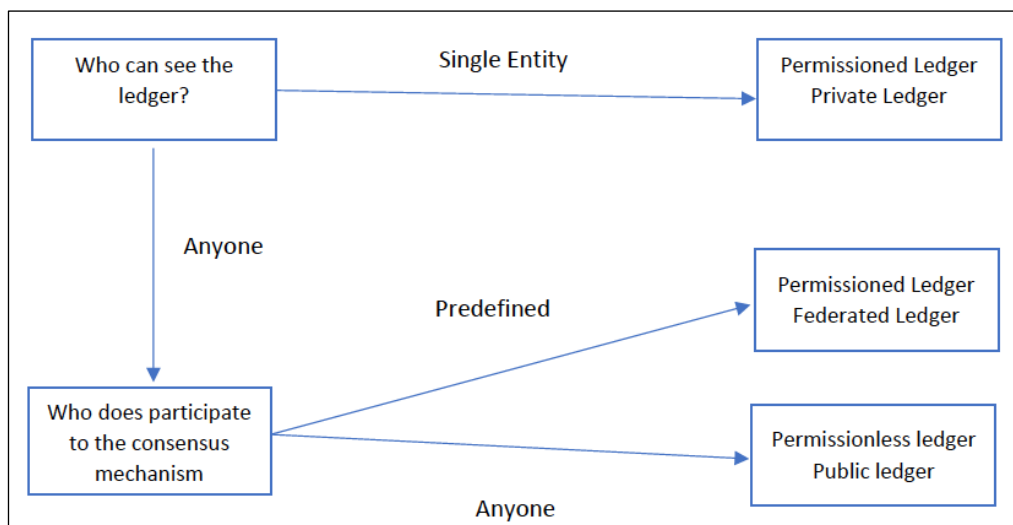


## 7.1. DESCRIPTION OF MHMD BLOCKCHAIN

As recommended by the *European Blockchain Observatory and Forum* (“**EU Observatory**”) in its ‘*Blockchain and the GDPR*’ report,<sup>31</sup> in case of need to store personal data, it is necessary to rely on private and permissioned blockchain.

It's worth making a quick ‘technical’ specification:

- i. in public, permissionless blockchains, anyone is allowed to join the network and become a participating node or a validating node;<sup>32</sup>
- ii. in public and permissioned blockchains, anyone can be a participating node and see all data, but only pre-approved actors can become validating nodes and add data to the ledger;
- iii. in private and permissioned blockchains, validating nodes and participating nodes must be preapproved by a governance of actors, generally in the form of a consortium of companies or government agencies. Furthermore, in some cases, there are rules in place that define who is able to see what data.

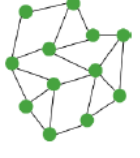

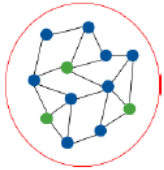
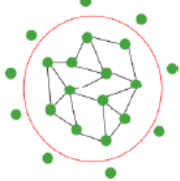


Accordingly, a private and permissioned blockchain – where each user must be formally admitted before joining the network – was implemented in MHMD, so as to ensure that any ‘transaction’ between Stakeholders is appropriately traceable and auditable, and

<sup>31</sup> European Blockchain Observatory and Forum, *Blockchain and the GDPR*, thematic report, October 2018.

<sup>32</sup> A blockchain network consists of a group of server nodes that store synchronized copies of the same data. There are usually two types of nodes: (i) validating nodes, are allowed to add data to the ledger, according to an agreed-upon algorithm called a consensus mechanism; (ii) participating nodes, which store synchronized copies of the data. Depending on the specific technology, not all nodes may necessarily store all data.

that all network participants commit to a set of specific terms and conditions. Read permissions may be public or restricted to an arbitrary extent.

<b>Blockchain type</b>	<b>Explanation</b>	<b>Example</b>	<b>Visualization</b>
<i>Public permissionless blockchains</i>	In these blockchain systems, everybody can participate in the consensus mechanism of the blockchain. Also, everyone in the world with a connection to the internet is able to transact and see the full transaction log.	Bitcoin, LiteCoin, Ethereum	
<i>Public permissioned blockchains</i>	These blockchain systems allow everyone with a connection to the internet to transact and see the transaction log of the blockchain, but only a restricted amount of nodes can participate in the consensus mechanism.	Ripple, private versions of Ethereum	
<i>Private permissioned blockchains</i>	These blockchain systems restrict both the ability to transact and view the transaction log to only the participating nodes in the system, and the architect or owner of the blockchain system is able to determine who can participate in the blockchain system and which node can participate in the consensus mechanism.	Rubix, Hyperledger	
<i>Private permissionless blockchains</i>	These blockchain systems are restricted in who can transact and see the transaction log, but the consensus mechanism is open to anyone.	(Partially) Exonum	

The most widely known instances of permissioned blockchain are Hyperledger Fabric and R3 Corda.

Considering this, the blockchain technology chosen for MHMD is Hyperledger Fabric, based on the modular characteristic and flexibility provided in view of implementing a more customized ledger according to the purpose, security, and performance needed for the Project.



HyperLedger Fabric	
Cryptocurrency required	None
Network	Permissioned
Transactions	Anonymous or private
Consensus <sup>33</sup>	PBFT (Practical Byzantine Fault Tolerance)
Smart contracts (business logic)	Yes (chaincode)
Time between blocks	Real-time
Language	Golang, Java
Companies behind	Linux foundation + IBM

*Characteristic of HyperLedger Fabric*

More specifically, in order to give effect to the principles of privacy-by-design and data minimization while shaping the Project, attention was drawn on the need to prevent anyone who may access the information stored on the blockchain from identifying the parties and the personal data involved in the relevant transactions (*i.e.* ensuring unlinkability).

To accomplish this goal, it was necessary to implement a new consensus protocol that allows to validate the essential parts of the transactional process (like endorsement and identity verification) while maintaining the privacy of the operation.

This issue was solved by using a new hybrid consensus scheme that maintains the principles of *Practical Byzantine Fault Tolerance*,<sup>34</sup> but which is able to authenticate the transactions without leaking any confidential information, by relying on *Zero Knowledge Proof* ('ZKP').

A privacy-preserving consensus algorithm was designed based on PBFT, named “proof-of-privacy”, that relies on Okamoto-Schnorr's blind signature scheme.

---

<sup>33</sup> Consensus mechanism is the core of the blockchain. In distributed systems, multiple processes communicate to enable system operation. Faults may occur anywhere throughout a distributed system, e.g. processes may crash or adversaries may send malicious messages to processes. Distributed systems use consensus protocols to achieve reliability despite faults. Through consensus, the shared state of the ledger comes to an agreement upon a global state, allowing all the nodes of the network to reach the same ledger state within a certain period of time. Achieving consensus in a distributed system is challenging, as it must be resilient to node failures, network delays and the existence of malicious nodes. There are three basic consensus mechanism categories: (i) Proof-of-Work (PoW); (ii) Proof-of-Stake (PoS); (iii) Practical Byzantine Fault Tolerance ('PBFT').

<sup>34</sup> This scheme is faster, scalable, and democratic (if 50% plus one of the nodes valid the new block, this is added to the chain). Nonetheless, this protocol needs the authentication of each node. To overcome the lack of anonymity, a Zero Knowledge Proof (ZKP) for node’s authentication has been implemented, maintaining the confidentiality of each node in the network. Consensus is reached once the node have received enough message with the same response.

The process is executed once the endorsing peers validate that the transaction is well formed before endorsing it. The new block validation process takes place as follows:

- i. one of the peers is elected as a 'leader';
- ii. the leader orders transactions candidates which should be included in a block and broadcasts this list of ordered transactions to all other validation peers in the network;
- iii. when each of validation peers receives an ordered list of transactions, then it starts executing them one by one;
- iv. as soon as all transactions are executed, each validation peer calculates a hash code for the newly created block (the hash code includes hashes for executed transactions and final state of the world);
- v. validation peer will verify the identity of the node that is proposing the new block by using ZKP. If the validation peer accepts the proof as valid, then the peer is accepting the block;
- vi. each validation peer broadcasts its answer to other peers in the network and starts counting responses from them;
- vii. if a node sees that 2/3 of all validation peers have the same hash code as a result of the transaction execution, it will commit the new block to their local copy of the ledger.

In addition to the above – as better explained in the preceding paragraphs – data lifecycle is managed through a Catalogue so that it can be referenced in the blockchain by storing a hash value of the indexed data items (PIDs).

Indexing data items means:

- first step: when new Datasets are indexed, an update of the central MHMD Catalogue must be pushed. This update shouldn't be made before the second step is complete, but it can be prepared (asynchronously) ahead of time;
- second step: a human intervention is necessary to establish permission settings for the registered Datasets to be exposed on the Catalogue. This step can be done in batches: a single permission setting can be used for all data item related to a specific data subject. Settings are then included into the relevant smart contract that will govern and authorize data transactions.

- third step: assign the blockchain identifier to each data item. The data provider – namely the Hospital or the User – keeps a mapping table between the blockchain identifier and the data item identifier, called ‘local mapping DB’,<sup>35</sup> which is of the utmost importance in order to ensure that individual rights are appropriately put into effect.

Therefore, a metadata description of the information registered on the Project’s blockchain appears safely in the Catalogue, which is freely open for browsing to all authorized Stakeholders. This process allows the blockchain to maintain the records of available data and its associated history without the need to record any personal data which, therefore, remain solely off-chain.

Using one-way cryptographic algorithms to describe data and transactions results in an anonymous ledger, which also prevents from statistical inference to locate data or individuals thanks to k-anonymity like models.

Following two years of intense prototyping, a GDPR-compliant permissioned blockchain is now deployed in pioneer Hospitals and research centres in Europe, to validate the concept.

## 7.2. BLOCKCHAIN AS A SECURITY MEASURE

As anticipated, the blockchain is a decentralized ledger that cannot be tampered with and its state replication through the network is based on protocols that ensure the broadcast of an agreed version of the last state. This is reached on the basis of the following properties:

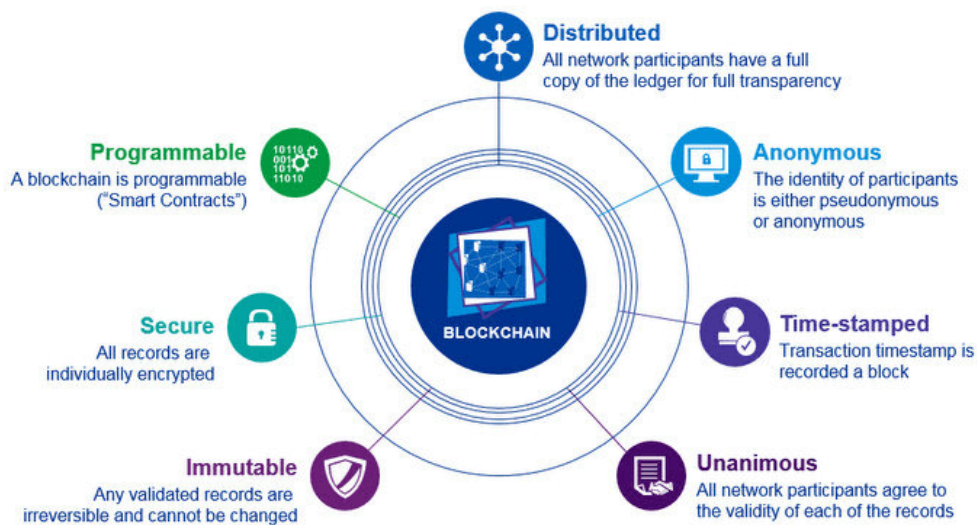
- ✓ **Consensus**: the protocol used to agree with the rest of the nodes about who will propose the next block to be added to the ledger. This process implies a distributed responsibility for the conformity of the entire chain and avoids any possibility of double-spending;
- ✓ **Validity**: when a member of a blockchain proposes to make a transaction and updates the system, each other member on the blockchain can check and validate whether it is a valid deed, valid state or a valid update;
- ✓ **Uniqueness**: updates to the blockchain are unique. From state A, the blockchain cannot go to both state B and state C, although both state B and state C are valid. The consensus-driven blockchain will have to agree on one among state B and state C, as the next state;

---

<sup>35</sup> The ‘Local Mapping DB’ is the database which contains the link between the data source identifier and the blockchain identifier.

- ✓ **Immutability:** this is ensured while the chain is growing, making it infeasible to change any data stored in a block (the probability of tampering the ledger is reduced as the chain grows);
- ✓ **Authentication:** all the transactions are validated by members of the network. This authentication process is based on digital signatures and depends on the implementation if the transactions are marked to users or network addresses.

## Properties of Digital Ledger Technology (DLT)



Authorized Stakeholders can browse the metadata made available on the Catalogue, then selecting a specific Dataset. In case of pseudonymized data, the smart contract accompanying the data would need to incorporate the relevant individual consent. Once this is checked, the chosen Dataset undergoes the further security safeguards automatically predefined according to its nature and the outcome is then made available to the authorized Stakeholder who entered the initial query into the Platform, so triggering the related smart contract.

Generally, even if strong encryption is applied on personal data, when needed, the result is in some cases to prove not fully anonymous given that, as long as the decryption key exists somewhere, the data can still be singled-out, leading to a reversal risk.

Another risk is that the linkability of encrypted data to an individual can be reached by further examining patterns of usage or context, or by comparison to other pieces of information.

On top of this, cryptography technologies and science are subject to a seamless evolution.

For these reasons, all such risks have been avoided by preventing the registration of any kind of personal/sensitive data on chain both in the Segregated Computation and the Secure Sharing Model, thus fulfilling the requirements of privacy and security in accordance with the GDPR and the EU Observatory's guidelines. The data are stored solely off-chain, in MHMD distributed database.

### 7.3. DATA SUBJECTS' RIGHTS IN CONNECTION WITH MHMD BLOCKCHAIN

As a consequence of the operational and technical choices explained above, data subject's rights are entirely and appropriately safeguarded.

In a recent report, the French Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés*, briefly "CNIL") explicitly revealed its concerns with regards to the exercise of data subjects' rights in the blockchain environment. Notwithstanding the choice of a private and permissioned blockchain, more in line with GDPR's principles and obligations, the CNIL focused its attention on the remaining unsolved issues at stake.

As a matter of fact, given the immutability of the data retained on a blockchain, compliance with the GDPR has to be ensured by means of technical loopholes, with specific reference to the rights of erasure, limitation and rectification.<sup>36</sup>

A similar position has been expressed also by the EU Observatory, which took a step forward by stating that «*these issues are not resolved just by moving to a private, permissioned blockchain network, unless that network is designed in a way that each and every piece of data is readable by only the parties that absolutely need to, and can be rectified or erased at the request of the data subject*».<sup>37</sup>

This is precisely the idea beyond MHMD blockchain. Data subjects' rights can indeed be easily and unhinderedly exercised off chain, by means of a specific request to the Hospitals or to the APP Operator or, under certain circumstances, by changing the related settings in the APP.

Needless to specify that the Platform is set up in such a way as to notify the Hospitals with no delay in those cases when, for whatever reason, an individual request of exercise of one or more rights in connection with the Clinical Datasets is not made directly to the Hospitals but to the Platform Operator (e.g. writing to the email address indicated in MHMD

<sup>36</sup> See CNIL, *Solutions for a responsible use of the blockchain in the context of personal data*, p. 9.

<sup>37</sup> *Blockchain and the GDPR*, p. 25.

Platform Privacy Policy or by modifying the relevant setting – such as in case of withdrawing one or more consents).<sup>38</sup>

However, even if no personal data are registered on the blockchain, the following actions are envisaged in the event of exercise of individual rights, to guarantee full accountability:

RIGHT	CONSEQUENT AUTOMATED ACTION
<p style="text-align: center;"><b><u>Access</u></b></p> <p>(obtaining confirmation as to whether or not personal data concerning the data subject are being processed and, in case, access to such data and to all relevant information regarding the processing)</p>	<p>A transaction is registered on the blockchain which indicates that access was requested for a specific data item. The extraction and delivery of the data will then take place off-chain (under the responsibility of the competent controller).</p>
<p style="text-align: center;"><b><u>Rectification</u></b></p> <p>(obtaining without undue delay the rectification of inaccurate personal data)</p>	<p>A specific smart contract is activated in order to prevent any party to the Project, including particularly the Stakeholder, to access the inaccurate data, while allowing to collect and process only the amended Datasets.</p>
<p style="text-align: center;"><b><u>Erasure</u></b></p> <p>(obtaining from the controller, when certain conditions laid down by Art. 17 of the GDPR are met, the erasure of personal data concerning him or her without undue delay)</p>	<p>The right to erasure is achieved by ‘breaking the link’, <i>i.e.</i> deleting the entries, in the Local mapping DB, so preventing anyone from being able to associate a data source identifier and the relevant blockchain identifier. As a result of this, a smart contract will forbid anyone from accessing the data on the Platform, thus guaranteeing a result whose effects are reasonably completely equivalent to those of material cancellation (which will obviously will be carried out off-chain with no delay, insofar at least one of the conditions set forth by Art. 17.1 is satisfied)</p>
<p style="text-align: center;"><b><u>Restriction of processing</u></b></p>	<p>Where the processing is restricted in the cases set out by Art. 18.1, a specific smart contract will be executed to</p>

<sup>38</sup> Reference is made solely to the Clinical data, because all requests to exercise the rights in relation to Individual Data can only be addressed to the Platform Operator.

<p>(obtaining from the controller restriction of processing, meaning that the personal data shall, with the exception of storage, only be processed, <i>inter alia</i>, with the data subject's consent or for the establishment, exercise or defence of legal claims)</p>	<p>prevent all Stakeholders from carrying out any kind of processing, with the exception of storage, unless (i) the data subject has given his/her consent, or (ii) for the establishment, exercise or defence of legal claims.</p>
<p style="text-align: center;"><b><u>Notification</u></b></p> <p>(the controller shall communicate any rectification or erasure of personal data or restriction of processing to each recipient to whom the personal data have been disclosed, unless this proves impossible or involves disproportionate effort. The controller shall inform the data subject about those recipients if the data subject requests it)</p>	<p>The measures described above in regards of the request of rectification or erasure of personal data, or restriction of processing, ensure that the relevant legal effects are appropriately extended to any user of MHMD blockchain, thus meeting notification requirement which, however, shall apply only where this does not prove impossible or does not involve a disproportionate effort.</p>
<p style="text-align: center;"><b><u>Portability</u></b></p> <p>(receiving the personal data provided to the controller in a structured, commonly used and machine-readable format and, where requested, having such data directly transmitted to another controller)</p>	<p>The same actions described above in regards of the right of access shall apply, mutatis mutandis, to the requests of portability. The duty to provide the data subject with the data in a structured, commonly used and machine-readable format lies exclusively with the controller (namely the Hospital for Clinical Data, or the Platform Operator for Individual data).</p>
<p style="text-align: center;"><b><u>Withdrawal of consent</u></b></p> <p>(revoking the consent(s) at any time, bearing in mind that it shall be as easy to withdraw as to give consent)</p>	<p>At any time a Patient or User should withdraw one or more of the consents previously provided – as repeatedly explained above in detail – a smart contract will be run to definitively (at least until the data subjects provides the consent again) prevent Stakeholders from accessing the associated Datasets for purposes which do not meet the applicable usage requirements.</p>
<p style="text-align: center;"><b><u>Automated decision-making process</u></b></p> <p>(not being subject to a decision based solely on automated processing, including profiling, which produces legal effects or similarly significantly affects the data subjects, unless this processing is based on his/her specific consent and provided that suitable measures to safeguard the data subject's rights and freedoms are in place, with</p>	<p>In no case the automated processing operations carried out through the Platform, as operationalized by both the blockchain and the associated smart contracts, may determine or result in a decision which produces legal effects for the data subjects, or which may in any case significantly affect them.</p>



<p><b>particular regard to the right to obtain human intervention, or to express the individual's point of view and to contest the decision)</b></p>	<p>Any such decision can be taken and is therefore exclusively left to the Stakeholders, without any chance that the Platform Operator carries out any fully automated individual decision-making as governed by Art. 22 of the GDPR.</p>
--	---

Another important property, in terms of compliance, regards the type of information transmitted in the whole process, which can be classified into the following groups:

- Source Data: no storage of data will be carried out by MHMD Platform. As better detailed below, all Datasets will be separately stored in the Hospitals' private repositories, under the Segregated Computation Model, and in the ways provided by each of the APPs and systems that each User will decide to connect and so 'make open' for the Project (e.g. each lifestyle and wellness APP's and social network platform's storage arrangements shall apply);
- Metadata: the MHMD Catalogue which features high-level descriptive statistics on encrypted Datasets and allows creating data queries, lies separately from the blockchain;
- Consent: the statement or the clear affirmative action by means of which the data subjects signify, in a free, specific, informed and unambiguous manner, their agreement to the processing of their personal data, will be operationalized by a dedicated smart contract which will enforce all access and usage restrictions outlined by Patients and the Users, so preventing any unauthorized processing of the Datasets;
- Dataset: any specifications relevant to the Datasets have been given in the preceding paragraphs.

Reliance on smart contracts permits the processing of both Clinical and Individual Datasets by the Stakeholders in full compliance with data protection and security principles, because the correspondence of the permissions granted (by the data subjects) and those requested (by Researchers and Private Businesses) – from which the legal agreement between such parties originates – is made tamper-proof and immediately enforceable through automated and self-executable codes.

In this scenario, the sharing of personal and medical data cannot be hampered or undermined by unauthorized third parties, as the access to such data is allowed only to those parties which, following the undeceivable checks carried out by the smart contracts, are



found to appropriately fulfill any applicable requirement (in terms of purpose envisaged in the query made to the Platform ‘against’ the consents given by the data subject).

Indeed, blockchain technology helps making online data transactions secure, by also eliminating any middlemen occurrences, as well as preventing the needed intervention of a trusted third party for validating the request made by Stakeholders for both Secure Sharing and Segregated Computation.

#### 7.4. PARTICIPANTS IN THE CONTEXT OF MHMD BLOCKCHAIN

In the traditional client-provider model, it is relatively easy to identify the data controller: there is almost always an entity that is offering some product or service and that consequently determines the purpose and means for the processing, sets up the systems to do it and processes the data made available by the data subjects.

On the contrary, identifying who is the controller and/or the processor is very challenging in a distributed ledger scenario.

Although MHMD, in line with the minimization principle set forth in the GDPR, is designed and set up in such a way as to prevent any processing of personal data from taking place on the blockchain, each participant shall be considered – and is specifically instructed in the applicable Terms and conditions to maintain the role – as an autonomous data controller *vis-à-vis* the others, with reference to both the data that are input and those contained in the blockchain that may be available on his/her device.

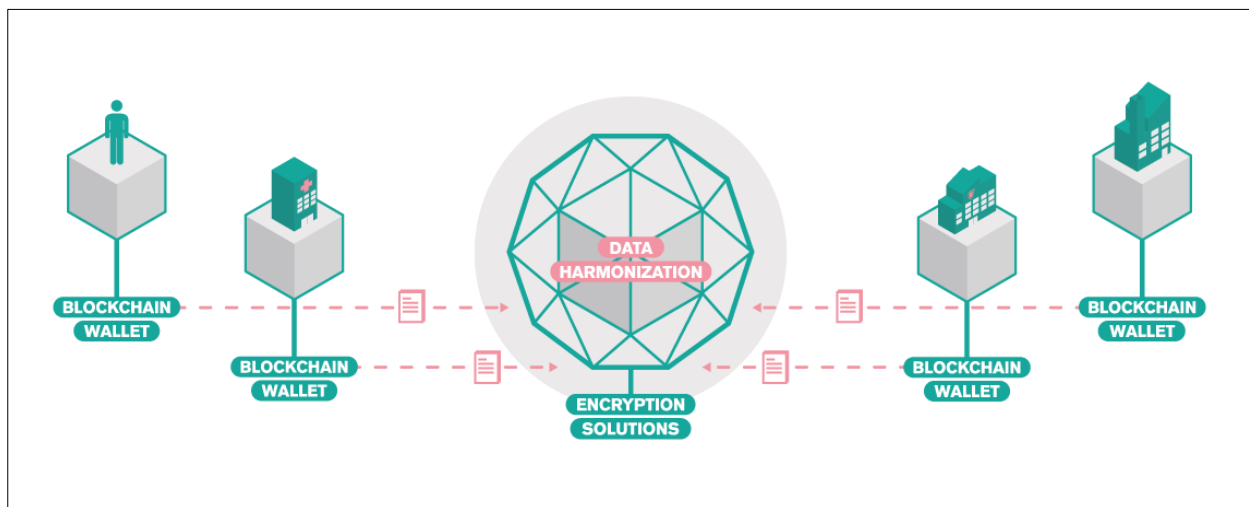
In this way, the most cautious and sound allocation of roles is ensured to avoid unreasonable sharing of responsibilities. Therefore, each participant to MHMD blockchain must be held individually liable, irrespective of its function off-chain, for any potential breach of the applicable legal, regulatory and contractual obligations.

Accordingly, the Platform Operator cannot take responsibility for breaches which are reasonably impossible to predict – also given the wide application of MHMD blockchain, from a territorial standpoint – and which arise from reckless or unlawful conducts by the participants.<sup>39</sup>

---

<sup>39</sup> In this context, it seems appropriate to provide a parallelism with the legislation applicable to the ‘Information society services’, laid down by the Directive 2000/31/EC. In fact, information society services provider shall not be deemed responsible (and so liable) for any users’ misconduct while they navigate on their online networks, unless it can be proved that i) such providers were somehow involved in the illegal activity or ii) once becoming aware of the breach, they had not taken the appropriate action to address it. Hence, also in this case, the Platform Operator – acting as a sort of ‘service provider’ – cannot be similarly considered responsible, since it is not possible to predict every and each potential unlawful action undertaken by Participants in the future.

The alternative scenario where all participants are considered jointly responsible for the actions and choices taken by the others was excluded entirely, due to the significant weakening of the system in terms of accountability and precise allocation of responsibilities. Moreover, it was considered that neither Hospitals nor any Stakeholder would be willing to participate in a project in which, as far as the individual responsibilities can be partitioned according to Art. 26 of the Regulation, a serious risk would still remain to suffer the detrimental effects arising from violations committed by others.



## 8. MHMD SECURITY FRAMEWORK

The security thresholds are set, from a regulatory standpoint, by the following provisions:

- a) *Art. 24 of the GDPR – Accountability*: on account of the nature, scope, context and purposes of processing, as well as the risks of varying likelihood and severity for the rights and freedoms of data subjects, appropriate technical and organisational measures must be implemented to ensure and to put the controller in condition to demonstrate that the data are processed in compliance with the Regulation.
- b) *Art. 25.1 of the GDPR – Privacy by design*: taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing, as well as the risks of varying likelihood and severity for rights and freedoms of the data subjects arising from the processing, both at the time of the determination of the means for processing and at the time of the processing itself, appropriate technical and organisational measures must be adopted, such as pseudonymization, to implement the data-protection principles in an effective manner in order to meet the requirements of the GDPR and protect the rights of the data subjects;
- c) *Art. 25.2 of the GDPR – Privacy by default*: appropriate technical and organisational measures must be implemented so that, by default, data minimization is put into effect, e.g. ensuring that only personal data which are necessary for each specific purpose of the processing are processed (this also applies to the amount of the data collected, the extent of their processing, the period of storage and their accessibility by duly authorized people);
- d) *Art. 32 of the GDPR – Risk-based approach*: taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing, as well as the risk of varying likelihood and severity for the rights and freedoms of data subjects, appropriate technical and organisational measures must be adopted to ensure a level of security appropriate to the risk, taking into particular account the risks that may arise from accidental or unlawful destruction, loss, alteration, unauthorized disclosure of or access to the personal data;
- e) *Art. 89 of the GDPR – Specific guarantees for research*: data processing for scientific research purposes must be subject to appropriate safeguards for the rights and freedoms of the data subjects, including in particular technical and organisational measures suitable to ensure data minimization, such as pseudonymization.

The design and implementation of the Project offered the opportunity to better investigate and test some innovative security techniques, aimed at ensuring suitable levels of protection for the data collected and strengthening the resistance and resilience of the entire infrastructure against both Platform-intrinsic and external threats.

## 8.1 DE-IDENTIFICATION MEASURES

To achieve both ‘computational privacy’, dealing with the process of computing a function in the safest way possible with a view to ensuring data minimization, as well as ‘output privacy’, *i.e.* preventing computation from leaking significant parts of the original inputs received, both centralized and distributed data processing models have been adopted:

- ✓ *Centralized*: under this model, data can be accessed after an additional security level has been applied that includes anonymization, or privacy preserving interactive data querying / data mining through a centralized differential privacy tool;
- ✓ *Distributed*: specific operations related to privacy-preserving data mining (PPDM) are executed through a secure distributed processing protocol which is usually based on Secure Multiparty Computation .

### 8.1.1 PSEUDONYMIZED AND ANONYMIZED DATA

Pseudonymization is a GDPR-approved technique that encodes personal data with artificial identifiers, such as a random alias or code.

The GDPR specifically describes pseudonymization in Art. 4(5), as the processing of personal data in such a manner that they «*can no longer be attributed to a specific data subject without the use of additional information*», provided that such 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 individual.<sup>40</sup>

Evidently, this does not amount to anonymization because the data can still be linked back to a person, by matching the pseudonymized dataset with other information stored elsewhere.<sup>41</sup>

---

<sup>40</sup> In brief, it is a privacy-enhancing technique where directly identifying data is held separately and securely from processed data to ensure non-attribution.

<sup>41</sup> Recital 26 of the GDPR states that «*personal data which have undergone pseudonymization, which could be attributed to a natural person by the use of additional information, should be considered to be information on an identifiable natural*

Nonetheless, pseudonymization may significantly reduce the risks associated with specific types of data processing, while also maintaining the data's utility. For this reason, the GDPR has introduced incentives to pseudonymize the data, taking a more flexible approach than the traditional 'binary' of the Directive, focusing on the risk that data will reveal identifiable individuals. Thus, the key distinction between pseudonymization and anonymization is whether the individuals can be still singled-out with reasonable effort.<sup>42</sup>

To illustrate the concept of reidentification risk, it is important to distinguish between direct and indirect identifiers (also called 'quasi-identifiers'). The International Organization for Standardization ('ISO') defines direct identifiers as "*data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain*",<sup>43</sup> namely data points that correspond directly to a person's identity (such as a name or social security number).

Indirect identifiers are data that do not allow to pick out a specific person, but are such as to reveal individual identities if combined with additional data points (e.g. a frequently-cited study found that 87 of U.S. citizens can be uniquely identified by combining three indirect identifiers: date of birth, gender and ZIP code. In other words, while no individual can be singled out based on just a date of birth, when combined with gender and ZIP code, the lens focuses on a specific identity).<sup>44</sup>

Pseudonymization involves removing or obscuring direct identifiers and, in some cases, also indirect identifiers that could combine to reveal a person's identity. These data points are then held in a separate database that could be linked to the de-identified data through the use of a key, such as a random identification number or some other pseudonym.

This process triggers the risk that a data breach may permit an attacker to obtain the key, or otherwise link the pseudonymized dataset to individual identities. To address this concern, as specified by Recital 75 of the GDPR, data controllers must implement appropriate safeguards to prevent the *«unauthorized reversal of pseudonymization»*, by relying on technical (e.g. encryption, hashing or tokenization) and organizational (e.g.

---

person» (i.e. personal data). See also the WP29's *Opinion 05/2014 on 'Anonymisation Techniques'* (WP216), dated 10 April 2014.

<sup>42</sup> «To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments» (Recital 26 of the Regulation).

<sup>43</sup> ISO 25237:2017 - *Health informatics – Pseudonymization* ([link](#)).

<sup>44</sup> L. Sweeney, *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, in *Data Privacy Working Paper 3*. Pittsburgh 2000 ([link](#)). See also L. Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, in *Journal of Law, Medicine & Ethics*, 25, nos. 2&3 (1997): 98-110.

agreements, policies, privacy-by-design) measures which ensure seamless separation between pseudonymous data and the re-identification key.

Under the Directive, even when controllers deleted all identifying information and could not themselves reidentify a dataset, the WP29 found that the data was still personal if any third party could conceivably reidentify the data sometime in the future.

In contrast, by focusing on whether reidentification is ‘reasonably likely’ (Recital 26 of the Regulation), taking into account all objective factors, such as the costs of and the amount of time required for identification, the available technology at the time of the processing and technological developments, the GDPR may provide greater flexibility than the Directive. For instance, where the controller deletes the identification key and the remaining indirect identifiers pose little risk of someone being able to identify an individual, the controller may argue that there is no reasonable risk of reidentification.

### 8.1.2 SYNTHETIC DATA

Synthetic data, i.e. faithful copies of original data sets that don’t contain any identifiable information have been used more recently for biomedical applications. Their development requires articulated quality control process to guarantee statistical soundness and conformity to the original. MHMD leverages mature synthetic data generation pipelines the results of which can be exchanged with minimal risk of re-identification, providing privacy protection substantially more robust than other anonymization techniques.

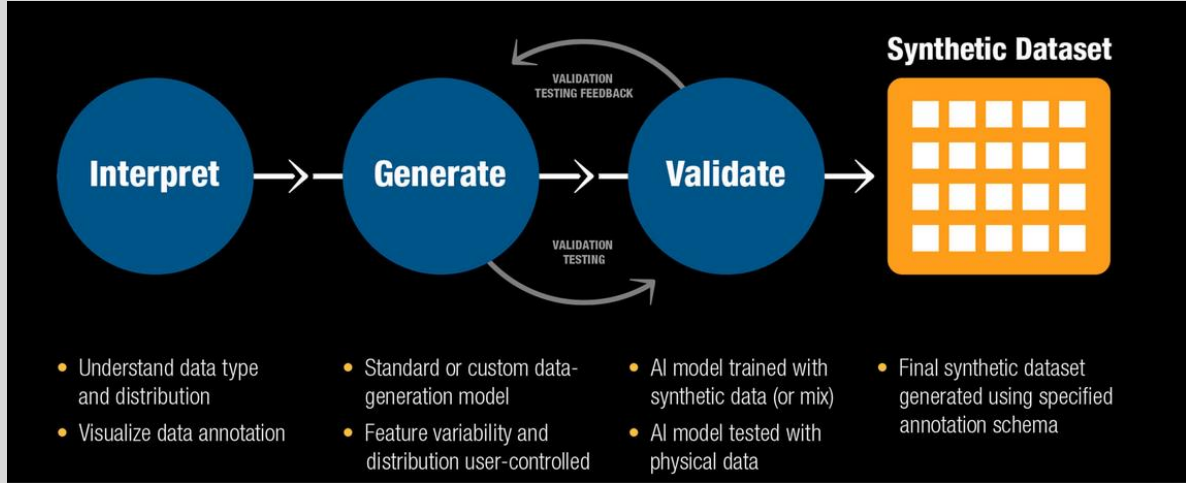
#### SYNTHETIC DATA

In addition to the de-identification techniques described above, another innovative privacy-enhancing process has been scrutinized and tested during the definition of the Project, to deal with those cases in which the other techniques described above may not be sufficient to guarantee individual privacy: data synthetization.

Synthetic data are generated using a combination of aggregate statistics from a known population. Using these inputs, virtual patients are created from scratch by drawing from the distributions, so that a significant amount of realistic data can be generated with an almost-zero risk of being able to identify the original data subjects.

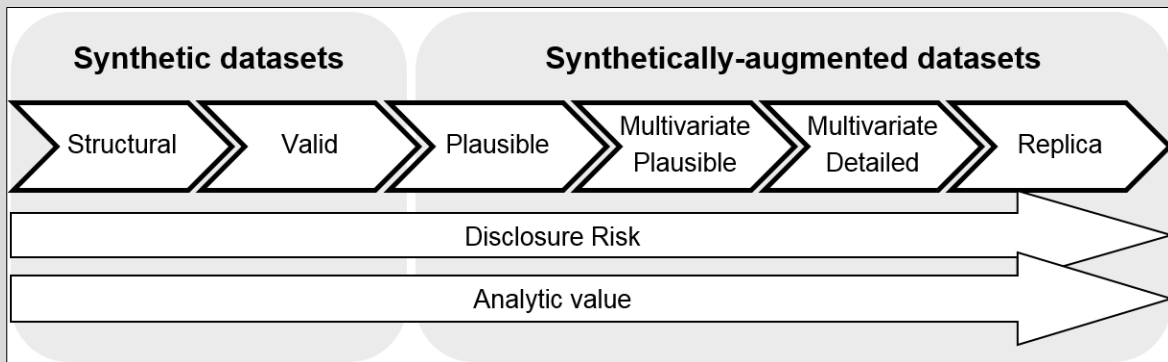
Although it is considered a strong security measure, pseudonymization always comes with the risks of someone being able to re-identify the individuals, while anonymization may somewhat diminish the utility of the data, as it is not always possible to carry out certain studies by means of fully de-identified datasets.

In this complex scenario, fully synthetic data help to overcome most of the issues at stake.



These data are indeed totally made-up and do not contain any of the original identifiable information. They are generated after the density function of the attributes in the original dataset is identified and their parameters has been estimated. Then for each attribute, privacy protected series are generated by randomly picking up the values from said estimated functions. Multiple imputation and bootstrap methods are few classical techniques used to generate fully synthetic data.

These data, automatically generated by making use of machine learning algorithms, are based on recursive conditional parameter aggregation, operating within global statistical models which, by definition, do not allow any personal re-identification of original individual datasets.



This image is taken from the ONS methodology working paper series number 16 - Synthetic data pilot<sup>45</sup>

<sup>45</sup> 'A pilot study investigating the demands and requirements for synthetic datasets, and exploring possible tools to produce synthetic data for specific user requirements' ([link](#)).



For this reason, synthetic data generation is being attracting the attention of experts in recent times, not only for its wide usage in privacy preserving environment, but also for its capability to support validation of new algorithms and applications which must be tested through data that are not available, or not accessible, due to privacy-related legal restrictions – which is especially the case in the context of processing special categories of personal data, as those relating to health.

Based on the current state-of-the-art of de-identification techniques, fully synthetic data – generated to replicate real patients’ datasets – can be considered as anonymized data even if, unlike this latter type of information, no significant differences can be spotted in the outcome of research based on the usage of synthetic data as opposed to real data.

### 8.1.3 SOLUTIONS IN MHMD

The de-identification tool developed for MHMD has been designed to apply one or multiple privacy-preserving techniques based on a number of intrinsic factors relevant to data sensitivity and consequent grade of risks.

#### A. *k*-anonymity (and *k<sup>m</sup>*-anonymity)

*k*-anonymity guarantees that every record in the de-identified dataset is undistinguishable from other *k*-1 records in the same dataset based on the indirect identifiers. In brief, this technique aims to prevent a data subject from being singled out by grouping him/her with, at least, *k* other individuals. To achieve this, the attribute values are generalized to an extent such that each individual shares the same value.


The figure below provides an example of *k*-anonymization (for *k*=4).

Consider the leftmost table: age and zipcode are quasi identifiers that can be used to re-identify a specific person in the anonymized data, while diagnosis is a sensitive personal data. The *k*-anonymization process transforms the indirect-identifiers to a form where each combination of values appears at least *k*-times.



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

k-anonymization



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

To foster security, Amnesia provides  $k^m$ -anonymity, which requires that each combination of up to  $m$  indirect-identifiers appear at least  $k$  times in the dataset.

**B. *l*-diversity**

A critical parameter when dealing with k-anonymity is the threshold of  $k$ : the higher the value of  $k$ , the stronger the privacy guarantees. A common mistake is to artificially augment the value  $k$  by reducing the considered set of quasi-identifiers, which makes it easier to build clusters of  $k$ -users due to the inherent power of identification associated to the other attributes (especially if some of them are sensitive or possess a very high entropy, as in the case of very rare attributes). Not considering all the quasi-identifiers when selecting the attribute to generalize is a critical mistake. If some attributes can be used to single out an individual in a cluster of  $k$ , then the solution fails to protect some individuals.

This issue was tackled by applying also *l*-diversity through Amnesia

L-diversity extends k-anonymity to ensure that deterministic inference attacks are no longer possible by making sure that in each equivalence class every attribute has at least  $l$  different values.

One basic goal to achieve is to limit the occurrence of equivalence classes with poor attribute variability, so that an attacker with background knowledge on a specific data subject is always left with a significant uncertainty.

L-diversity is useful to protect data against inference attacks when the values of attributes are well distributed, meaning that leakages of information are not prevented if

the attributes within a partition are unevenly distributed or belong to a small range of values or semantic meanings.

Anonymization irrevocably changes the data, but it is such to ensure that, even considering all means reasonably usable by a malicious third party, Patients and APP Users cannot be re-identified. This makes data anonymization suitable for making large datasets available to a wider audience.

An important property of data anonymization techniques is that the strength of the privacy guarantee is parametric: by changing the de-identification parameter, e.g.  $k$  in  $k$ -anonymity, data with different trade-offs between data quality and levels of privacy protection may be created. This allows to define policies for data sharing based on the degree of trust the controllers have in the Stakeholders (e.g. if the data are open to a closed group of experts, a low-level privacy guarantee might be sufficient; while if they are made available to an undetermined number of people, a higher level of privacy is required).

### **C. Secure Multiparty Computation**

Secure Multiparty Computation is a subfield of cryptography which provides the ability to compute values of interest from multiple encrypted data sources without any of the parties involved having to reveal its private data.

Through SMPC protocol, controllers enter the data which are split into separate pieces and masked with other random numbers; the encoded data pieces are then sent to multiple servers, enforcing data privacy and allowing organizations to work together without ever knowing one another's confidential datasets.

Computation is secure if, at the end of the process, no party knows anything except its own input and the final result.<sup>46</sup>

To leverage these guarantees, an additional privacy-preserving layer based on SMPC has been developed with the aim to efficiently support MHMD data analysis requirements. This layer will also be combined with other security techniques such as differential privacy, to ensure multi-level protection of all Datasets.

---

<sup>46</sup> For instance, a group of persons decide to calculate their average salary, without involving any third party to do it and without any of these persons being willing to reveal his/her salary to the others. The first person picks a random element, which is known only by him, to modify his salary – imagine that he adds the random number 2.016.679 to his salary, and then shares the result of that addition with the person next to him. The next person then adds his salary to the first result, and passes the result of that computation on to the third person, and so on, until all the salaries have been summed and returned to the first person. The intermediate results are passed securely from the previous person to the next one, and only the final sum is returned to the first person, who may then subtract the random element 2016679 from the total and thus calculate the average salary of the group without knowing the salary of any one individual.

#### **D. Differential privacy**

Applying secure computation techniques to carry out various data analysis tasks do not ensure, alone, that the relevant results do not contain any identifiable information which may allow to trace back specific individuals and even the release of purely aggregated statistics (e.g. the output of a machine learning model) might risk, under certain circumstances, compromising individual privacy.

Differential privacy falls within the family of randomization techniques, with a different approach: while, in fact, noise insertion comes into play beforehand, when dataset is supposed to be published, differential privacy can be used when the data controller generates anonymized views of a dataset – typically generated through a subset of queries for a specific third party – whilst retaining a copy of the original data.

In other words, differential privacy suggests the controller how much noise must be added, and in which form, to achieve the necessary privacy guarantees.<sup>47</sup>

One major benefit of differential privacy lies in the fact that datasets are provided to authorized third parties in response to specific queries rather than through the *una tantum* release of a single dataset.

#### **E. Federated learning**

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development.

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian inference more popular than ever: growing volumes and varieties of available data, cheaper and more powerful computational processing and affordable data storage.

In general, centralized machine learning is far from being perfect. Indeed, training the models requires companies to amass mountains of relevant data to central servers or data

---

<sup>47</sup> A key strength of DP, compared to combinatorial techniques (e.g. k-anonymity, l-diversity, etc.), is that it provides a formal mathematical framework to reason about and quantify privacy.

centers. This implies that all data must be transferred to a central entity and that data providers fully trust said entity (the 'simple' task of gathering all the data is always expensive and time-consuming).

Moreover, a significant amount of valuable training data is created on hardware at the edges of slow and unreliable networks, such as smartphones or equipment in industrial facilities.

Distributed learning achieves the opposite result, by enabling a collective model to be constructed from data that are scattered across data providers (e.g. mobile APP users, clinical institutions, etc).

Algorithm training moves to the edge of the networks, so that data never leaves the device, whether it's a mobile phone or a Hospital's data center. Once the model learns from the data, the results are uploaded and aggregated with updates from all the other devices and the improved model is then shared with the entire network.

Therefore, distributed learning takes into effect the approach of 'bringing the code to the data, instead of the data to the code' advocated by the EU Commission in the "*Guidance on sharing private sector data in the European data economy*".<sup>48</sup>

This kind of training protocol takes place in three phases:

- i. selection: devices reports the server their availability for training a federated learning task. In more detail, periodically, devices that meet the eligibility criteria check in to the server by opening a bidirectional stream used to track liveness and orchestrate multi-step communication. The server then selects a subset of connected devices based on certain goals, like the optimal number of participating devices;
- ii. configuration: the server sends the distributed learning plan to selected (for training or not) devices which meet certain predefined conditions, along with various hyperparameters and instructions for batching;

---

<sup>48</sup> Said Guidance accompanies the Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the Regions "*Towards a common European data space*" (COM(2018) 232 final): «*Algorithm-to-the-data: Bringing the algorithm to the data can be a solution to the security, data protection and privacy challenges of data. It would respect one of the main considerations for ensuring protection of personal data and privacy, which is to move data as little as possible. Using this solution means that the algorithm is installed within the IT environment of the private company and the analysis takes place there. Only the anonymous insights derived by the algorithm are transferred back to the public sector body. The data query interface and analytics possibilities could be co-designed by the company and/or the public organisation in question (or by a trusted intermediary)*» ([link](#)).

- iii. *reporting*: the server first aggregates the updates using the federated averaging algorithm and then modifies the global model which is then used in the next round (if enough devices report the needed updates in time, the round will be successfully completed and the server will update its global model, otherwise the round is abandoned).

The entire process is repeated until a specific termination criterion is met.

The most common form of distributed learning to date is federated learning ('FL'), which integrates all the phases described above, with the only and main difference that the reporting step additionally fulfills differential privacy.<sup>49</sup>

One of the main objectives (or challenges) for FL is to preserve the privacy associated with data. It appears that even when the raw data are not exposed, the repeated model weight updates can be exploited to reveal properties not global to the data but specific to individual contributors (this inference can be performed on both the server-side and the client-side). This is why differential privacy is leveraged to mitigate the relevant risks.

The updates are typically gradients which are clipped, noise added and aggregated over many participants, so to make very hard to draw any conclusion about the data samples that have been used to produce these updates. Furthermore, aggregation can be performed with SMPC.

On the other side, model is typically a machine-executable program code signed by the trusted server to guarantee authenticity. In MHMD, as an additional security layer, an untrusted black box model is deployed, whereby the server receives the model from an untrusted third party and cannot verify whether the program code in the black box is benign, thus performing machine learning.

The advantages of FL are tangible:

- ✓ data providers (Hospitals and Users) have no need to move their Datasets from the repositories where they currently reside;
- ✓ data providers are aware of each data access, keeping full control over their data;
- ✓ costs for data curation and enforcement of individual rights are reduced.

---

<sup>49</sup> *Towards federated learning at scale: system design*, Bonawitz et al., SysML 2019 ([link](#)).

## F. Watermarking techniques for data publishing

In order to help controllers managing the risks which may stem from data breach events (namely accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data), technical measures were developed and adopted allowing the inclusion of unique and unrecoverable fingerprints in anonymized datasets, without overly altering the information content, so to keep a seamless trace of each data flow.

### 8.1.4 MULTI-LAYERED PRIVACY-PRESERVING TECHNIQUES

The most suitable combination of the different techniques listed above is calibrated and then applied by MHMD dedicated de-identification engine to achieve data minimization and prevent – or at least minimize – the risks that are automatically identified on the basis of the degree of sensitivity of the Dataset that has from time to time to be shared or computed.

This allows mixing and stratifying, among others,  $k$ -anonymity, differential privacy, SMPC and homomorphic encryption, thus achieving both computational and output privacy for distributed data processing (e.g. the amount of noise required to ensure differential privacy guarantees can be reduced if noise is added to a  $k$ -anonymous version of the dataset, so long as  $k$ -anonymity is reached through a specially designed micro-aggregation of all attributes) and ensuring high levels of protection for all Patients' and Users' personal data.

In sum, following are the main measures that have been developed for securing the Datasets under MHMD:

1. a privacy-preserving data publication engine implementing privacy-by-design analytics and data anonymization procedures (MHMD Catalogue);
2. an automated differential privacy adaptive interface, capable of triggering the adoption of the most appropriate privacy preserving and anonymization method;
3. applying watermarks and fingerprints to each Dataset, providing solutions for proper provenance tracking and versioning of evolving data sources for data subset identification and citation;
4. assigning a unique Persistent Identifier (PID) to each Dataset;
5. making use of PIDs on MHMD blockchain ledger, providing a second level of anonymization and data replication services, physically deployed over the network of the Hospitals;

6. identifying all users in the system and mapping them to anonymous blockchain accounts;
7. providing blockchain mining service, API, Data Catalogue (PID indexing) and core libraries.

This innovative technical and legal framework allows to leverage securely anonymised or encrypted data for advanced data analytics and patient-specific model-based prediction applications, by *inter alia*:

- a. enabling the automated retrieval of clinical similarities and clinical annotations from the distributed database;
- b. estimating clinical risks making use of personalized physiological modelling, and more specifically demonstrating the feasibility of patient-specific modelling on securely anonymized data in order to predict the effects of treatments on patients;
- c. allowing Stakeholders to visualize data, explore patient graphs and perform patient stratification, while training a deep learning network on the data;
- d. making it possible to run analytics, enabling knowledge discovery and similarity analysis, on anonymized and encrypted data;
- e. allowing to automatically estimate the sensitiveness degree and consequent risk scale of a given Dataset, by then making available only personal data which have already undergone de-identification measures needed to ensure the lawfulness of the processing operations envisaged and, more generally, compliance with applicable laws.

In light of the above, MHMD system proves to be secure, interoperable, accountable, traceable, trustable, resilient, scalable, distributed, non-repudiable, transparent and unlinkable.

In parallel, dynamic consent functionalities make individual consent policies cryptographically bound and the resulting data access and usage restrictions self-enforceable and controlled by a hierarchy of semantically defined policies, with managed control of precedence and conflict resolution, enabled through smart contracts and made tamper-proof and auditable thanks to the blockchain-based transaction oversight.

## 8.2 INFRASTRUCTURE SECURITY

A solid security infrastructure was designed, including an *ad-hoc* distributed intrusion detection framework, to protect the entire MHMD system from any risks and cyber-attacks.

All reasonable state-of-the-art threats were considered, in conjunction with well-known defense systems, in order to identify the most significant risks and the most appropriate measures to fix them, it being understood that each Stakeholder that will implement the Platform shall remain free, as autonomous data controller in connection with its own research purposes, to adopt additional safeguards based on its peculiar security needs, also in the light of the outcomes of the DPIA that each Stakeholder will be strictly required to carry out on its own.

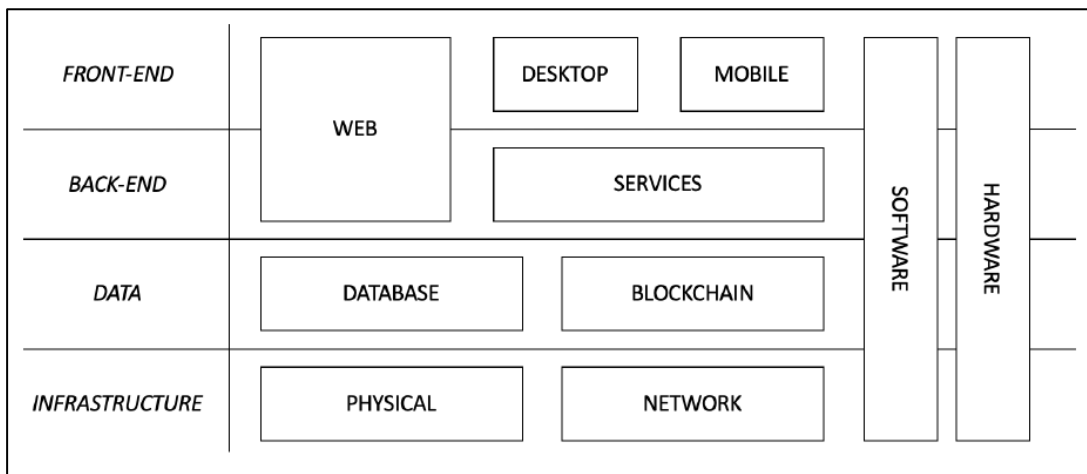
In more detail, the following elements were investigated in relation to MHMD security infrastructure ('MSI'):

- *physical*: concerning components protection from a physical point of view, for instance, safeguarding the servers and rooms where hardware components are hosted, or restricting the relevant access only to duly-authorized personnel;
- *network*: related to protection of the interactions between the various components of the Platform to avoid abuses from malicious users, both internal and external;
- *database*: aimed to minimize the risks of accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to the personal data collected and processed in connection with the Project, as well as to guarantee the integrity and availability of the system;
- *blockchain*: ensuring protection of the blockchain component from any use for malicious purposes, for instance by the injection of malicious software/smart contracts/transactions;
- *services*: regarding the protection of back-end services which represent the link between front-end components – directly accessible by Stakeholders and Users – and storage information or other components;
- *web*: guaranteeing the protection of web services and applications (web hosts) providing direct communication with the Platform user and access to the data



- *mobile*: covering different aspects relating to mobile security, such as the access to the Platform from remote, or through BYOD, or mobile applications interacting with the MHMD system and PDA;
- *desktop*: concerning the protection of various components such as the workstations placed inside of the premises of a Stakeholder, or desktop applications interacting with the MHMD system.

These components can be grouped into four main categories, as illustrated below: (i) infrastructure security; (ii) data security; (iii) back-end security; (iv) frond-end security.



Taking into account the nature, scope, context and purposes of the processing operations carried out within MHMD and that cyber-risks scenario is going through a ceaseless and daily evolution, state-of-the-art technologies were evaluated to implement the most appropriate infrastructure security framework, including a distributed ‘Intrusion Detection System’.

Attention was focused particularly on the following profiles:

**A. Network and communication security**

IT networks are usually made up of different components, such as servers, workstations, virtual hosts, accessory, IoT and mobile devices, but also network firewalls, switches and accelerators, wired, wireless, VPN, DMZ and VLAN networks, honey nets, etc.

In this context, there are essentially two kind of attacks to consider (resulting, for instance, in denial of service, malicious software, packet forging and replay attacks, covert channels):

- well-known/exploit-based attacks, often characterized by a Common Vulnerabilities and Exposures number (CVE) and related to an available exploit code;
- novel/0-day attacks.

Other kind of attacks can be executed for specific purposes, such as phishing (often involved to inject malicious software on a host) or spam. Nevertheless, given the context and nature of the MHMD Platform, this type of social engineering threats are likely to pose minor risks compared to the network attacks described above.

## **B. Blockchain and transactions security**

Regarding blockchain network and components security, an important and well-known attack is majority attack. In particular, in this case, if the attacker controls more than 51% of the nodes (or the resources, such as computing power, in function of the mining algorithm adopted), it is possible to mine blocks quicker, hence having the authority to discard/accept specific blocks and including fake transactions. Although this vulnerability mainly affects public blockchains, it may become relevant even for private ones, since mining hosts may be targeted by cyber-criminals that may take control of the entire network and pass unobserved even for a very long time.

Also distributed denial of service (DDoS) attacks must be considered, such as volumetric DDoS executed by overwhelming network resources of the targeted nodes, aimed to dismantle the entire network or parts of it. Similar attacks are carried out by executing hijacking activities to network traffic interception, by exploiting the Border Gateway Protocol (BGP) routing protocol, e.g. to divert the traffic to a node under the control of the attacker.

With reference to the security of blockchain nodes, eclipse attack may allow an attacker to take control of incoming and outgoing connections of a given target, thus potentially isolating the victim from other peers and manipulating its view of the blockchain.

Finally, focusing on blockchain users' security, the main risk is that an attacker installs a malicious software (in the host) able to target and leak the private encryption keys adopted by the users, allowing the attacker to impersonate the victim

Furthermore, the phenomenon known as 'criminal smart contracts' (CSC) must be properly addressed, referring to the injection of smart contracts specifically designed to leak personal data or to trigger real-world crimes.<sup>50</sup>

### C. Web services and applications security

The most important profile regards cloud security, which may require different solutions depending on the specific solution at stake, such as infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS) and software-as-a-service (SaaS). Although the MHMD Platform can in practice be considered as a cloud-based platform, since most of its services are web based, also web applications security plays a crucial role.

In this respect, the major threats may be classified in three different categories:

- a) attacks to web forms, strictly related to phishing, exploiting web forms to retrieve users' personal datasets;
- b) Structured Query Language (SQL) attacks, which mainly take place in the form of SQL injection attacks (SQLi), aimed at entering strings leading to the execution of illegal queries on the underlying database systems;
- a) cross-site scripting (XSS), concerning the injection of client-side scripts into web pages accessed by other users.



<sup>50</sup> As an example, in 2016 cyber-criminals targeted the DAO smart contract, exploiting a recursive call vulnerability to stole crypto-currencies.

#### D. Database security

Database protection is strictly linked to SQL injection threats (point b) of let. C. above), which may be prevented particularly by implementing an Intrusion Detection System that works by logging the activities of intruders performing SQL injection attack.

Additional security measures may include the adoption of:

- database encryption;
- (multi-factor) database authentication, as well as of *ad-hoc* authentication policies and management rules;
- watermarking techniques to manage data access and integrity;
- storage of auditing records from database management systems (DBMS) logs directories in order to identify suspicious activities.

#### E. Other security aspects

Finally, innovative 0-day threats must be addressed (a zero-day vulnerability is a software security flaw that does not have – yet – a security patch in place because developers are oblivious to the threat and which can hence be easily exploited by cybercriminals), particularly as follows:

- carrying out continuous security updates in order to protect the system from novel threats and vulnerabilities;
- being impossible to implement a complete protection system, since any system is potentially vulnerable to 0-day threats, the adoption of an anomaly detection systems helps timely identifying unknown behaviours, potentially related to novel cyber-attacks.

### 8.2.1 MHMD INFRASTRUCTURE SECURITY

The MSI was designed and implemented taking into account three main factors: (i) the peculiarities of the MHMD platform, as well as its components and their interactions, (ii) state-of-the-art solutions in the security field (as summarized above from A to E) and (iii) the best practices set by the most important standards providers in the cyber-security context, such as the *International Standards Organization* (with particular reference to

ISO/IEC 27000-series, ISO/IEC 15408 and ISO/IEC 18045), the *National Institute of Standards and Technology* (NIST), the *European Union Agency for Network and Information Security* (ENISA) and the *Open Web Application Security Project* (OWASP).

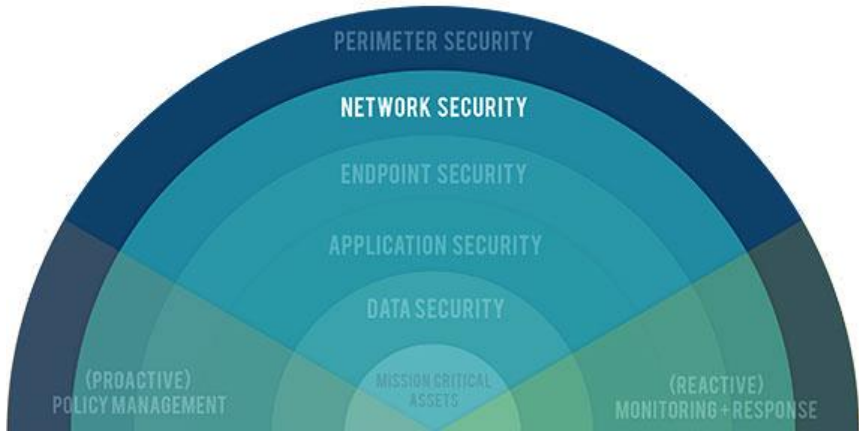
In light of the above, a number of safeguards have been put in place to protect the Project’s infrastructure, illustrated in the table below:

COMPONENT	STATE-OF-THE-ART MAJOR THREATS	COUNTERMEASURES ADOPTED
<p><b>1. Encryption</b></p>	<p>Data confidentiality breaking</p>	<ul style="list-style-type: none"> <li>▪ All communications between the infrastructure components are encrypted;</li> <li>▪ End-to-end encryption is adopted for communications;</li> <li>▪ All data stored by the components are encrypted;</li> <li>▪ Encryption/decryption keys are bound to a specific entity (e.g. dataset, user, etc.);</li> <li>▪ Encryption/decryption keys are stored in memory for very short periods;</li> <li>▪ Adopted encryption algorithms do not include unreliable algorithms (such as DES and RC4);</li> <li>▪ The adoption of “custom” encryption methods is accurately evaluated, from the security point of view (e.g. encryption of a portion of a file, double encryption using different algorithms, etc.)</li> </ul>
<p><b>2. Communications and network</b></p>	<p>Cyber-attacks like denial of service, malicious software, packet forging and replay, covert channels.</p>	<ul style="list-style-type: none"> <li>▪ Component’s hosts are placed in a dedicated and segmented network separated from other nodes (e.g. workstations, other server systems, etc.);</li> <li>▪ Component communications are protected by a network firewall (e.g. allowed ports, network limits, filtering, IDS/IPS, etc.);</li> <li>▪ Components connections adopt secure encrypted protocols (see point 1 above, under ‘Encryption’);</li> <li>▪ Adoption of user authentication methods, combining both server-side and client-side authentication methods;</li> <li>▪ Adoption of strong authentication credentials;</li> <li>▪ Adoption of certificate pinning methodologies;</li> </ul>

		<ul style="list-style-type: none"> <li>▪ Anonymizing and/or tokenization servers are isolated from the network, except the (single) node they are supposed to communicate with;</li> <li>▪ Network communications are protected by a network firewall;</li> <li>▪ Network policy and access control rules are in place;</li> <li>▪ A network Intrusion Detection and Prevention System is deployed on the network (see below);</li> <li>▪ Critical nodes are replicated on the network, in order to improve availability.</li> </ul>
<b>3. Blockchain and transactions</b>	Network attacks to the system (e.g. denial of service), exploitation/injection of malicious transactions	<ul style="list-style-type: none"> <li>▪ Protection of the blockchain network (see point 2 above);</li> <li>▪ Adoption of secure network designing approaches (e.g. to counter majority attacks);</li> <li>▪ Adoption of secure smart contracts development approaches (security by design, identification of code vulnerabilities and bugs, testing, etc.).</li> </ul>
<b>4. Services</b>	Exploitation of insecure communications, SQL injection, cross-site scripting	<ul style="list-style-type: none"> <li>▪ Services protection at the network level (see point 2 above);</li> <li>▪ Adoption of host or network tools/modules/approaches to properly address well-known attacks (e.g. instance, SQLi, XSS, etc.).</li> </ul>
<b>5. Mobile applications</b>	Exploitation of software bugs	<ul style="list-style-type: none"> <li>▪ Mobile application security is strictly dependent on services and smart contracts security (see. Points 3 and 4 above);</li> <li>▪ Secure software development;</li> <li>▪ Deep testing activities;</li> <li>▪ User access and communication restrictions;</li> <li>▪ Adoption of strong and secure communications and strong data encryption;</li> <li>▪ Accurate evaluation of external interfaces (like intents or content providers).</li> </ul>
<b>6. Database</b>	SQL injection, accidental or unlawful destruction,	<ul style="list-style-type: none"> <li>▪ Adoption of access control and user restriction methods (principle of the least privilege);</li> <li>▪ Adoption of strong user credentials;</li> </ul>

	<p>loss, alteration, unauthorized disclosure of or access to personal data</p>	<ul style="list-style-type: none"> <li>▪ Data encryption;</li> <li>▪ Prevention of access to original datasets after de-identification is applied;</li> <li>▪ Adoption of a database Intrusion Detection System;</li> <li>▪ Implementation of data integrity procedures.</li> </ul>
<p><b>7. Software</b></p>	<p>Exploiting vulnerable software components and vulnerabilities, privilege escalation attacks.</p>	<ul style="list-style-type: none"> <li>▪ Adoption of secure software development approaches (security by design, identification of code vulnerabilities and bugs, testing, etc.);</li> <li>▪ Adoption of third-party well-known software/software libraries/modules/code snippets (if any), implemented by trusted developers;</li> <li>▪ Adoption of code obfuscation techniques;</li> <li>▪ Avoidance of sensitive debugging logs printed in output and available to the user, even if through dedicated consoles;</li> <li>▪ Adoption of a security by design approach to design the workflows implemented in the Platform.</li> </ul>
<p><b>8. Hosting and organization</b></p>	<p>Malicious software (e.g. aimed at gaining administrator privileges), phishing, majority attack,</p>	<ul style="list-style-type: none"> <li>▪ An operating system with updated security modules is adopted;</li> <li>▪ The host component is not running useless services (e.g. not needed web service);</li> <li>▪ The host component is not running other services external to the Project;</li> <li>▪ Users and administrator accounts are protected by strong passwords;</li> <li>▪ No sensitive password or connection data are stored on the system;</li> <li>▪ Remote connection services (e.g. remote desktop, remote shell, file sharing, etc.) are not running, or their network access is restricted to the only nodes connecting to them;</li> <li>▪ Network logs are collected and maintained.</li> </ul>
<p><b>9. Continuous security</b></p>	<p>All the above</p>	<ul style="list-style-type: none"> <li>▪ Encryption keys are replaced, in case of a data breach;</li> <li>▪ Stronger encryption measures are implemented;</li> </ul>

		<ul style="list-style-type: none"> <li>Seamless security updates are carried out on the system;</li> <li>Security issues and data breaches are promptly reported to controllers, to competent Data Protection Authorities and, where applicable, to the data subjects;</li> <li>Backup procedures are in place;</li> <li>Network security controls are periodically executed;</li> <li>Network traffic monitoring activities are periodically carried out.</li> </ul>
--	--	---



**8.2.2 MHMD DISTRIBUTED INTRUSION DETECTION SYSTEM**

In order to strengthen the overall security of the system, a specific distributed Intrusion Detection System (**'dIDS'**) was designed and deployed which is able to detect intrusions on the Platform and, thus, to prevent exploitation of weaknesses by malicious users.

The MHMD dIDS is made up of two components integrated into the system and transparently analyzing the network traffic on sensitive locations (e.g. by physically interrupting and forwarding the network traffic, or by working on mirrored traffic in a passive and less invasive way):

- ✓ MHMD-dIDS Collector: it is the main component of the MHMD-dIDS, aimed to provide a graphical interface to the user. This component is supposed to collect

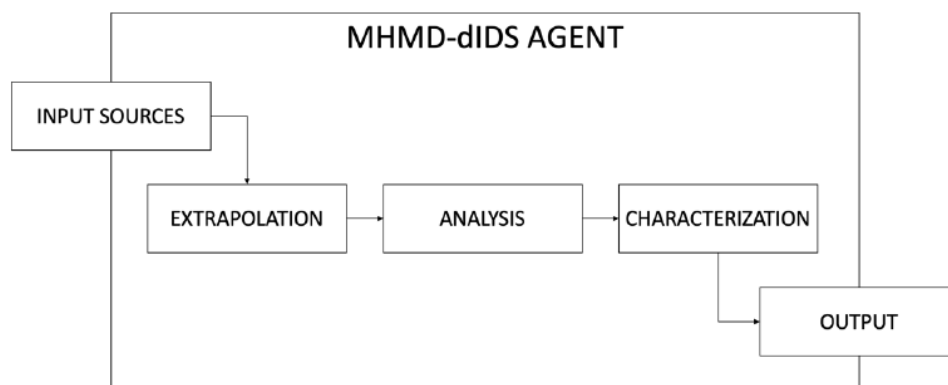


alert data from the secondary components distributed throughout the Platform, in order to show anomaly-triggered messages to the user;

- ✓ *MHMD-dIDS Agents*: such components are replicated within the infrastructure of each data provider (Hospitals and APP Users) and execute the following activities: (i) communication with the internal (distributed) modules and/or network components of the system; (ii) run the (local) Intrusion Detection System on the relevant Datasets; (iii) send relative reports (alerts, mainly) to the MHMD-dIDS Collector component.

The dIDS operates thanks to data computation executed directly inside the data providers' systems, thus guaranteeing privacy because no data is ever pulled out of the local repositories. The intrusion detection activities carried out by the MHMD-dIDS Agent involve (as illustrated below):

- a. the extrapolation and retrieval of data by adopting specific pre-defined metrics;
- b. data analysis, by elaborating the retrieved data and comparing the context with an analogous one related to a no-anomaly (*i.e.* lawful) situation;
- c. the characterization of the current situation as legitimate or anomalous, by adopting specific thresholds.



The adopted architecture allows to accomplish local replication of MHMD-dIDS Agents inside the same data provider's organization, thus being able to monitor different data sources and making MHMD-dIDS a distributed and multi-contextual tool.