This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare

Deliverable 8.7
"Report on Data Protection Impact Assessment"

Work Package 8
"Testing and evaluation in clinical translation"

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Document information

Grant Agreement Number: 826078			**Acronym: FeatureCloud**	
Full title	Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare			
Topic	Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 January 2019	**Duration**	60 months	
Project URL	https://featurecloud.eu/			
EU Project Officer	Christos Maramis, Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Jan Baumbach, University of Hamburg (UHAM)			
Deliverable	D8.7 - Report on Data Protection Impact Assessment			
Work Package	WP8 - Testing and evaluation in clinical translation			
Date of Delivery	**Contractual**	31/12/2023	**Actual**	19/01/2024
Nature	Report	**Dissemination Level**	DoA: Confidential, now: Public	
Lead Beneficiary	RI			
Responsible Author(s)	Walter Hötzendorfer (RI), Jan Hospes (RI), Christof Tschohl (RI)			
Keywords	DPIA, data protection impact assessment, GDPR, safety, security			

History of changes

Version	Date	Contributions	Contributors (name and institution)
V0.1	09/03/2023	First draft structure	Walter Hötzendorfer (RI), Jan Hospes (RI), Christof Tschohl (RI)
V0.2	10/08/2023	Draft version Scope / stakeholders / legal admission / PTA	Walter Hötzendorfer (RI), Jan Hospes (RI), Christof Tschohl (RI)
V0.3	20/10/2023	Draft Risk Analysis	Walter Hötzendorfer (RI), Jan Hospes (RI), Christof Tschohl (RI)
V0.4	12/09/2023	Input form Risk Analysis Workshop	Mohammad Bakhtiari (UHAM Walid Fdhila (SBA) Rudolf Mayer (SBA) Bela Bihari (GND) Sandor Fejer (GND) Balazs Orban (GND)
V0.5	09/10/2023	Draft AI Act	Tuende Fueloep (RI), David Schneeberger (RI)
V0.6	20/10/2023	Risk Analysis	Walter Hötzendorfer (RI), Jan Hospes (RI), Christof Tschohl (RI)
V0.7	30/11/2023	writing of the deliverable report	Walter Hötzendorfer (RI), Jan Hospes (RI), Christof Tschohl (RI)
V0.8	15/01/2024	Completion of internal review and quality control	Nina Donner (concentris)
V0.9	18/01/2024	Final version	Walter Hötzendorfer (RI)
V1.0	19/01/2024	Final edits, approval by the project coordinator, and submission to EC	Nina Donner (concentris), Jan Baumbach (UHAM)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Table of acronyms and definitions

AI	Artificial Intelligence
API	Application Programming Interface
CFR	Charter of Fundamental Rights
CLI	Call Level Interface
CNIL	Commission nationale de l'informatique et des libertés (French data protection authority)
concentris	concentris research management GmbH
CVSS	Common Vulnerability Scoring System
DMOP	Data Mining Output Privacy
DoA	Description of Action (of the FeatureCloud Horizon2020-funded project)
DP	Differential Privacy
DPIA	Data Protection Impact Assessment
DPO	Data protection officer
DPSGD	differentially private stochastic gradient descent
DuD	*Datenschutz und Datensicherheit* (German for Data protection and data security)
EC	European Commission
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EP	European Parliament
FC	FeatureCloud
FL	Federated Learning
GANs	generative adversarial networks
GDPR	General Data Protection Regulation
GUI	Graphical user interface
eDPSGD	extended DPSGD
GND	Gnome Design SRL
GUI	Graphical user interface

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

HE	Homomorphic encryption
ICDPPC	International Conference of Data Protection and Privacy Commissioners
ICO	Information Commissioner's Office (UK data protection authority)
IEEE	Institute of Electrical and Electronics Engineers
IVDR	In Vitro Diagnostic Medical Device Regulation (Regulation (EU) 2017/746)
KPI	Key Performance Indicator
MDR	Medical Device Regulation (Regulation (EU) 2017/745)
ML	Machine Learning
MS	Milestone
MUG	Medizinische Universität Graz
NISD	EU Network and Information Security directive
OECD	Organisation for Economic Co-operation and Development
PAML	Privacy-Aware-Machine-Learning
Patients	In this deliverable, we use the term “patients” for all research subjects. In Feature-Cloud, we will focus on patients, as this is already the most vulnerable case scenario and this is where most primary data is available to us. Admittedly, some research subjects participate in clinical trials but not as patients but as healthy individuals, usually on a voluntary basis and are therefore not dependent on the physicians who care for them. Thus to increase readability, we simply refer to them as “patients”.
PII	personally identifiable information
PPDM	Privacy-preserving Data Mining
PPDP	Privacy-preserving Data Publishing
RI	Research Institute AG & Co KG
SBA	SBA Research Gemeinnützige GmbH
SDU	Syddansk Universitet
SDV	Synthetic Data Vault
SMPC	Secure multi-party computation
TFEU	Treaty on the Functioning of the European Union
TUM	Technische Universität Muenchen
UHAM	University of Hamburg

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

UM	Universiteit Maastricht
UMR	Philipps Universität Marburg
WP	Work package

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

1 Objectives of the deliverable based on the Description of Action (DoA)

The main objectives of the deliverable are (1) to provide guidance to partners on how to conduct a Data Protection Impact Assessment (DPIA) as part of a comprehensive risk management, (2) to conduct a DPIA in the applicable scope of the deliverable and (3) to thereby ensure the protection of the fundamental rights and freedoms of natural persons affected by the FeatureCloud data processing activities. Consequently, the deliverable seeks to guarantee that all partners in the project are able to implement state-of-the-art (technical and organisational) measures for mitigating potential risks. In doing so, the present framework particularly promotes the approach of 'data protection by design and by default'.

The main objective of WP8 is to evaluate the applicability of the FeatureCloud platform in a real-world setting, and to successfully validate it on clinical data to ensure a wide acceptance beyond the research community and to convince clinicians, clinical scientists, and patients to use the FeatureCloud platform, this WP will make sure that it can be used as intuitively as possible while providing all relevant regulatory, scientific and computational features and guaranteeing data safety in an unprecedented manner.

Task 6: Data protection impact assessment (RI, UHAM)

In alignment with and based on the results of the security and privacy risk assessment and the GDPR compliance and ethical evaluation in WP2, WP6 and D10.1 and D1.3, RI will carry out a Data Protection Impact Assessment in accordance with the provisions of Art 35 GDPR. This includes an analysis of data protection law in scientific research and medical research in particular and of the admissibility of data processing activities in the FeatureCloud platform. The results will clearly delineate the boundaries between the platform and the applications, state the responsibilities about privacy measures on the components and modules of the FeatureCloud platform and its applications, and include recommendations, guidelines and policies for FeatureCloud application developers to guarantee that applications developed and deployed on top of FeatureCloud will produce the expected privacy guarantees.

2 Executive Summary

Deliverable D8.7 of the FeatureCloud project compiles a lot of knowledge developed throughout the project in the form of a DPIA report compliant with Article 35 GDPR. It contains, among other things, a comprehensive description of the FeatureCloud system, an analysis with regard to data protection law and the proposed AI Act, a comprehensive risk analysis, including the identified risk mitigation measures and, in Annex I, the FeatureCloud deployment manual comprising those risk mitigation measures and recommendations which by their nature go beyond what was possible to implement already during the development stage following a privacy-by-design approach but can only be put into practice by the respective stakeholders during the different phases of an actual study.

This report, therefore, is able to serve two major purposes: Different stakeholders can use it as a comprehensive guidance document for deploying and using FeatureCloud after the official conclusion of the FeatureCloud H2020 research project. In addition, those stakeholders who are obliged to carry out a DPIA in the context of applying FeatureCloud can use this report as a basis for a DPIA report compliant with Article 35 GDPR. It already contains everything, which could be collected and analysed until the end of the FeatureCloud project, i.e. end of 2023, that is necessary for conducting a DPIA for the actual use of FeatureCloud in a particular use case. By extending it in relation to the circumstances of the individual case a well-founded DPIA for the specific use case can be conducted relatively quickly.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The key findings of the DPIA documented in this report are that federated learning actually leads to the expected privacy and security gain and that additional privacy and security risks which are not mitigated by the federated approach have been dealt with or can be dealt with appropriately. In other words, when the results of the FeatureCloud project are applied properly, no unmitigated high risks remain: the identified risk mitigation measures, which are either already implemented in FeatureCloud as far as this has been possible by their nature or are otherwise included in the FeatureCloud deployment manual, are able to reduce all identified risks to a level below the critical threshold ("high" according to Article 35 GDPR).

3 Introduction (Challenge)

The FeatureCloud project has developed a secure, novel architecture and infrastructure based thereon for federated machine learning in the medical domain governed by a tailored system for immutable access control. FeatureCloud can be implemented in hospitals` IT infrastructure and using federated machine learning, globally distributed data can be leveraged to learn a global computational model for healthcare and medical research, but without the need to send confidential data over a communication network.

Through this innovative privacy-by-architecture approach, FeatureCloud intends to substantially reduce privacy and security risks for medical data utilised for research purposes, essentially through two key characteristics of this approach: (1) no sensitive data is communicated through any communication channels, and (2) no sensitive data is stored in an additional location and in particular not in one central point, which would be a central point of attack. The overall aim is, with these guarantees intrinsically tied to the fundamental design of the system, to lower the hurdles for acquiring more data for research purposes in order to aid in diagnostics, understand disease mechanisms or assess risk factors.

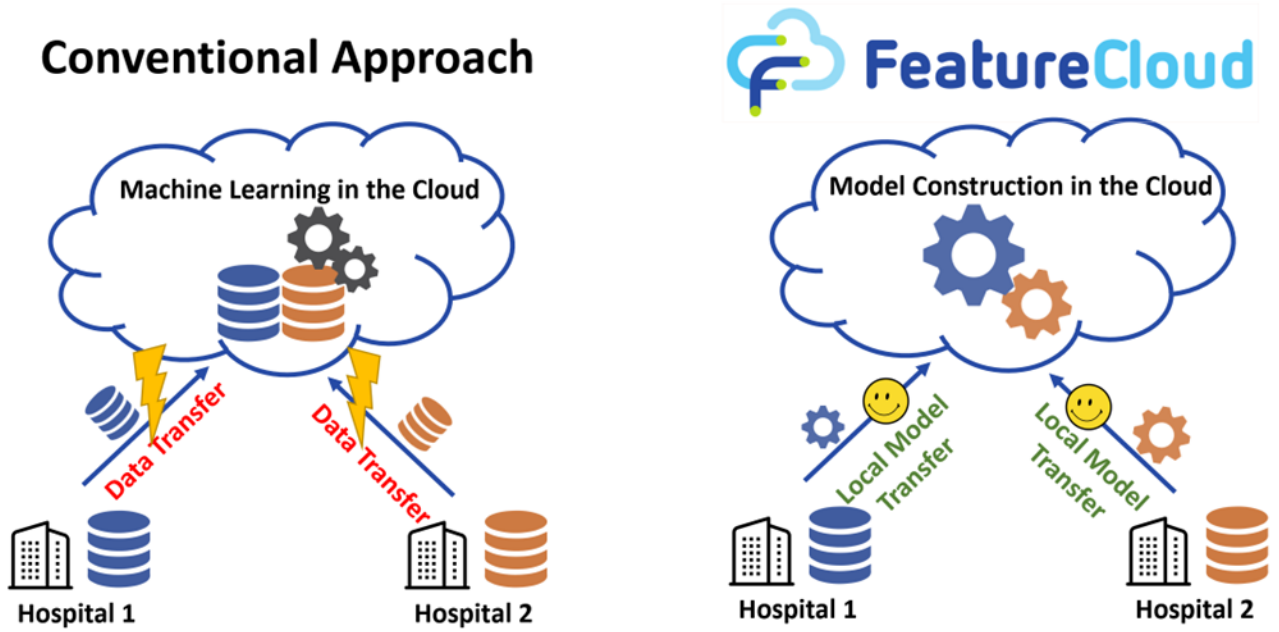

Figure 1. *Comparing schematics of conventional machine learning in the cloud (left) and federated machine learning FeatureCloud (right).*

Two of the major objectives of this deliverable are therefore (1) to investigate whether these guarantees are actually functioning as intended, i.e. that the federated approach leads to the claimed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

privacy and security gain and (2) to demonstrate that additional privacy and security risks which are not mitigated by the federated approach have been dealt with or can be dealt with appropriately.

This is documented in the form of a Data Protection Impact Assessment (DPIA) report compliant with Article 35 of the GDPR, because the literature and practice related to Article 35 of the GDPR offers a well-established body of methodology and because the future use of the system developed in the FeatureCloud project on real-world clinical data will likely require that a DPIA compliant with Article 35 of the GDPR has been carried out. According to Art 35 of the GDPR, a DPIA must be carried out if a type of processing is likely to result in a high risk to the rights and freedoms of natural persons due to the nature, scope, context and purposes of the processing. In particular, Art 35 (3) GDPR provides for a mandatory DPIA if processing on a large scale of special categories of personal data pursuant to Art 9 (1) of the GDPR takes place.

It can be assumed that the practical application of the results of the FeatureCloud research project will concern medical data of a large number of persons and the criterion of processing on a large scale is interpreted in practice as not very high. For instance, according to jurisdiction of the data protection authority in Austria, the recording of health data in a narcotics book, which contains data records of (only) about 150 patients (first name, surname, physical state of health, narcotics administered and quantity dispensed) and about 60 rescue service employees (personnel number and signature), may already constitute processing on a large scale of sensitive data.

Therefore, in many instances, future practical application of FeatureCloud will require that a DPIA compliant with Article 35 of the GDPR has been carried out. However, according to this provision, a DPIA must be carried out in relation to a specific processing activity (or several specific processing activities) of a specific controller (or several specific controllers). The specific purpose of a future practical application of FeatureCloud infrastructure and to which data it will be applied by whom depends on the individual case and can at this stage not be determined. If this is determined, the complete DPIA can be carried out. The present document, compiling a lot of knowledge developed in the FeatureCloud project, as will be described in the next section in more detail, contains everything, which could be collected and analysed until the end of the FeatureCloud project, i.e. end of 2023, that is necessary for that purpose. Aspects of the individual case, however, by their nature, can only be analysed once they can be determined. Therefore, the respective parts of this report can be considered what is sometimes called a framework DPIA, while all aspects where this was possible are already fully analysed and elaborated according to the requirements of Article 35 of the GDPR. This report can be used as a basis and extensive body of knowledge for future DPIAs in this context, which only need to extend the present report by specific aspects related to the individual case.

4 Methodology

Based on WP8 task 6, this is a report on the data protection impact assessment (DPIA) regarding the use of the FeatureCloud platform and app store developed in the course of the FeatureCloud research project, along the object of consideration described in chapter 5. With the obligation to conduct a DPIA, the legislator intends to identify and evaluate possible risks to the rights and freedoms of natural persons associated with the planned processing, in particular with regard to their cause, nature, specificity and severity, before the start of the processing. In this way, the processing can be designed from the outset in such a way that these risks are minimised as far as possible. The DPIA pursuant to Article 35 of the GDPR is closely related to Articles 24, 25 and 32 of the GDPR. In particular, the results of a DPIA are to be reflected in the privacy-by-design measures to be taken pursuant to Article 25 GDPR.

In principle, a DPIA must be carried out in relation to a specific processing activity (or several specific processing activities) of a specific controller (or several specific controllers). In the FeatureCloud project, a specific infrastructure has been developed for federated machine learning, but how in

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

particular and especially by whom it will be used is intentionally left open. Nevertheless, the use is to a certain extent already determined in advance by their architecture and functionality.

With the aim of providing an analysis that is as close to practice as possible, the present report is methodologically structured on the basis of a data protection impact assessment in accordance with Article 35 of the GDPR. With regard to the circumstances not yet determined, a framework data protection impact assessment (framework DPIA) is carried out. Before the actual use of the systems, this framework DPIA can be specified in relation to the circumstances of the individual case and thus a well-founded DPIA for the actual use case can be created relatively quickly.

Ethical considerations guide the evaluation of potential risks associated with the processing of personal data. This involves ensuring transparency and minimising data collection to only what is necessary, maintaining data accuracy, implementing robust security measures, and safeguarding against biases, discrimination, and, in the present context, wrong medical decisions.

The focal point of ethics lies in the assessment of human action, a pursuit shared with law and related disciplines. While ethics aims for universally applicable norms and rules, law pertains to a specific, factually grounded order whose norms it interprets and enforces, known as positive law. Legal norms, distinct from ethical norms, serve as codified directives for conduct, which may draw from ethical norms but are not bound to them. Ethics comes into play, particularly in legal sciences, when contemplating the justification of these orders and norms, as seen in fields like philosophy of law, legal history, and legislative theory. The convergence of law and ethics in content emerges prominently in the deliberation of fundamental principles and the rule of law. These encompass:

- The principle of legality: constraining state authority by laws (formal rule of law) and the mandate to achieve substantive justice (substantive rule of law).
- The principle of separation of powers: delineating functional and organisational separation of legislative, executive, and judicial branches.
- The principle of certainty: necessitating clarity in the content of state action (legal clarity).
- The principle of the protection of legitimate expectations: ensuring predictability and reliability in state action (legal certainty).
- The principle of proportionality: upholding a balanced relationship between the end goal (common good) and the means (restriction of freedom) in state interventions. [D10.1: 5.3]

These principles of the rule of law find expression in fundamental rights, enshrined in pivotal legal documents of nations, such as the Constitution for the Federal Republic of Germany (*Grundgesetz für die Bundesrepublik Deutschland, GG*) or in Austria, divided into the Federal Constitutional Law (*Bundes-Verfassungsgesetz, B-VG*), the Basic Law on the General Rights of Citizens of 1867 (*Staatsgrundgesetz über die Allgemeinen Rechte der Staatsbürger, StGG*), and further legislation implementing international human rights conventions. [D10.1: 5.3]

This common foundation is predominantly codified in international treaties, or, due to ongoing European integration, on a supranational level, as seen in the EU or the European Convention of Human Rights (ECHR). The contents of international human rights conventions are also entrenched as fundamental rights in constitutions or statutory laws, exemplified by the Universal Declaration of Human Rights, the European Convention on Human Rights, the Geneva Convention, and the EU Charter of Fundamental Rights. Moreover, in the Member States of the European Union, fundamental rights can be directly established by legal acts of the EU, as exemplified by the General Data Protection Regulation (GDPR). [D10.1: 5.3]

This document follows the methodological framework for conducting a Data Protection Impact Assessment (DPIA) of the data processing operations taking place in the FeatureCloud system. The following framework is based on the requirements of Article 35 of the GDPR and on common international standards and approaches in regard to impact-, risk- and technology assessment and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

adapts these methodologies to enable an adequate analysis of the AI-based data processing activities within the FeatureCloud system. Most importantly, the analysis will not be limited to data protection aspects but considers other relevant fundamental and human rights as well. Thereby we are striving to present and implement a human rights-based approach (United Nations, 2006) to assess the impact of data processing operations, which complements traditional DPIA methodologies. An important reference has to be emphasized here to Deliverables D2.1 and D1.3, which contain many details on general risk assessment methodology and partly provide the basis for the risk analysis that is documented in this deliverable. Specific and further methodological considerations for risk analysis are discussed in chapter 9.

A constant exchange with the technical partners in this project on a substantial and qualitative level was indispensable for an understanding of the potential risks coming along with the FeatureCloud system as well as for devising measures.

Providing the methodical DPIA framework presents the first step in a (challenging and yet worthwhile) process that will endure as long as the life cycle of the FeatureCloud system. As such, the assessment process is essential to ensure the legal and ethical compliance of the AI-based technology in question.

In the course of writing we closely analysed existing guidelines (issued by the OECD, High-Level Expert Group on Artificial Intelligence, Council of Europe, Amnesty International, EDPS, German Data Ethics Commission, CNIL, ICO, ICDPPC, IEEE etc.) and actively follow the European and international development of guidelines on AI and Ethics, with a special focus on aspects of AI in (bio-)medicine and health.

The present report has been written as deliverable D8.7 of the FeatureCloud project as an ongoing endeavour throughout the final phase of the project. Its content, however, has been developed in all phases of the project from the outset. The processes of conducting privacy and data protection analyses, impact assessment and risk assessment in particular started from the project's beginning and were carried out throughout the project, strongly interlinked with all relevant design and development activities in the project. All substantial results of these processes are documented in the present report, which is also meant to be a compilation of all project results in this context that have already been documented in other deliverables and publications of the project. A significant portion of the content of this document has therefore been compiled from other deliverables of the FeatureCloud project, which is indicated in square brackets in the following format: [<deliverable number>:<section number>]. A particular case is the description of the FeatureCloud architecture and implementation in the following chapter, which is spread over several deliverables created in different phases of the project and is now for the first time collected in one place and in accordance to the final state at the end of the project. In addition, the present report contains references to other FeatureCloud deliverables, in particular for further details or in order to point to specific risk mitigation measures described there.

5 Data Processing Operations in Scope

The overview of the structure and functioning of the FeatureCloud platform and app store in this chapter is based on the contents of the FeatureCloud deliverables D2.1, D2.2, D2.3, D6.1, D6.2, D6.5, D7.1, D7.2 and D10.1. Please refer to these deliverables for further details.

Machine learning depends on the availability of large amounts of data. In biomedicine, vast amounts of data exist and could aid in diagnostics, understanding disease mechanisms or assessing risk factors. However, potentially valuable data, such as electronic health records (EHR) or omics data, is often scattered across multiple facilities, rendering large-scale ML infeasible without sharing of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

personal data. In most cases, sharing of data is not an option due to privacy considerations or lack of trust.

FeatureCloud employs the approach of federated learning to utilise that treasure of scattered data for such research while still maintaining the required level of privacy. Federated learning allows to carry out machine learning on data distributed across different sources without having to copy the data into a large common database. Instead, as shown in the figure below, only data of models which are trained locally are transferred to a central instance in order to build a global model. This is achieved by pushing the appropriate algorithms to local execution platforms inside the data holders' premises, calculating the feature vectors there and subsequently pushing them back to the overall platform, possibly in an iterative process going back and forth until convergence has been reached (Yang et al. 2019). [D2.2:3.1]

The following figure, based on the simplified example of a linear regression, illustrates how such a global model could be trained using federated learning on the basis of data distributed over different hospitals.

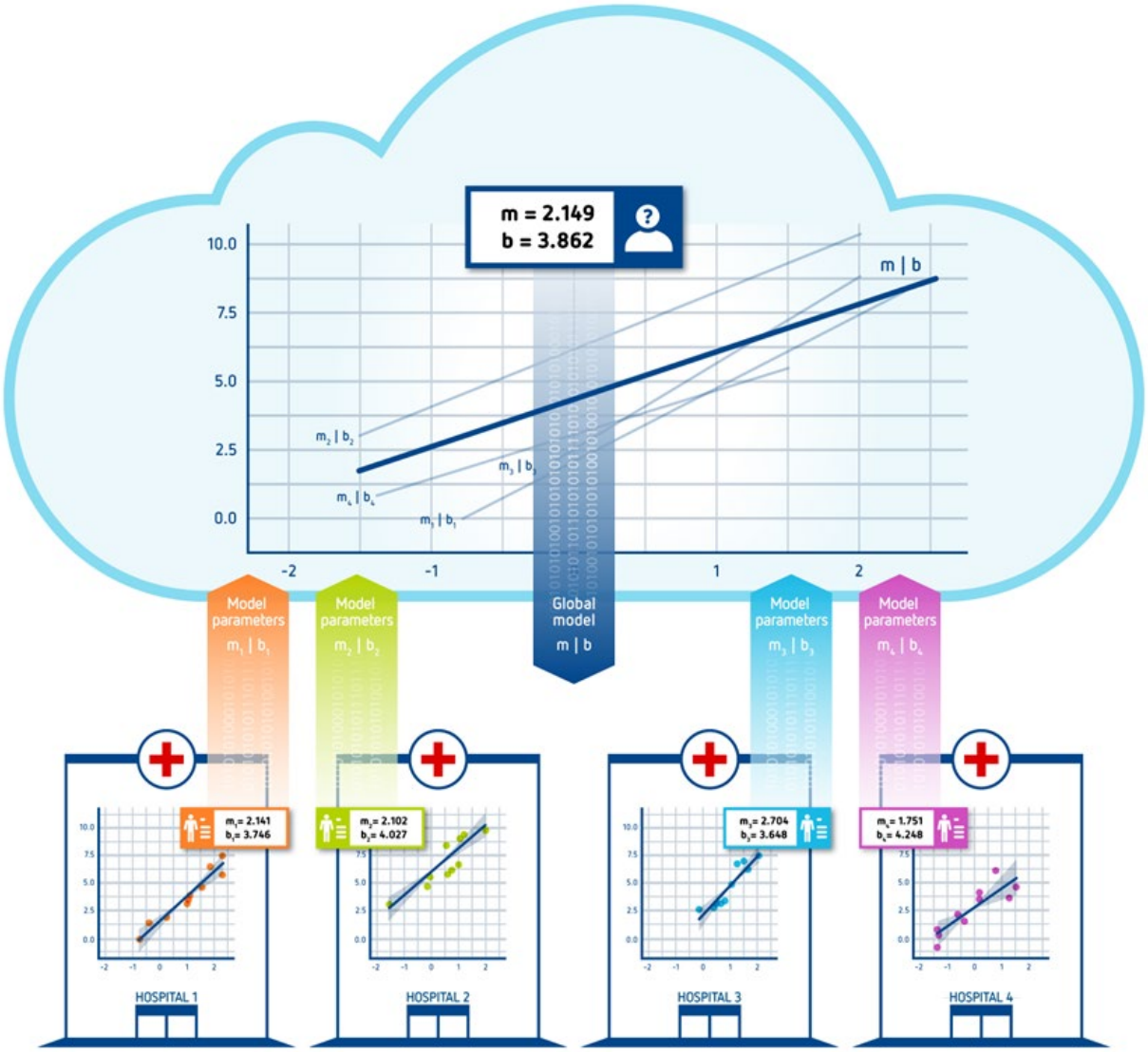

Figure 2. *Global platform schematics (simplified) where participating hospitals send local models to coordinators who computes and distributes the global model back to participants.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

In machine learning, two fundamental stages can be distinguished: (1) the actual machine learning, i.e. the training of a model based on a particular dataset and (2) the application of this model on data, i.e. the inference, e.g. to classify that data. The FeatureCloud project centres around the training stage, developing methods, infrastructure and apps for federated machine learning. The actual training of models that may be applied in medicine and even more so, the application of those models is not carried out within the FeatureCloud Project. The specificities of such training and application will be determined in the future and depend on the individual case. Nevertheless, this DPIA shall address the risks to those data subjects to whom the models will eventually be applied, but limited to the perspective of training the model.

Federated ML broadly involves two general operations, possibly alternating in multiple iterations: local optimization and global aggregation. In FeatureCloud, all running instances of a federated application (app) take one of two roles: participant and coordinator, performing the respective federated operation. FeatureCloud expects precisely one coordinator and an arbitrary number of participants, leading to a star-based architecture.

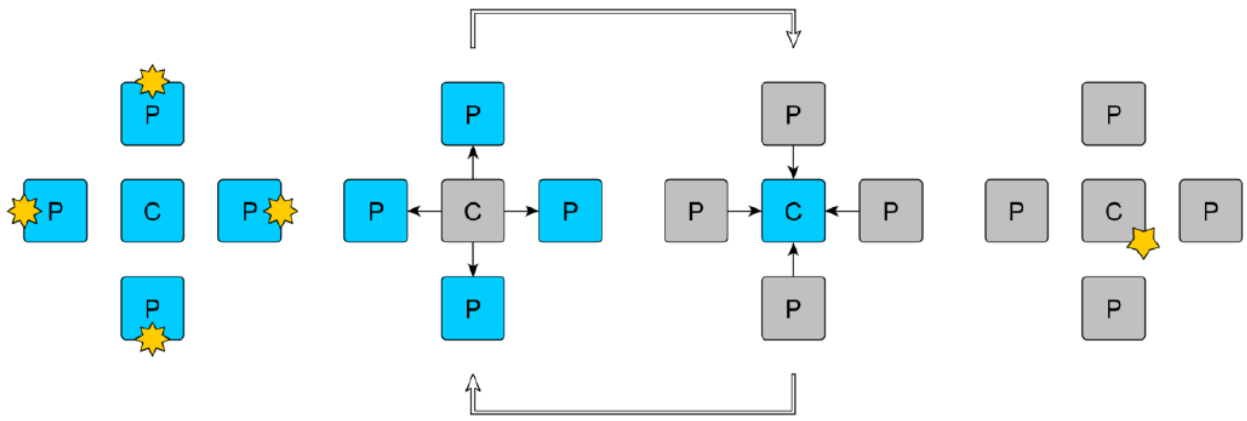

Figure 3. *Four stages of federated execution in FeatureCloud. The four main stages are 1) local data loading, 2) broadcasting a global model, 3) gathering local models, 4) compiling results. Stage 2 and 3 can be repeated depending on the executed algorithm. 'C' and 'P' stand for coordinator and participant, respectively. The yellow stars in stage 1 and 4 represent local training data and global parameters, respectively.*

After a local optimization operation has been completed by a participant, it sends the local parameters to the coordinator. The coordinator collects these parameters and aggregates them into a common (global) model, which is shared with the participants. Depending on the type of ML algorithm, these two operations can alternate a couple of times, e.g. until convergence or a predefined number of iterations has been reached. For some algorithms (e.g. random forest, linear regression), only one iteration is necessary. However, this strict separation between optimization and aggregation is not actively enforced by FeatureCloud. In many cases, aggregation can already start after the first parameters have been received, thereby increasing efficiency through parallelization of the computation. Figure 1 shows the logical roles of coordinator and participant, however in practice the coordinator usually has local data as well. Therefore, FeatureCloud also allows the coordinator to additionally assume the logical role of a participant [D.7.2:2.2].

5.1 Architecture

In this section, the technical details of the FeatureCloud system are described. It is split into an overview of the system architecture, i.e. the high-level constellation of the system components and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

their tasks, and the respective software architectures and details on behaviour and applied technologies of these components. [D7.2:4]

The FeatureCloud architecture consists of the following system components (see Fig. 4): Local Controller (deployed at every data holder's site, i.e. participant), relay server, global backend, frontend, and App Store Server (formerly known as AI Store Server). [D7.2:4.1]

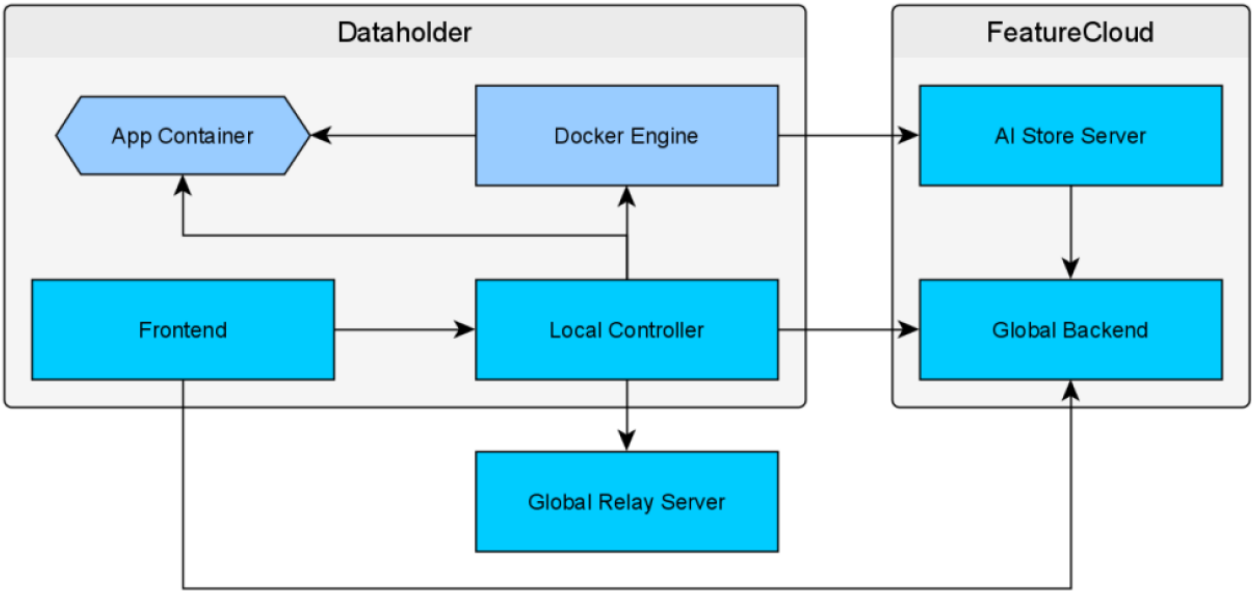

Figure 4. *Interactions between the FeatureCloud system components. Frontend and local controller are at the data holder's site (i.e. participant), App Store server and global backend run on FeatureCloud servers.*

On the data holder's (participant) site, the controller and frontend web application are running. On the FeatureCloud servers, the App Store Server, including a Docker registry, and the global backend, are running. Optionally, a global relay server is provided by FeatureCloud as well, in case setting up a custom relay server is not required or not possible. [D7.2:4.1]

The controller orchestrates app execution by instructing the Docker engine to create or shut down app containers, create and mount input and output volumes, and expose the required ports for the FeatureCloud API. It also serves as a proxy between the frontend and the app containers to decouple containers from the frontend. The frontend is used to access the controller and manage the FeatureCloud account, federated apps, and projects, which involves the global backend. The relay server acts as a communication hub for all participants of a workflow. Since it relays model parameters and has access to this data, users might want to use their own relay server instead of using the one provided by FeatureCloud. The App Store server is used to host app images and is described in section 5.2.2 in more detail. The global backend stores all user information, information about data holders, apps, projects and workflows, and is involved during workflow execution by saving the current step and progress. However, it never has access to any raw data or traffic between apps participating in a workflow, which is one of the crucial properties of FeatureCloud. [D7.2:4.1]

The FeatureCloud platform intends to accelerate all steps involved in federated learning, in particular the development of federated algorithms by providing an open API, the deployment and distribution of the algorithms though an App Store (https://featurecloud.ai/app-store) and the application of the algorithms for individual use cases in the form of configurable workflows.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The FeatureCloud platform consists of a number of nodes (sites) running the machine learning algorithms on the locally available data, communicating with a number of other services that orchestrate the process (see Figure 5). [D2.1:4.3] The platform is controlled via a web application (frontend) involving user rights management. For a more detailed description of the system, the reader is referred to deliverable D7.2 “App store ready and extendible by developers”.

The following conceptual figure [D6.5:Figure 2] also illustrates how FeatureCloud components could be deployed in every participating hospital (participant) and the related data processing and data flows from a technical point of view. The grey area represents the FeatureCloud components which are deployed in the participant site. The database on the right represents the existing patient data participant site. There is no direct access to that data for processing it for FeatureCloud purposes by other participants or the coordinator. Instead, through a separate manual or automated loading process upstream only data for which a permission exists is copied to a separate data store for FeatureCloud purposes. This data store is still a local data store at the participant’s site. As mentioned above, it is the key feature of FeatureCloud that this patient data never leaves the local site. However, the separation by way of that data loading process is absolutely necessary because a separation of patient treatment on the one hand and research on the other hand is strictly required. In particular for safety and security reasons, research cannot be carried out directly on the data in the local patient data management system.

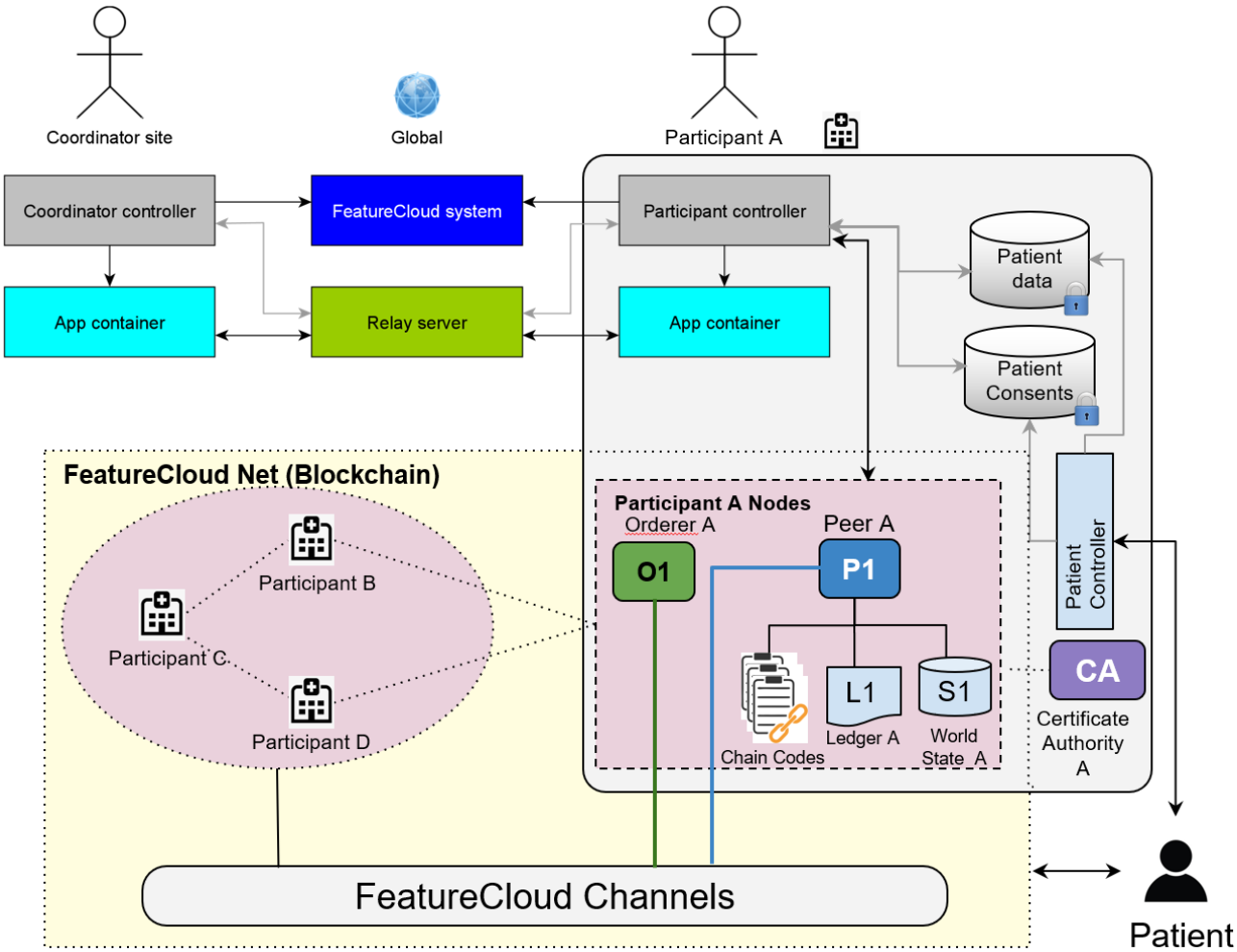

Figure 5. *Integration of the blockchain-based consent solution into the FC architectural concepts including the participant site. The figure focuses on the integration of the blockchain-based consent solution developed in WP6 and summarised in chapter 5.2.6.*

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

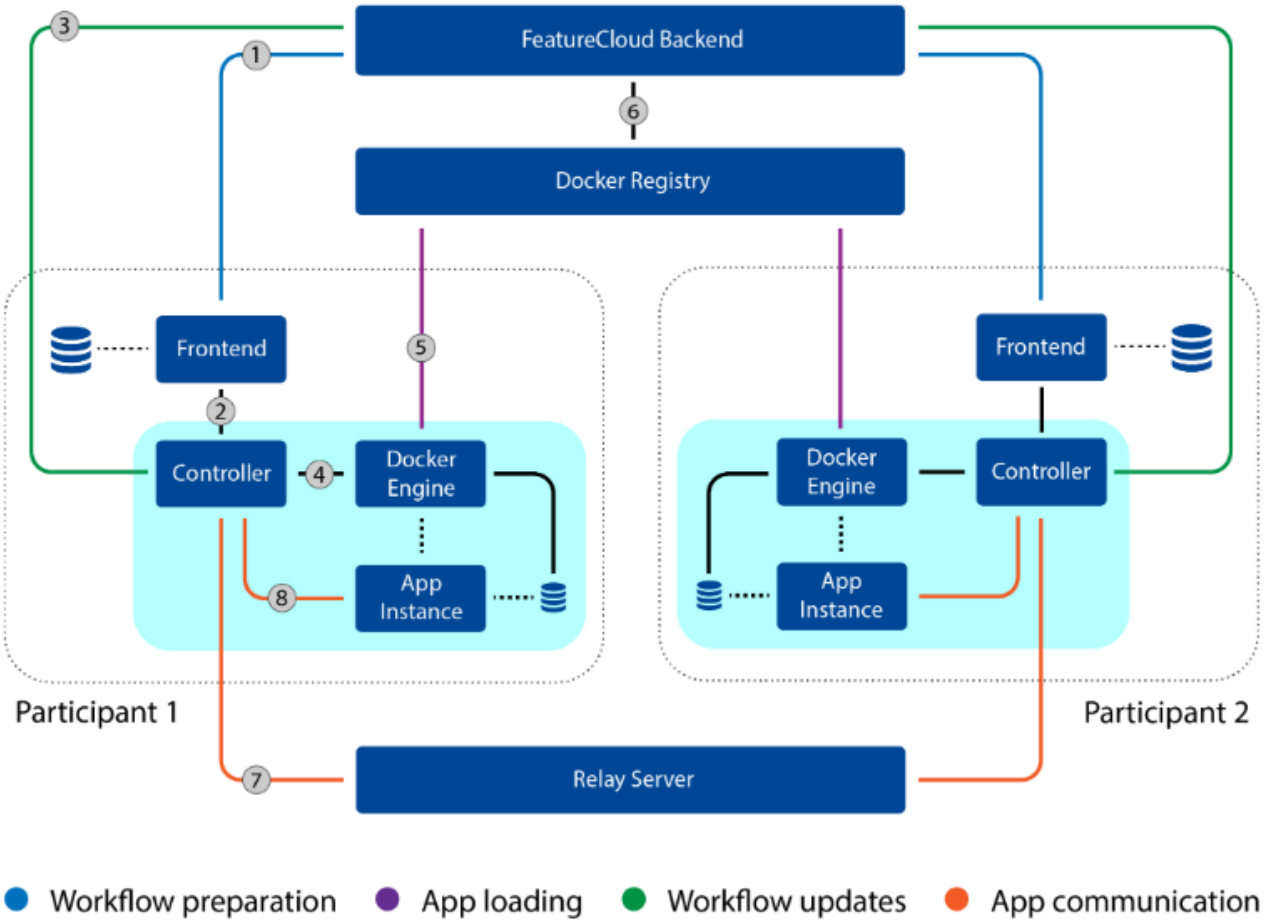

Figure 6. *FC architectural components and their interplay.*

5.2 Components and Concepts

FeatureCloud's primary goal is to simplify the development and usage of federated ML algorithms. This involves the following concepts and three phases: development of apps, distribution of apps, and usage of apps. Each of these is described in the following subsections. Furthermore, the technical details of the FeatureCloud system (system and software architecture) are described. [D7.2:3]. The goal of the KPIs presented in deliverable D2.2 is to measure how well data security is ensured and privacy leakage is mitigated.

5.2.1 FeatureCloud Apps

Since FeatureCloud does not impose restrictions on the kinds of algorithms it supports, the execution environment of the federated apps is kept very general. It allows for implementing any type of ML algorithm and an optional custom graphical user interface (GUI) for user interaction, also referred to as app frontend. [D7.2:3.1]

From a technical point of view, a FeatureCloud app acts as a web server (see section 5.2.5), responding to requests sent from the FeatureCloud system or the app frontend. No direct Internet access is granted to apps, as that would pose a security risk.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

System access of apps should be as limited as possible to rule out several attack vectors, such as leakage of sensitive data or access of data that should not be included in an ML algorithm (e.g. due to lack of consent). FeatureCloud uses Docker as a virtualization technique. Docker has been widely adopted in the developer community, particularly in the area of web development. It is available for all major operating systems (Linux, Microsoft Windows, macOS), making FeatureCloud nearly platform-independent. Docker also offers the necessary level of isolation, preventing Internet and file access if not explicitly granted, and sandboxing to limit the usage of compute and memory resources if necessary. These isolated running environments (containers) are created from pre-defined images, which are the federated apps in our case. [D7.2:3.1]

ML apps need access to training data to optimise their models. As depicted in Fig. 7, apps cannot access sensitive data directly. Instead, the app user needs to provide the data to the FeatureCloud system, making it available to an app. From an app perspective, the data can be expected to reside inside a dedicated input directory mounted to the app container. Analogously, all results generated by an app must be put inside an output directory, which is provided by FeatureCloud as well, whose contents can be downloaded via the FeatureCloud user interface. [D7.2:3.1]

This output data can also be picked up by another app that is executed successively and finds the output data of the previously run app inside its input folder. Chaining these apps is a feature that allows the composition of multiple apps into a workflow, a concept that is further described in section 5.2.5. [D7.2:3.1]

The FeatureCloud controller (Fig. 7) regularly asks a running app whether it has new model parameters to share with the other members of the federation. If this is the case, it loads them from the app and hands it over to the other apps, depending on whether it is a participant or a coordinator, as described in section 2. If the global model parameters need to be shared with the app, the controller actively sends them to the app. [D7.2:3.1]

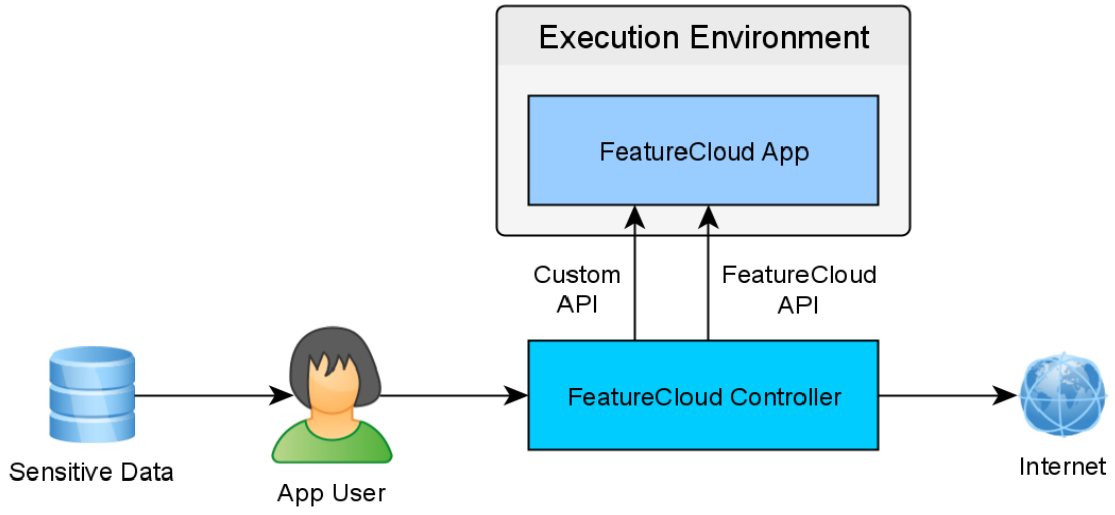

***Figure 7**. Execution environment for FeatureCloud apps. App users decide what data to load into the system. FeatureCloud apps cannot directly access the file system or the internet.*

The API is based on the TCP/IP-based HTTP protocol, which is asynchronous by its nature. In web terms, the FeatureCloud app acts as a web server and the FeatureCloud controller acts as a web client. [D7.2:3.1]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

An app can also provide its custom user interface to allow for monitoring the computation or for interaction with the user. To this end, additional endpoints need to be defined, which can then be accessed from the app user's browser. Web technologies are used for GUI design, i.e. HTML/CSS/JavaScript. Custom endpoints cannot be accessed directly due to technical and security reasons. In a typical scenario, the FeatureCloud controller runs on a central server inside a data holder's local network and users access the frontend from a different machine inside the network. To make this possible, the FeatureCloud controller listens on only one port, which can e.g. be accessed through an SSH tunnel, redirecting all app-targeted traffic to the correct container.

The development of algorithms involves intensive testing and debugging. For rapid development, it is crucial that these testing and debugging cycles are as quick as possible. Therefore, FeatureCloud comes with a local test framework that enables app developers to instantly run their application on their machine without deploying it first. When using this functionality, one has to specify the number of participants, i.e. app instances to simulate, and a data directory for each instance containing the respective input data. When started, the FeatureCloud controller creates one container for each instance and connects them logically identically on the developer's machine to a truly federated setup on different machines. [D7.2:3.1]

The API has deliberately been designed in an algorithm and usage agnostic way. This leads to high flexibility but requires the app developer to implement all algorithm-specific functionality by themselves. To quickly introduce developers to the API and provide a convenient starting point for app development, FeatureCloud comes with a collection of easily extendable templates. This collection includes a minimal template with a demo Python/Flask implementation, stubs for all API calls and a blank demo frontend, and a federated mean app. [D7.2:3.1]

5.2.2 FeatureCloud App Store

FeatureCloud presents all implemented apps in an easily searchable App Store to make federated ML available for as many users as possible. Conversely, app developers need a simple way to share their apps and attach important information, such as the required input data format, the format of the produced output, usage instructions, and privacy considerations. [D7.2:3.2]

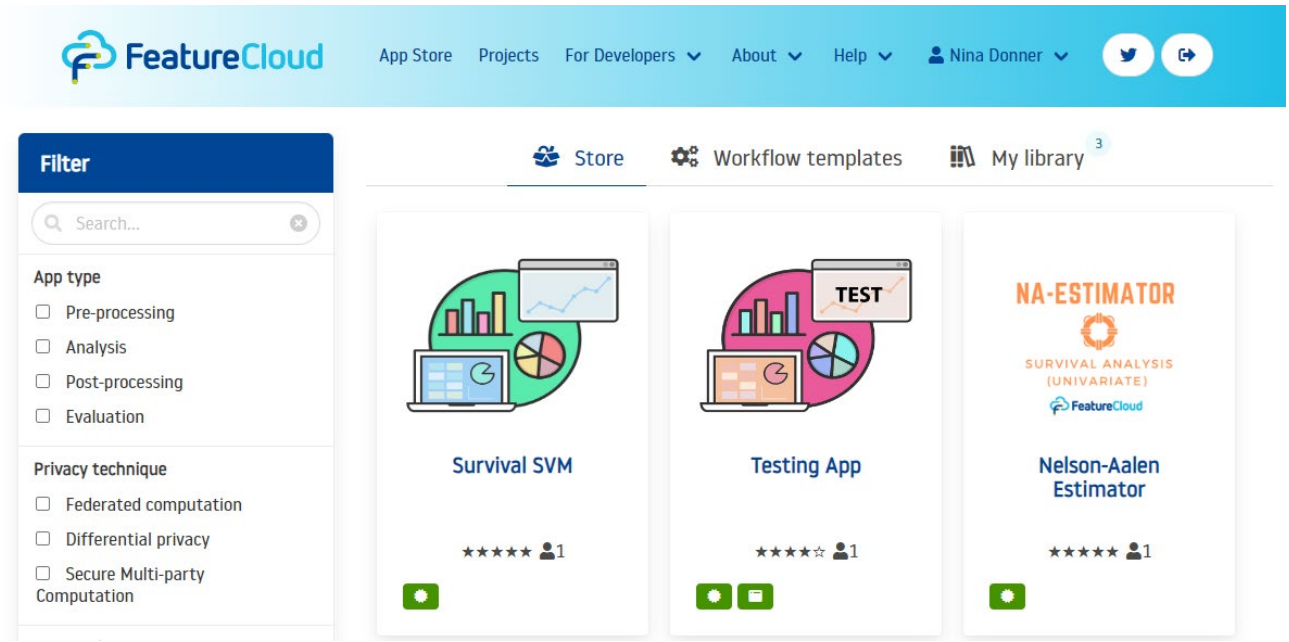

Figure 8. *FeatureCloud App Store. Users can select from a variety of ready-to-use apps.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

As described in section 5.2.1, all apps are stored as Docker images. Conventionally, docker images are shared using a Docker registry, to which new or updated images can be pushed and existing images can be pulled. FeatureCloud uses a standard Docker registry and controls its access through a proxy server (see Fig. 9).

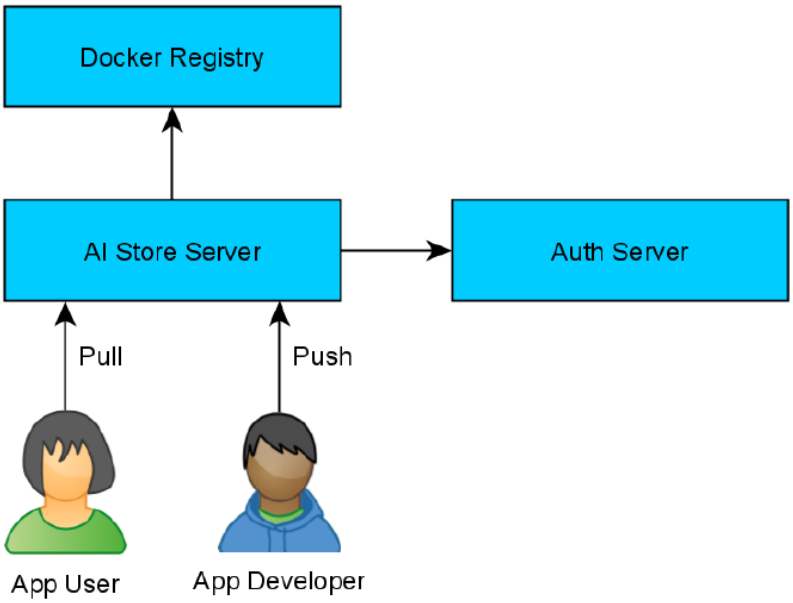

Figure 9. *Users and developers access the Docker registry through an App Store Server. App users can only pull, app developers can also push new images.*

Sharing the app after implementation on the FeatureCloud App Store involves the FeatureCloud website and usage of the Docker CLI [See D7.2:E/2 Manual for App Developers]. When a new image is pushed to the registry, the App Store Server assures that the developer has the required permissions to push the respective image by connecting to the Auth Server. If it is successfully authenticated, the image is uploaded to the FeatureCloud Docker registry and becomes available to other users. [D7.2:3.2]

Pulling the images is done automatically when the workflow starts running [See D7.2:E/3 Manual for App Developers]. The only technical requirement is a Docker installation on the user end. All the other steps are performed without user interaction. [D7.2:3.2]

After the app image is available in the FeatureCloud App Registry, developers can publish their app in the App Store. As a first step, a FeatureCloud account needs to be created and verified by a confirmation link sent to the corresponding email address. As soon as the user activates the developer mode in the profile settings, the App Store provides an additional tab “Developed” which will list all of the user’s developed apps. Furthermore, a developer sidebar, including a “Create App” button, becomes visible. [D7.2:3.2]

Users can search the App Store using full-text search or by choosing an app category or tag (see Fig. 10). Apps that have been reviewed by FeatureCloud are marked as such and shown by default. Other apps are only shown if the user explicitly accepts unsafe apps. Before using an app, users need to add it to their personal library of apps. This serves as bookmarking and allows for adding an extra licensing step in the future. Once added to their library, users can include an app in their workflow and provide the developer with feedback, i.e. a star-based rating and a clarifying comment.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

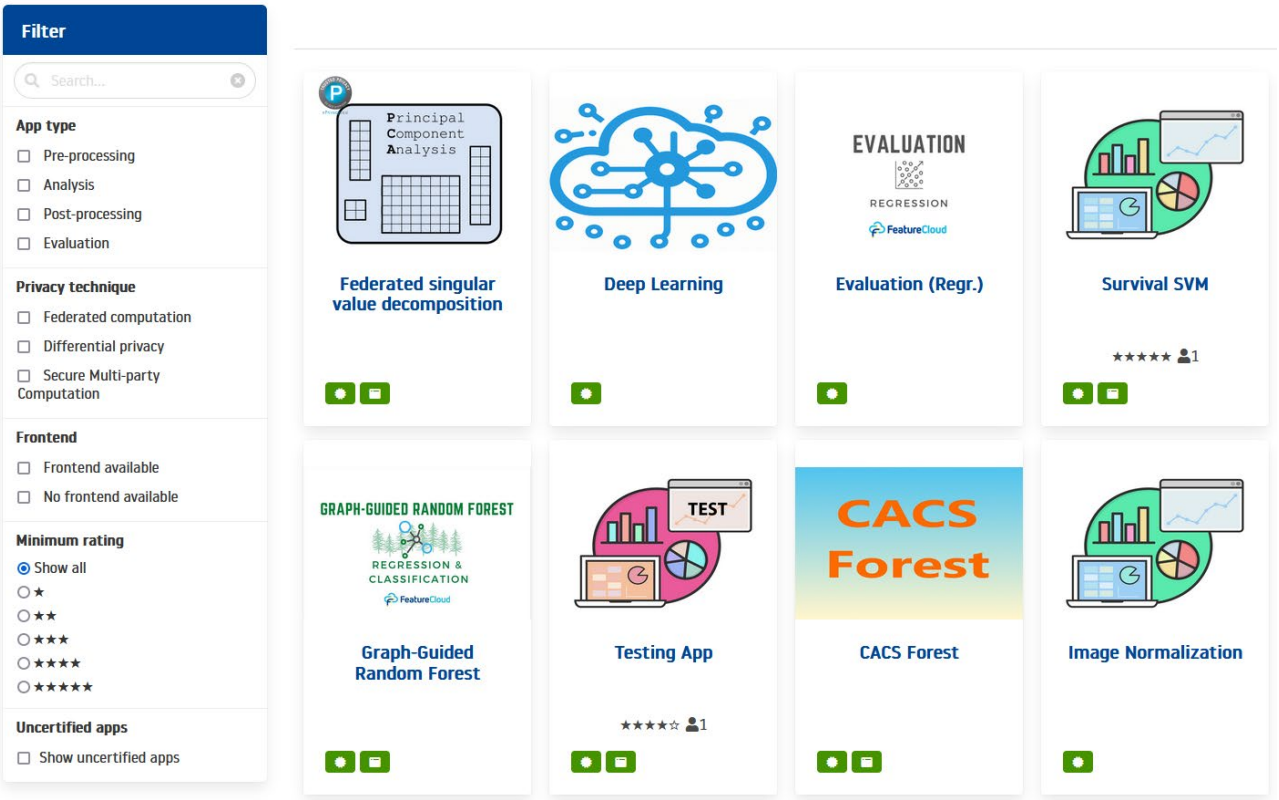

Figure 10. *App Store for developers. Developers can see the apps they already published in the "Developed" tab.*

Clicking the "Create App" button forwards the user to an input form (see Fig. 11) that is used to describe the information about the app. An app consists of a name, short and large description, an icon, and one or multiple labels that can be used as search tags. Furthermore, an app type needs to be selected that is either "Preprocessing", "Analysis", or "Visualization". The privacy technique defines what methods are used in the app to preserve privacy. For now, this can be "Federated Learning", "Differential Privacy", "Secure Multi-Party Computation" or combinations of them. Finally, the image name needs to be defined to connect it to the corresponding app in the FeatureCloud App Registry. The app information can always be updated by the app author at a later point.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

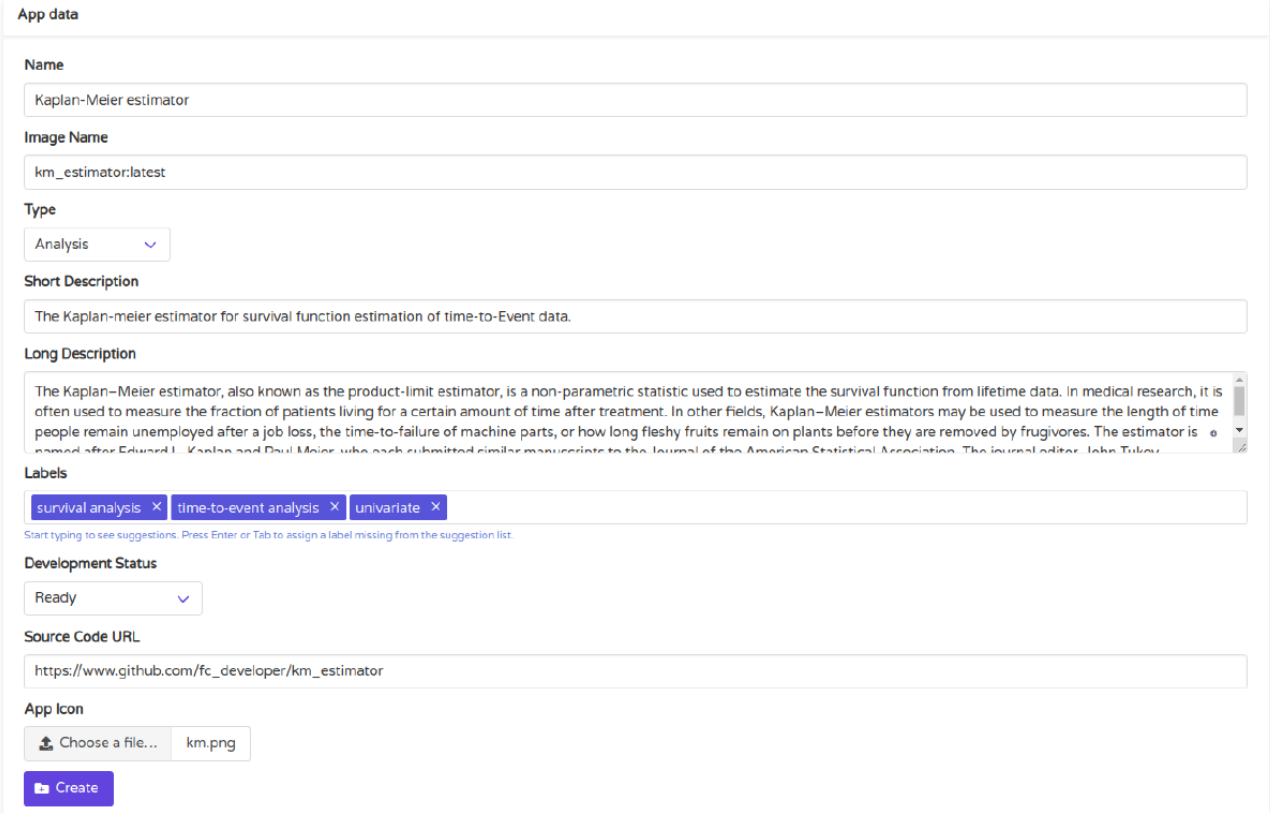

Figure 11. *Publishing an app. Developers can publish an app by defining the app information and link it to a Docker image in the FeatureCloud App Registry.*

5.2.3 App Certification

Allowing third-party developers to quickly and easily push apps to the app store and use them for collaborative studies is one of FeatureCloud's selling points. However, it is difficult to automatically ensure privacy awareness of such apps (see deliverable D2.2, KPI 'Privacy Requirements'). Therefore, FeatureCloud distinguishes between two types of apps: 1) certified ones and 2) uncertified ones. By default, the app store only displays apps that have been certified by the responsible FeatureCloud member (currently) or auditor (in the future). The user needs to actively choose to display uncertified apps and is warned and informed about the risks. In general, users are advised to only use uncertified apps from a source they trust, e.g. a collaboration partner they already work together with. [D7.2:3.2]

If developers want their apps to be certified, the source code needs to be completely accessible to the responsible FeatureCloud member (currently) or auditor (in the future). After the code has been reviewed and deemed secure, the Docker image is built by the consortium member/auditor and pushed to the app store via the Docker CLI. The FeatureCloud system recognizes the author of the image as a trusted party and marks the corresponding image version, identified by its SHA256 digest, as certified. After a successful certification, third-party developers can still push new uncertified versions to update the app or fix bugs. However, each update of the app, and respectively each new version, needs to be certified again to make sure that the update does not raise any privacy issues. In the meantime, all users who added a particular version of an app from the app store to their personal library of apps will automatically keep this version in their library. At this point, if a user wants to update their app to a new (certified) version, they need to remove it and add it again. This

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

process will be simplified in future App Store releases and replaced with a convenient update functionality as it is already established for mobile phone app updates. [D7.2:3.2]

5.2.4 Workflows

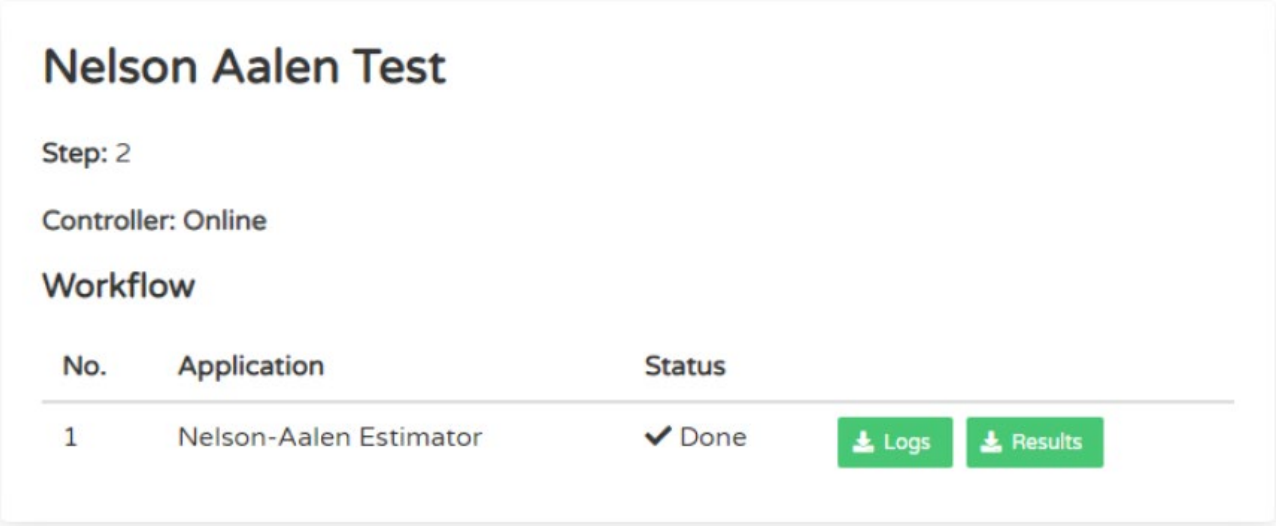

Figure 12. *Workflow has finished successfully. A workflow which has successfully completed offers its results as download in the FeatureCloud frontend. Logs can also be downloaded for debugging purposes.*

To run a study with other collaborators, a project needs to be created in the FeatureCloud frontend first. Projects consist of a name and a brief description to provide information for invited collaborators. Additionally, they contain a workflow, defining which apps will be executed in which order. The creation of a project is only possible for FeatureCloud users assigned to a data holder site (see section 5.2.5, Fig. 16). When they create a project, they act on behalf of their site. In practice, users typically are medical doctors or academic researchers who administer a FeatureCloud project, and sites are medical facilities or academic institutions. [D7.2:3.3]

The following two subsections describe how a workflow can be composed from a user (coordinator) perspective and how the consecutive execution is performed from a technical perspective. [D7.2:3.3]

All apps that should be part of the workflow need to be added from the user's library of apps (see Fig. 10). However, it is not required that all participants later have the apps in their library. After all apps have been added, the project is finalized and becomes immutable. To invite other collaborators, tokens (i.e. large random strings) need to be shared with them. Tokens are uniquely linked to a project and allow for joining the project. They can only be used once for security reasons and need to be entered in the FeatureCloud frontend. Once all participants have joined, the coordinator can take a final look and start the project. From that moment on, no one can join anymore and the execution begins. [D7.2:3.3]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

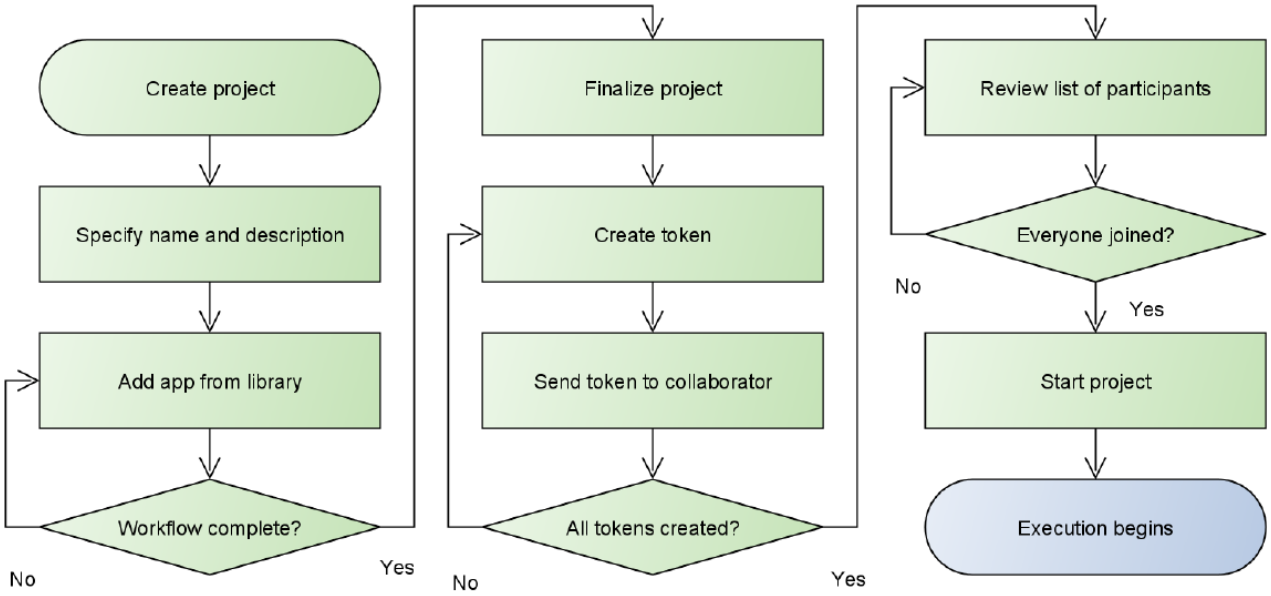

Figure 13. *Process of composing a project, inviting participants and starting a project. Green symbolizes human interaction; blue symbolizes automatic behaviour.*

Once the coordinator has triggered the execution, the FeatureCloud controller creates the input volume for the first app in the workflow at each participating site. This volume needs to be provided with the actual data relevant to the study, which is processed by the workflow. The users need to select the data via the FeatureCloud frontend, which is then sent to their local controller, importantly not leaving the data holder's domain. As described in section 5.2.1, each app has an input and output directory, which serves as a file-based data interface to FeatureCloud. [D7.2:3.3]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

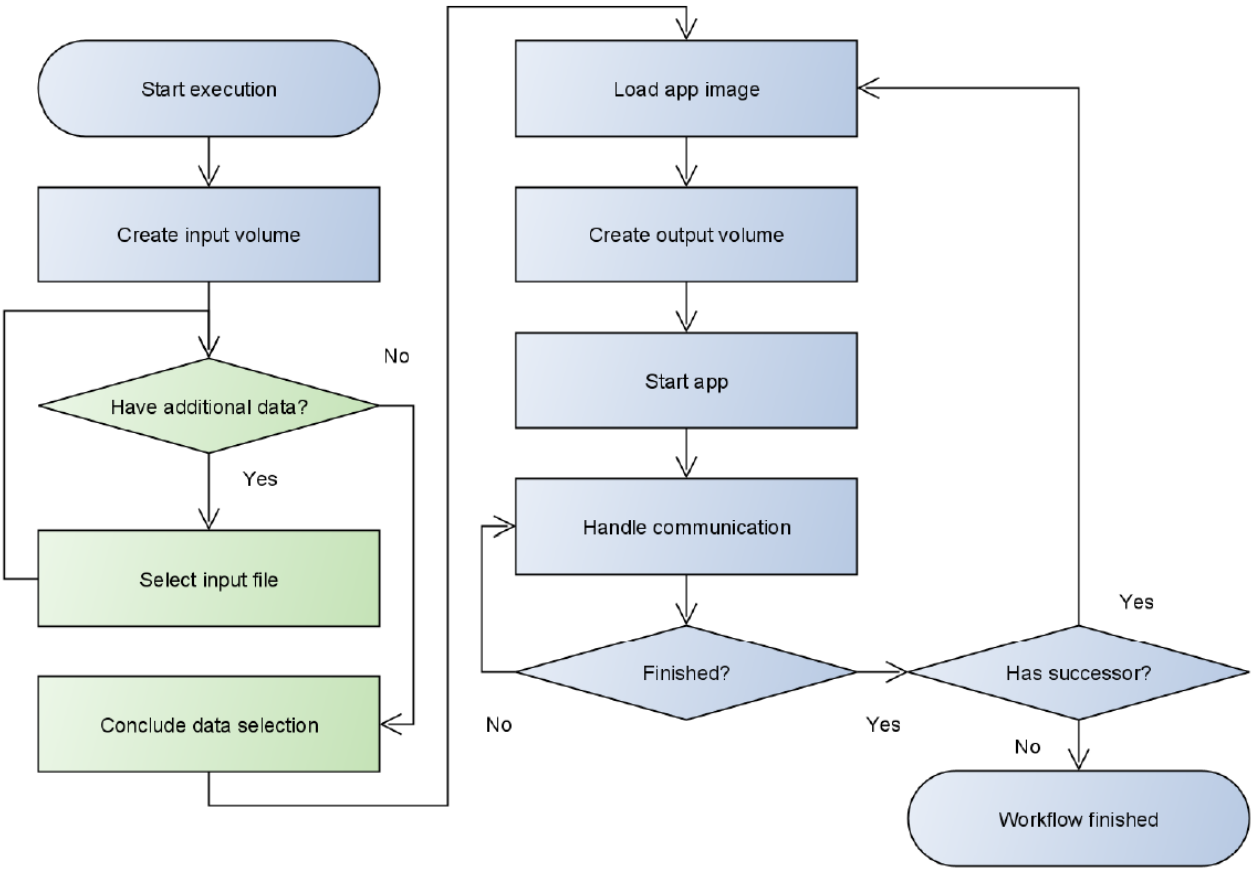

Figure 14. *Workflow execution managed by the FeatureCloud system. Green symbolizes human interaction; blue symbolizes automatic behaviour.*

After each participant has selected their input data, the first app is started as a docker container. The current progress of the workflow can be monitored in the frontend, showing the currently executed step (i.e. app) and providing its container logs if required (see Fig. 12). In general, no user interaction is necessary from this point on unless an app in the workflow actively requires so through its custom frontend. The app frontends can be accessed from the workflow page as well, usually to monitor app-specific events or view visualizations provided by the app. When the computation of a step has been completed, indicated by the coordinator app instance, all containers of this step are shut down and the contents of the output directory are placed inside the input directory of the next app in the workflow. These intermediate results can also be downloaded from the frontend for later investigation or detecting potential errors in the analysis. All debugging output produced by apps is stored in a directory on the controller machine, to investigate errors that might occur during execution. [D7.2:3.3]

A workflow can thus be regarded as a processing pipeline, composed of apps provided by Feature-Cloud, enabling an additional level of customizability (For a selection of the currently available apps, see D7.2:5; for a complete workflow sequence diagram, see D7.2:E/3 Manual for App Developers). [D7.2:3.3]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

5.2.5 Implementation

This section contains information about technology, software architecture and implementation details for each of the integral FeatureCloud system components. [D7.2:4.1]

Local Controller. The local controller needs to be able to handle large amounts of data and asynchronous tasks as well as keep up multiple socket connections and support HTTP-based and raw byte traffic. For this reason, this component has been implemented in Go (aka Golang), a native programming language developed for server applications. It allows for lightweight co-routines to monitor app containers and regularly query for updates from the global backend. [D7.2:4.1]

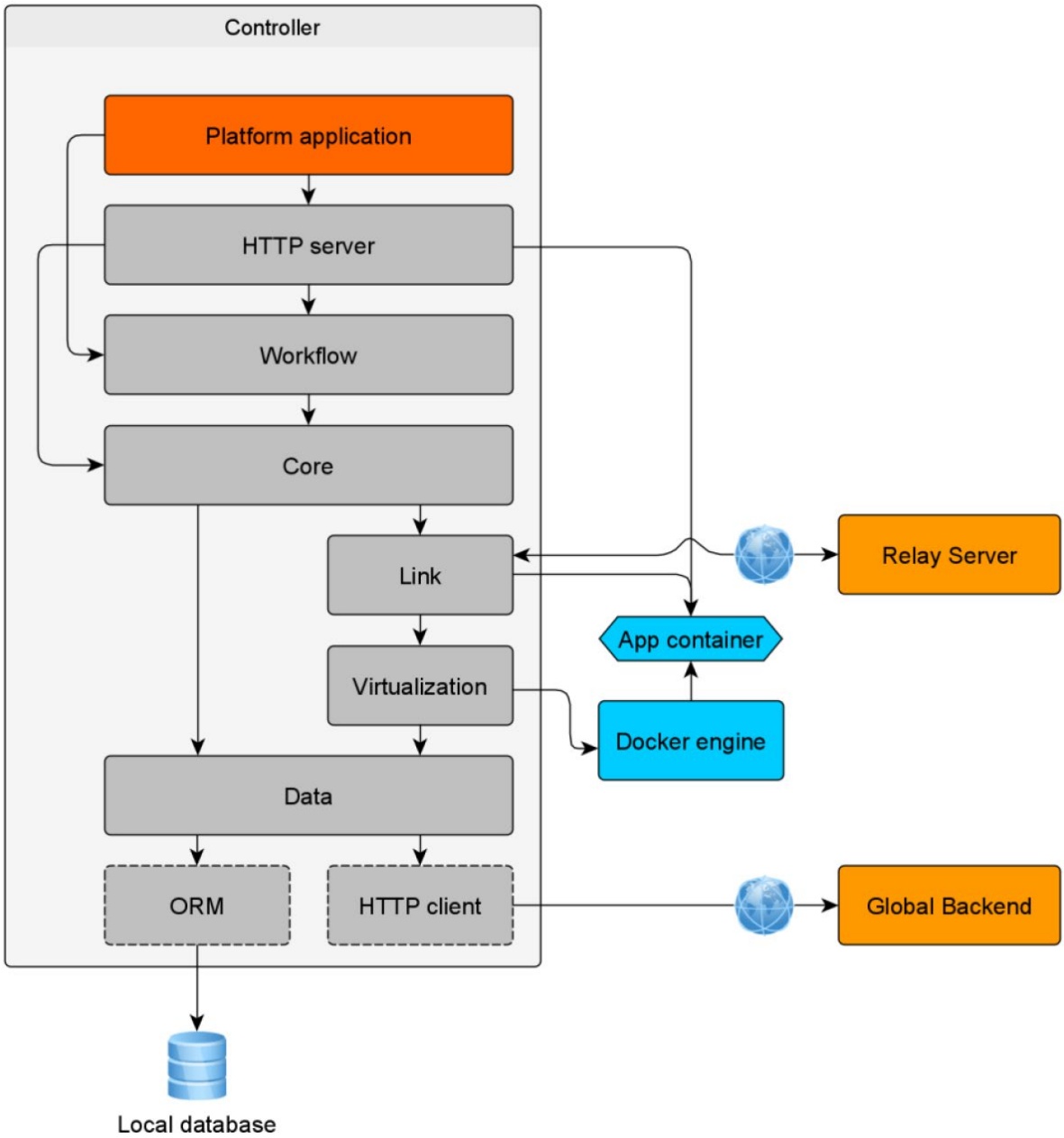

Figure 15. *Software architecture of the local controller. It uses a layered architecture preventing arbitrary access across layers by enforcing a partially ordered access hierarchy.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The software architecture has a layered structure, with a decreasing level of abstraction from top to bottom (see Fig. 15). The platform application layer is the main entry point responsible for reading configuration values (e.g. local database credentials, address of the global backend) and starting an HTTP server and polling routines. The HTTP server provides endpoints for the frontend to control workflow-related tasks, such as loading data into the first input volume, showing container logs. It also relays traffic to the app-specific frontends. The workflow layer offers abstract functions for the HTTP server and takes care of workflow management, such as setting up and attaching volumes, starting containers, shutting them down, reacting to updates from the global backend (by using the data layer through the core layer). The core layer provides an abstraction of the core business logic, especially app container management and functions for testing apps during development. The link layer handles communication between app containers and the relay server, translating raw byte-traffic from the relay server to HTTP-based traffic for the containers and vice versa. The controller acts as an HTTP client in this case, and the app containers as HTTP servers. This way, active access by the app containers to the Internet can be avoided. The virtualization layer is a direct abstraction of Docker, which allows for replacing the virtualization technique in the future if needed for security or compatibility reasons. [D7.2:4.2]

Relay Server. The relay server implements basic relay functionality for star-based federations of clients. It knows the role of each client (i.e. participant or coordinator) and treats their traffic accordingly. If data is received from a participant, it relays it to the coordinator. If it is received from the coordinator, it is broadcast to all clients. A relay server can handle multiple workflows at once. For that, it uses workflow-specific credentials chosen by the coordinator and automatically distributed to the participants by the global API. Like the controller, it is written in Go since it needs to efficiently handle large amounts of binary data, which Go is capable of. [D7.2:4.2]

Global Backend. The global backend mainly offers an HTTP API for controllers and the frontends. It is responsible for managing all necessary data related to projects, apps, users and data holders (sites). It is implemented in Django, a Python web framework that offers the functionality for this kind of task, particularly database abstraction, URL routing and web-related utilities (e.g. JSON serialization, HTTP abstraction). [D7.2:4.2]

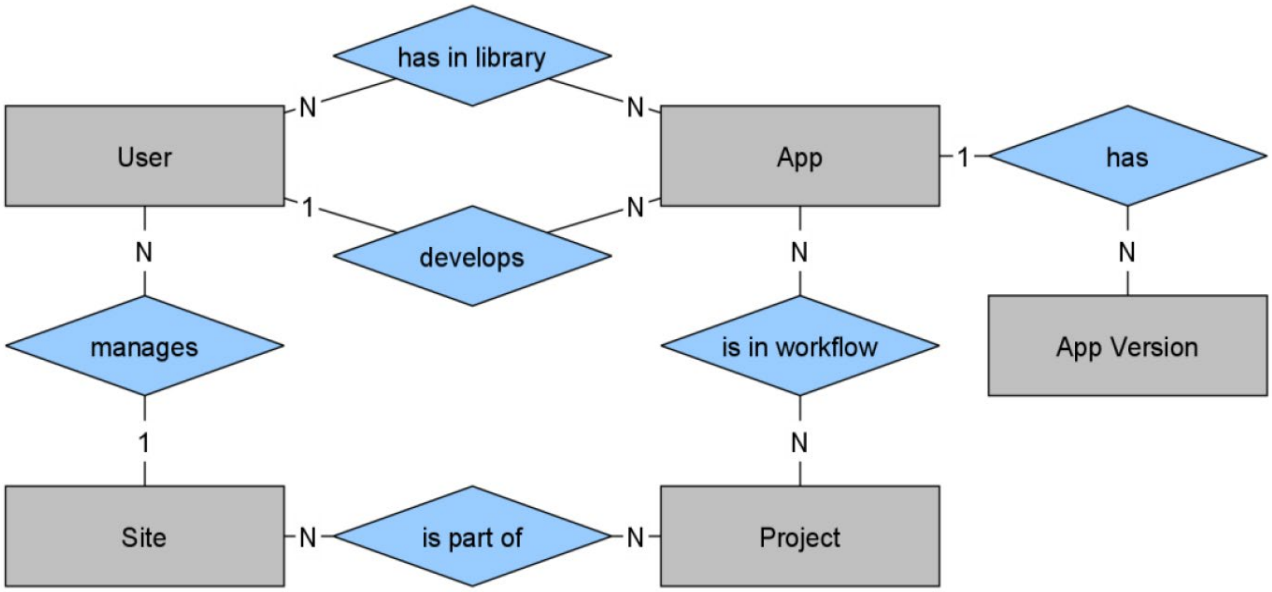

Figure 16: *E/R diagram of the data model in the backend. Grey boxes represent entities, blue diamonds represent relationships.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The E/R diagram of the data model is shown in figure 16. The global backend allows controlled access to instances of these entities. [D7.2:4.2]

User. Users have an email address and a hashed and salted password allowing them to log in to the FeatureCloud frontend, which then queries the global backend. In practice, a user is either a developer who has apps linked to them through the 'develops' relation, or an end user. Both, developers and end users, can add apps to their library (relation 'has in library') and manage a site (relation 'manages').

Site. Sites have necessary contact information and represent a data holder location, e.g. a hospital or academic research institution. Each site needs to run a controller instance (see Fig.15) to participate in projects (relation 'is part of'). When a site is part of a project, it can either assume the role of the coordinator or a participant.

Project. Projects encompass a workflow, descriptive information and a set of tokens allowing for joining a project (see section 5.2.4). Tokens are not modelled explicitly. Instead, the 'is part of' table is used, which can have entries with a token string and where the related site is NULL. Once a site joins a project, this entry is linked accordingly and can no longer be used by anyone else.

App. Apps are AI applications which appear in the app store. They contain an image name, which needs to be used when pushing new versions of the app, an icon, a short and long description, tags, a category and link to the source code. They are linked to a developer through the 'develops' relation and workflows they are part of through the 'is in workflow' relation.

App Version. New versions of apps are tracked automatically when pushing a new version via Docker by the developer and are linked to the respective app through the 'has' relation.

Frontend. The frontend serves as a graphical user interface (GUI) for FeatureCloud users and developers. It is the only component FeatureCloud users directly interact with. It then calls the API of the controller or the global backend on behalf of the user, depending on the nature of the task (local controller/ global backend). Since the frontend needs to be platform-independent, it has been implemented as a web application running inside a browser. This enforces a clear separation between GUI-related concerns and backend-related tasks by employing an HTTP-based API, as described earlier. Angular has been chosen as a web framework due to its high popularity, long-term support and extensive functionality. [D7.2:4.2]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

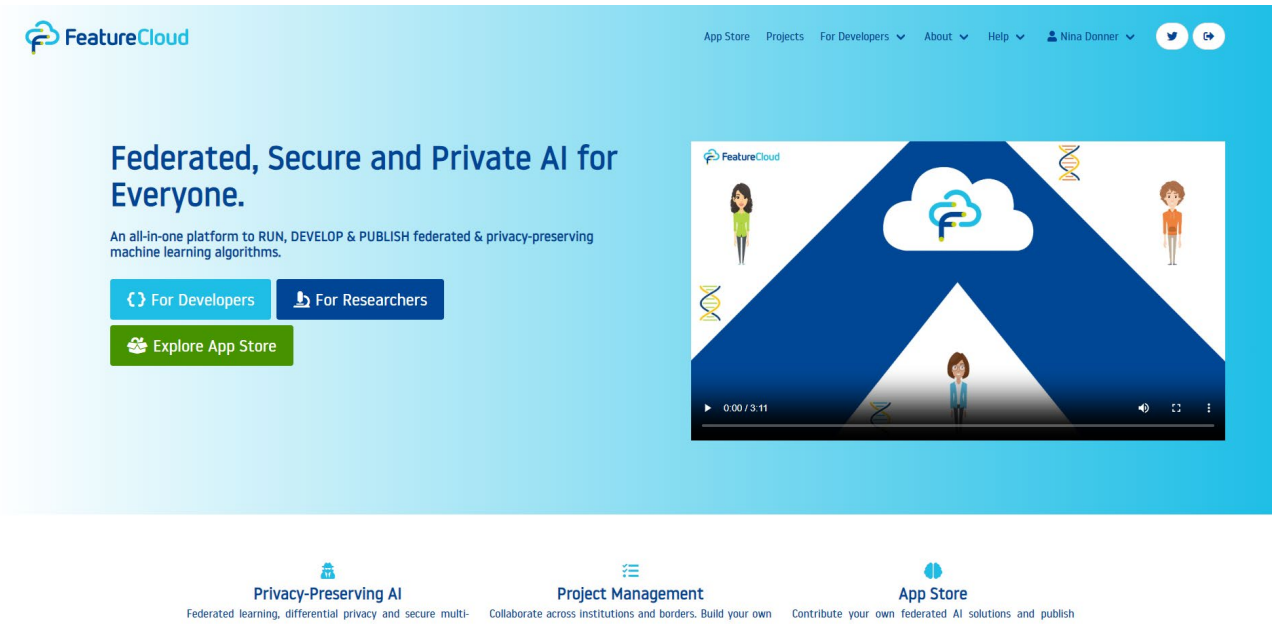

Figure 17. *FeatureCloud frontend. The frontend serves as a GUI for the users and allows intuitive project management, workflow execution, presentation of apps in the App Store, and many more.*

The GUI is structured into the following sections (accessible through the menu): 1) Account management, 2) Site management, 3) App management, 4) Project management and 5) App testing, each divided into subsections again. For more details and walkthroughs, see supplement, section 1. [D7.2:4.2]

App Store Server. As described in section 5.2.2, the App Store server is connected to the global backend that serves as an auth server and a Docker registry (see Fig. 9). It performs two main tasks: relay queries from the local Docker engines using the Docker registry API (https://docs.docker.com/registry/spec/api/) and protecting images from unpermitted access, in particular restricting pushing of images to the respective app developers. For that, the App Store server provides endpoints to request a JWT token which is then attached automatically by the Docker CLI to authenticate consecutive actions. App developers need to be FeatureCloud users and use their FeatureCloud credentials to login. That way, the global backend acting as an auth server can check whether the user pushing an image is the corresponding app owner. [D7.2:4.2]

Like the controller and relay server it is written in Go for performance reasons. App images can be several GB large and pulling images is a task performed each time before a workflow step is executed, making performance a critical consideration. [D7.2:4.2]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

5.2.6 Blockchain-based mechanism for logging and auditing of data usage

For the purpose of enabling complete control of all uses of personal data through FeatureCloud, a blockchain-based logging and auditing mechanism has been developed in Work Package 6. As outlined in D6.1, the general fundamentals and requirements for this development in WP6 were as follows:

- The system is expected to work in a minimal trust environment where some of the actors can behave maliciously (excluding the auditor).
- The system is expected to reduce trust assumptions on centralized services, and is able to tolerate or minimize the effects from maliciously acting or compromised participants.
- The system should make the audit process easier for detecting wrongdoings.
- The system should ensure traceability, confidentiality and integrity of healthcare data used for studies.
- It should be analysed if and how patient consents and identities could be managed digitally in a secure way.

Based on this, a comprehensive and structured list of 19 requirements with a particular focus on consent management from different perspectives such as legal, technical and privacy/ethical was compiled in D6.3.

An architecture was chosen that allows to create a permanent record of the machine learning activities managed by the platform. For this purpose, in particular the local manager executing the machine learning algorithms is required to create a new log entry whenever a new study is to be performed. The key elements for such entries include:

- input data (training set) for the machine learning model used
- matching consent confirmations for all patient in the input dataset
- intermediate states
- output data
- meta-information about the study itself, including a unique identifier
- specification of the machine learning algorithm and the corresponding hyper-parameters

To hide all sensitive information, the above elements are not stored directly. Rather the record is composed of fingerprints of the data calculated using cryptographic hash functions in the form of a Merkle tree (as described in Section 4 of D6.1). The properties of the cryptographic hash function ensure that it is not possible to derive any of the information using the publicly stored fingerprint, while the record serves as a binding commitment. Upon request by an auditor, the local manager discloses the requested record(s) (identified by its fingerprints) to the audit, which first verifies that record against the publicly stored fingerprint, and then checks the validity of the consent confirmations to ensure only data for which consent was given are actually used to train the machine learning model. The data is only required/extracted to be managed by the auditor during the duration of the actual audit. This drastically reduces the attack surface, as there is no single party where all data is required to be stored. The separation ensures that, in case of a data leak or compromise of one of the parties, the other parties are unaffected. To hold the local project manager accountable, each record (fingerprint) is further associated with the manager's identity. This identity is established by using asymmetric cryptography, in particular digital signatures. Prior to the use of the system, the local manager generates a cryptographic keypair and registers the public key at the auditor. Upon writing a new record to the blockchain, the record is signed using the corresponding private key.

A key characteristic of this solution, which follows from the above, is that an audit can only be carried out as long as the original data (i.e. the records of patient data on which the research was performed) has not yet been deleted (see also section 5.1.6.b of D6.2). In other words, this solution enables proving that the use of a particular data record for a study was lawfully based on valid consent as long as this record exists. This characteristic, following necessarily from the way this prove is provided cryptographically, has some similarities to the underlying principle behind Article 11 GDPR,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

which limits the obligation to store data only for compliance purposes and prioritizes data minimisation instead. A prototype based on the architecture and requirements described above was implemented in WP6 on the basis of a permissioned private blockchain based on Hyperledger Fabric and x.509 certificate-based identities. The underlying concepts are outlined in detail in D6.2, and the implementation is outlined in D6.4.

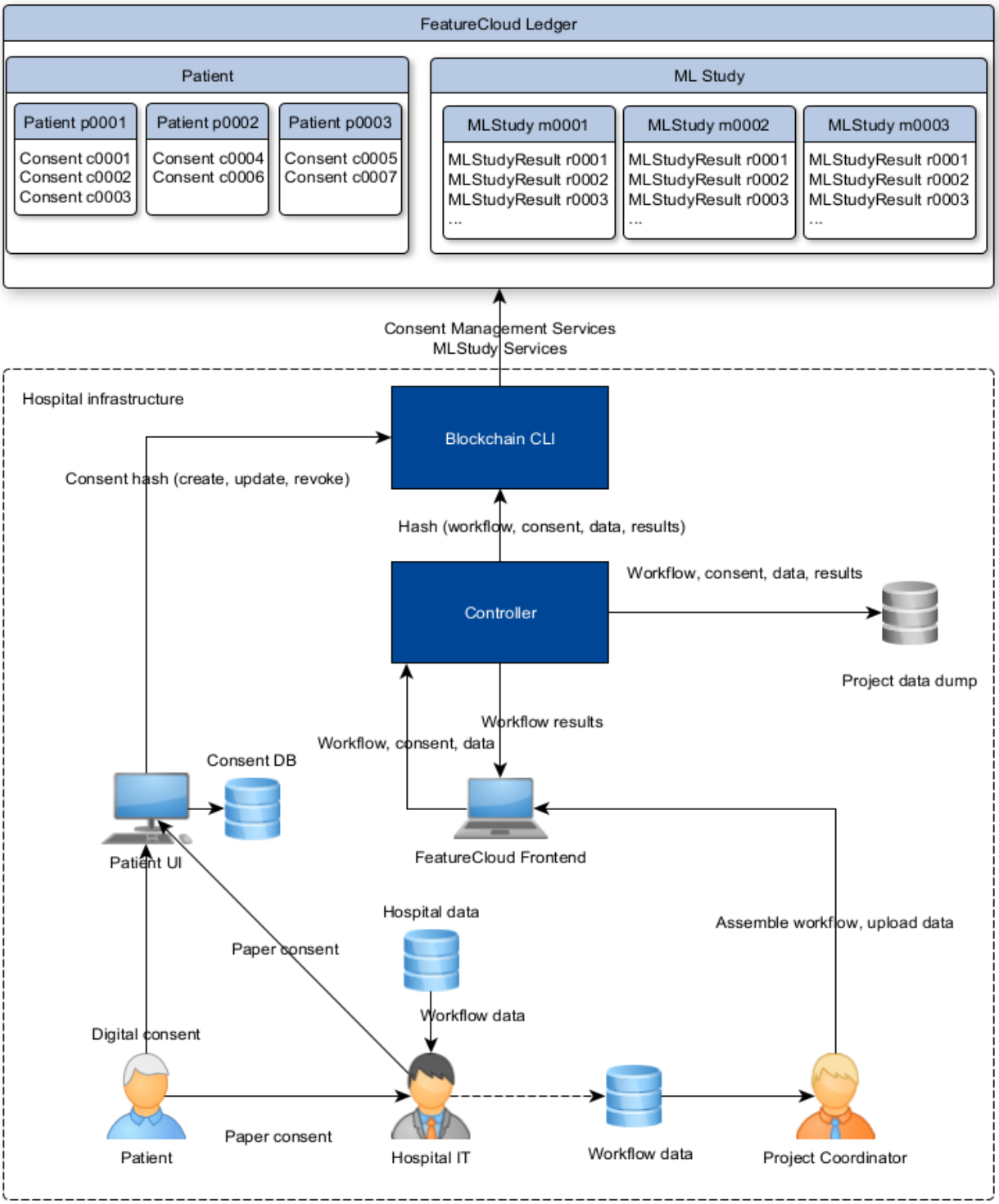

Figure 18. Fundamental elements of the blockchain-based consent management solution and their interplay.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

5.2.7 Specific technical and organisational Measures (particularly for App Developers)

In terms of privacy, the following fundamental technical measures need to be considered during deployment of FeatureCloud:

Although federated learning preserves privacy to some extent, the intermediate results (model parameters) exchanged between the parties may be abused and reveal some private information about the individuals [D7.1: 5.7, 5.8]. To address this and decrease risks to confidentiality, differential privacy and cryptographic techniques can be employed [D2.1:9.2.2].

The efforts in making ML models privacy-preserving can be categorized into four groups based on the method they employ: (1) federated learning (FL), (2) cryptographic techniques (including HE and SMPC), (3) differential privacy (DP) and (4) hybrid approaches. Each of these categories has its strengths and weaknesses in terms of computational and communication efficiency, utility and privacy guarantee. For example, FL - which is built into the architecture of FeatureCloud by default - suffers from high communication cost compared to HE and SMPC. However, as FL is based on the "moving computation to data" methodology rather than "moving data to computation", it is computationally more efficient than HE and SMPC. As another example, an FL model does not provide a privacy guarantee while a differentially private ML model does so (namely epsilon and delta). On the other hand, FL is a more utility-aware technique than DP as it does not inject any noise perturbation to the data or the training process. [D7.2:2.1.1]

Differential privacy (DP)

Differential privacy is a mathematical technique to quantify privacy and has attracted a lot of attention in recent years. The approach strives to ensure a chosen level of privacy by adding noise before, during or after the learning process of the function to make it more difficult to determine an individual in the dataset. The trade-off is that adding noise to the collected data may reduce accuracy. The technique can be applied to different machine learning algorithms (Shokri and Shmatikov 2015). This can be a defence mechanism for specific data analysis applications. Further, even single participants in the learning process may choose to employ this measure if they want to additionally protect their input.

Secure multiparty computation (SMPC)

This cryptographic protocol allows several collaborators to compute a common function of interest without revealing their private inputs to other parties. A SMPC protocol is considered secure if the parties learn only the final result, and no other information (Mugunthan et al. 2019). It can be used in addition to federated learning to compute models' average (Bonawitz et al. 2017) instead of relying on a central, trusted coordinator.

Homomorphic encryption (HE)

While traditional encryption does not generally allow for computation over encrypted data points, fully homomorphic encryption (FHE) enables arbitrary computation over encrypted data, opening the door to privacy-preserving applications of computational techniques such as machine learning and statistical analysis to genomic and medical data and outsourcing of computation (Wood, Najarani et al. 2020). Homomorphic encryption schemes, either fully or partially homomorphic, can be the solution to mitigate privacy risks in federated learning. The idea is to encrypt models' parameters before sending them to the aggregator, who performs operations on them. Additively homomorphic encryption was shown to ensure the security of federated learning for an honest-but-curious coordinator while preserving identical accuracy of a federated learning system without homomorphic encryption

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

(Phong et al. 2017). Several works proposed privacy preserving federated learning with homomorphic encryption on different regression models, e.g. ridge regression (Chen et al. 2018), logistic regression (Hardy et al. 2017). [D2.4:5.1.5]

Federated Learning

The only information being shared between hospital platforms are model parameters. This, however, does not guarantee that no private information can be retrieved from this data [Li et al. 2019]. The application developers need to make sure that the information they send around is unproblematic and this is also being checked again by a certification authority before the application becomes admitted to the app store [D.2.2:3.3.3].

Discussion

Each of these privacy-preserving techniques (DP, SMPC and HE) has its own limitations and this should be considered in choosing the proper technique. Specifically, differential privacy suffers from low utility in real-world applications, while the cryptographic techniques (SMPC and HE) have communication and computation overhead in real-world settings [5.9, 5.10]. In order not to lose too much accuracy, one can consider using only global DP, as it transfers the actual raw data with differential privacy mechanisms. As this highly depends on the algorithm, global DP cannot be integrated into the platform itself. The only way perceivable at the moment is to make the data itself differentially private. By using FL the health institutions (or the clients in general) can collaborate in training a common ML model while ensuring that the individuals' private data will not move out of their local sites (even in encrypted form). Moreover, if there is a case in which enhanced privacy is required, they can also privatise the FL model parameters using DP or other types of obfuscating techniques [D.7.2: 2.1.1].

Homomorphic Encryption (HE) (Rivest et al. 1978) schemes, in general, try not to only protect the confidentiality of data, but in addition, to allow for performing mathematical operations on the ciphertexts. When decoding the result, this is equivalent to having performed the operations on the plaintext. Fully homomorphic schemes (Gentry, 2009) allow arbitrary computation. Partially or somewhat homomorphic schemes allow only a certain subset of operations but provide increased efficiency. Newer schemes include e.g., CKKS (Benhamouda et al., 2017). This would in principle allow envisioning an architecture where the coordinator receives all local models in the ciphertext of a homomorphic encryption scheme, and computes (averages) the global model still in ciphertext, before sharing it back to the clients, which can then decrypt the global model for further use. However, even with recent improvements, HE is in general less efficient than SMPC. State-of-the-art HE implementations are thousands of times slower than SMPC in typical application (Evans et al., 2018). Another aspect that would need to be solved when employing HE in a federated setting is key sharing, which would require a secure scheme for key exchange. If all clients use the same encryption key, leakage of such a key would allow deciphering all local models. Thus, SMPC seems more promising to utilise within the FeatureCloud architecture. [D2.4:5.1.5] In addition, the weak spot of SMPC is collusion of participants, which is unrealistic in a setting where these participants are different hospitals.

Certification of apps

FeatureCloud distinguishes between two types of apps: 1) certified ones and 2) uncertified ones. By default, the app store only displays apps that have been certified by a privacy expert of the FeatureCloud consortium (currently) or auditor (in the future). The user needs to actively choose to display uncertified apps and is warned and informed about the risks. In general, users are advised to only use uncertified apps from a source they trust, e.g. a collaboration partner they already work together with. [See in detail 5.2.2; 5.2.3]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Guidance for developers

Based on D7.2, chapter E “Manual for App Developers” and the considerations outlined above the FeatureCloud consortium provides comprehensive guidance for FeatureCloud application developers to better enable them to develop and deploy applications that meet the expected privacy guarantees (see https://featurecloud.ai/assets/developer_documentation).

6 Identification of Stakeholder and Role Distribution

Based on the systematic description of the (envisaged) data processing operations, the required identification of relevant stakeholders including the clarification of their legal and organisational roles and accountabilities can be carried out (Vemou and Kadyra 2020).

On the one hand, the identified actors can be divided into internal and external stakeholders. On the other hand, stakeholders should be defined according to their role as either rights-holders or duty-bearers. The former encompasses future users, patients, consumers, or any other person/group affected by the data processing, with a particular focus on those in vulnerable situations. The latter is primarily the controller of the respective data processing activities as defined in Article 4 (7) GDPR.

Therefore, besides the identification of affected data subjects, the controller of the data processing has to be specified. The term controller, in the sense of Article 4 (7) GDPR, means ‘the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data’.

To systematically identify the relevant stakeholders, it is advised to check whether the processing activity in question involves/affects internal (e.g. data controllers, data processors, data protection officers, recipients) and external stakeholders (e.g. data subjects, third parties from the public and private sector representing data subjects) identify the type and level of their involvement/concernment (Kloza et al. 2020). On this basis, they can be addressed in the appropriate place within the DPIA.

6.1 Role distribution

The allocation of roles under the GDPR must always be aimed towards the isolated data processing operation. In order to be able to achieve an assignment of the central role of the controller based on "factual elements or circumstances" (EDPB 2020, para 12), first the specific processing purpose of the respective processing activity should first be considered and determined in isolation in order to find out why this processing is determined in the first place. From this first information it can subsequently also be derived in a supportive manner who is "served" by the processing of the personal data. In this way, the key question "Who initiated it?", which is essential for determining the controller, can be solved in a practicable manner in conjunction with all the facts (Article 29 Data Protection Working Party 2010, p. 11). According to Art 4 ref. 7 GDPR the controller decides on the purposes and essential means of the processing of personal data.

6.1.1 Governance Body

The governance body provides support and guidance to developers, coordinators and participants. It provides the FeatureCloud servers (see in detail Section 5.1) and will review and certify Apps that are uploaded to the FeatureCloud App Store. The governance body does not intervene in specific data processing operations and will not process personal or anonymous data of data subjects itself.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

The role allocation shall be based, above all, on the characteristic features of a controller within the meaning of the GDPR, since only the controller or controllers have the decision-making power regarding the purposes and means of the processing. The governance body exerts a certain influence on the data processing operations by the coordinators and participants and in a broader sense enables these operations by providing information and resources which might be interpreted as an exercise of the described decision-making power. Joint controllership may already be assumed between two organisational units when one unit influences the processing of another unit in its own interest (*Fashion ID v Verbraucherzentrale 2019, para 74*). It also not at all necessary that a configuring organisational unit has access to the data (*Tietosuojavaltuutettu v Jehovan todistajat 2018, paras 69 - 73*). Thus it might be argued that the governance body acts as a joint controller with the coordinator and/or participants which factually process the data. Due to the fact that the governance body does not configure any applications - this role falls to the developers - there is no comparable factual basis for the above decisions. The certification which is carried out by the governance body should not have a sufficiently intensive effect on the data processing to change this. Since the governance body itself does not process any data nor does it control the data processing by the participants, there are in fact no further indications which point to allocating the role of a (joint) controller regarding the machine learning operations.

However, the governance body is the controller regarding the data stored on the Global Backend [5.2.4] insofar as the GDPR is applicable to the stored data.

6.1.2 Developer

The developer uses the FeatureCloud API to develop federated algorithms which may be uploaded to the FeatureCloud App Store for further use. The developer may simultaneously assume another role. Developers can see the apps they already published in the “Developed” tab.

The developer programs applications for a later user - primarily the coordinator - and therefore does not ultimately determine the purpose or means of the processing of the coordinator for the latter has free choice over which applications he will use. The developer also has no own interest in the data processing (Fashion ID v Verbraucherzentrale 2019, para 74) and shall there not be categorised as joint controller with other actors in FeatureCloud.

Of course, in case the developer uses personal data for testing an app she develops, she will be in the role of the controller for this processing operation and must ensure to have a legal basis for it but since anyone can become a developer of FeatureCloud app this is out of scope of the responsibility of the Governance Body, the Coordinators and the Participants and therefore out of scope of this DPIA.

6.1.3 Coordinator

The coordinator is the site that compiles the project, assembles the workflow, and invites other participants. The coordinating site is the central aggregation entity in an FL workflow. The coordinator simultaneously may assume the role of a participant - for the purpose of role allocation the coordination activities will be considered first and foremost.

When federated learning is applied, while enforcing anonymity of the shared models, the coordinator is acting as a technical messaging system for data between the participants using the relay server as a communication hub for all participants of a workflow. The relay server implements basic relay functionality for star-based federations of clients. It knows the role of each client (i.e. participant or coordinator) and treats their traffic accordingly. If data is received from a participant, it relays it to the

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

coordinator. If it is received from the coordinator, it is broadcast to all clients. A relay server can handle multiple workflows at once. [D.7.2:4.2.2] From this observation alone the coordinator is not a processor and not a (joint) controller, since he is only acting to transport encrypted data between the other participants. Thus it appears that no role in the sense of data protection may be assigned.

The concept of controller and joint controllership has often been expanded by case law and it is often uncertain how the criteria for identifying these actors should be applied in practice (Millard et al. 2019). When interpreted widely the caselaw Fashion ID, taken at its extreme, this ruling may result in every actor that makes the processing of personal data possible qualifying as a joint controller (Bobek, Fashion ID v Verbraucherzentrale 2018, para 74). According to the EDPB, “the overarching criterion for joint controllership to exist is the joint participation of two or more entities in the determination of the purposes and means of a processing operation”. A ‘joint determination’ means (among others) a common decision, which implies that the actors decide together and have a common intention. This could be the case “when there is a mutual benefit arising from the same processing operation [...]” (EDPB 2020, p 17, 18).

Whether there is joint controllership between the coordinator and participants will largely depend on the design of the pre-project phase. By selecting apps and compiling workflows, the coordinator essentially determines the means of processing. By default, the coordinator initiates the data processing by setting up the project and providing the project infrastructure, thus also enabling the project partners to process data and therefore enters in a joint controllership with participants.

6.1.4 Participant

Participants join projects via a unique token they received from the coordinator and choose the data they want to contribute to the analysis. The participants are joint controllers with the coordinator regarding this data [see 6.1.3]. Determination of means and purpose of processing is done by each participant with the coordinator and not with the other participants.

The cooperation with the coordinator resulting from the selection of the concrete data by the participant leads to a joint controllership. Since one participant never determines the purpose and broad means in coordination or cooperation with other participants, no joint controllership amongst the participants should be assumed. In the case of a parameterizing organisation and several data contributors, the ECJ does not assume joint controllership between the data contributors, but assumes a star-shaped joint responsibility - emanating from the parameterizing organisation (Tietosuojavaltuutettu v Jehovan todistajat 2018).

In their joint paper (AEPD, EDPS, 2022) the Spanish data protection agency ('AEPD') and the European Data Protection Supervisor ('EDPS') imply that participants in a federated machine learning system should be qualified as separate controllers in their own right, as they deliberately describe them as "each controller", notably without including the possibility of the existence of a coordinator.

Indeed, in the basic case of participation, no joint controllership should be assumed, since participants determined the means of processing by selecting the training data and the purpose of processing by selecting the projects.

6.1.5 Model user

Model users, doctors in general, apply AI models on data of individual patients. This inference stage is carried out not in the course of the FeatureCloud project, but in the course of future use of the methods, infrastructure and apps developed in FeatureCloud. Nevertheless, this impact assessment

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

also considers the application of the methods, infrastructure and apps developed in FeatureCloud in the future. However, as the specific use will be determined in the future and depends on the individual case, at this stage this can be only carried out in a generic manner from the perspective of the training and not the model application.

6.2 Views of data subjects or their representatives (Art 35 para 9 GDPR)

Data subjects are individuals whose medical data (medical information) is processed during federated machine learning. These individuals may participate in clinical trials as patients or healthy individuals. Data subjects may also be individuals whose medical data is analysed with the help of a (already trained) application. Data subjects whose data is used to train applications will usually differ from those whose data will be analysed by said applications.

When FeatureCloud is used for a specific research project (workflow), this DPIA must be adopted and completed according to the specifics of the individual project, in the course of which the views of the data subjects or their representatives (e.g. patient representative organisations) on the intended processing shall be obtained.

6.3 Involvement of the data protection officers

According to Article 35 (2) of the GDPR, the controller must seek the advice of the data protection officer when carrying out a DPIA. Whether the advice of the data protection officer must be obtained and to what extent the advice obtained from the data protection officer must be followed is subject of discussion. *Trieb*, for example, assumes that the GDPR does not stipulate such an obligation. (Trieb in Knyrim (Ed.), DatKomm Art 35 para 124). Jandt, on the other hand, sees an obligation in the provision, but the provision does not make any statement as to whether the advice of the data protection officer must also be followed and does not provide for a right of veto or similar for the data protection officer (Jandt in Kühling and Buchner (Eds.), DS-GVO/BDSG Art 35 para 18).

If the controller does not agree with the advice (or parts of it) obtained by the DPO, the Article 29 Working Party considers that a (comprehensible) justification for the lack of compliance with the advice should be included in the DPIA report.

When FeatureCloud is used for a specific research project (workflow), this DPIA must be adopted and completed according to the specifics of the individual project, in the course of which the data protection officers of the relevant institutions shall be involved.

7 Applicable Data Protection Law and Legal Admissibility

7.1 Personal Data

The concept of personal data is the starting point for clarifying the question of whether the GDPR is applicable or not. Whenever personal data are processed (within the material scope as laid down in Article 2 and the territorial scope as laid down in Article 3 GDPR), the Regulation is applicable.

Article 4 (1) GDPR defines the term ‘personal data’, as

> “any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

Thus, personal data can be information such as name, address or date of birth, but also information about private and family life (such as marital status, leisure activities, consumer behaviour or dietary habits) as well as professional or economic activities (such as employment or property relations). Other examples often include external characteristics of a person (such as height, weight, skin colour, papillary lines, iris or vein structure), which in turn are contrasted by inner attitudes (such as motives, desires and ideological convictions) (*Klar* and *Kühling* (2018) Art 4 ref. 1 ff).

Especially in the context of medicine, it must be mentioned that the GDPR further defines so called special categories of personal data, also known as sensitive data, in Article 9 (1). These data are subject to a stricter data processing regime than ordinary (non-sensitive) personal data. Health data, as well as genetic and biometric data fall within this special category. It is irrelevant whether the information is true or not, or whether it is only true with a certain statistical probability (Recitals 59 and 65 GDPR).

As the definition in Article 4 (1) GDPR implies, it is necessary that the respective information can be linked to a specific person in order to be able to speak of personal data. Information can be considered to relate to a person when it is about the respective individual, or also, when it is about an object which itself belongs to the individual or relates to it in another way (Article 29 Data Protection Working Party 2007, p 6 ff).

Article 4 (1) further clarifies that only natural persons count as such data subjects. In addition, Recitals 14 and 17 GDPR specify that data subjects must be living human beings, which means that data protection does not extend to deceased persons (Recital 27 GDPR) or to legal persons (Recital 14 GDPR). However, data on deceased persons, especially in the medical domain, might also contain some information relating to living persons. One example is information regarding the relatives of the deceased person, as in the case of hereditary genetic dispositions for certain diseases. Such data fall within the scope of protection of the GDPR, not because it is relating to the deceased person but because it is relating to another person that is still alive (Rothmann, Kastelitz and Rothmund-Burgwall 2022).

The data subject to whom the respective information relates must be identified or at least be identifiable. The identification of a person can be derived directly from the given information or indirectly; that is, if the existing information is not sufficient to unambiguously identify an individual but the concerned subject can still be identified by linking the existing information with additional information or by using additional criteria or means of identification such as those listed in Article 4 (1) GDPR (Ennöckl 2014, p 107 ff).

The most common identifier is a person's name. However, a widely used name may not be sufficient to uniquely identify a person. In such cases, a second piece of information such as the location (address), other specific factors or a specific context are required. Article 4 (1) GDPR also refers to 'identification numbers' and 'online identifiers'; beside IP addresses, these numbers and identifiers can also be codes such as MAC addresses, the 'International Mobile Station Equipment Identity' (IMEI) or company codes such as Apple's 'Unique Device ID' (UDID).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

If the additional information (e.g. the identification number, the online identifier or another specific identifying factor) is stored separately to ensure that no personal reference can be made, the respective data is called pseudonymous data (see subsequent section) (Karg 2015, p 520 ff). According to Article 4 (5) GDPR the term 'pseudonymisation' means that 'the personal data can no longer be attributed to a specific data subject without the use of additional information' and such additional information 'is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'.

Recital 26 GDPR also clarifies that personal data, which have undergone pseudonymisation, 'should be considered to be information on an identifiable natural person' if these data 'could be attributed to a natural person by the use of additional information'. This means that pseudonymised data is still considered personal data if the additional information makes it possible to attribute the data to a natural person.

The GDPR further states that

> "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 GDPR).

In this regard, the European Court of Justice (ECJ) has stated that the reference to a natural person can be assumed even if an allocation is not directly possible but the relevant body has legal means to obtain the additional information required to identify the person (Patrick Breyer v Bundesrepublik Deutschland 2016, para 49). However, the principles of data protection should not apply to anonymous information, i.e. information that does not relate to an identified or identifiable natural person (Recital 26 GDPR). To determine whether a natural person is identifiable, not every theoretical possibility to identify the person must be taken into account but only means reasonably likely to be used to do so. To ascertain whether means are reasonably likely to be used, all objective factors should be taken into account, 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 (Esayas 2015).

From this, it can be concluded that in order to determine whether data is personal data under GDPR a practical, not a theoretical standpoint must be taken. Means reasonably likely to be used, are means that not only exist theoretically but that would be used practically.

In the context of FeatureCloud, this means that a practical assessment must be carried out: From the attack vectors on the anonymity of the data only those are legally relevant that are reasonably likely to be used by an actual attacker in practice. This must be assessed on the basis of objective factors such as the costs and the amount of time required, the required skills, the potential gain and the available technology but also possible technological developments in the future.

In order to assess whether an attack vector is relevant from a legal perspective, attacks that are reasonably unlikely can be ignored. An attack can be considered reasonably unlikely if it cannot be imagined that it will happen in practice in the given context because the attacker will shy away from the required effort.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

For determining whether personal data is processed, in the context of pseudonymisation and anonymisation of the data, it can be important to distinguish between exactly reproducible and not exactly reproducible data. This is relevant for trying to re-identify individuals in the data by reproducing the same data from a known individual. Raw data that cannot be exactly reproduced by repeating the medical examination at a later point, in particular information which is measured where there is some measurement inaccuracy and every measurement leads to a slightly different result and information which changes over time, cannot be reproduced in a way to exactly match the original raw data, which could render such an attack impossible or at least unreasonably unlikely. The case is different if there is data, such as binary data or genetic data in particular, which is (theoretically) exactly reproducible when a new examination or other act of data collection is carried out.

7.1.1 Governance Body

The governance body controls the Global Backend [5.2.5] which especially contains personal data on participants. It does not process any personal or anonymous data of patients. Regarding the possibility of a joint controllership see section 6.1.1.

7.1.2 Developer

As far as the scope of this DPIA is concerned, the developer does not process any personal or anonymous data of patients. Regarding the possibility of a joint controllership see section 6.1.2.

7.1.3 Coordinator

After a local learning operation has been completed by a participant, it sends the local parameters to the coordinator. The coordinator collects these parameters and aggregates them into a common (global) model, which is shared with the participants. The coordinating site is the central aggregation entity in an FL workflow and will process models. Models may memorize details about the training data that are completely unrelated to the intended task (Carlini et. al 2019).

Malicious machine learning (ML) algorithms can create models that are leaking a significant amount of information about their training datasets, even if the adversary has only black-box access to the model (Song et al. 2017).

However, not every attack which is possible in theory immediately leads to a model being considered personal data. As laid down above, according to Recital 26 of the GDPR, in order to determine whether a natural person is identifiable, only those means should be taken into account that are reasonably likely to be used by the controller to identify the natural person directly or indirectly. FeatureCloud presents KPIs to measure how well data security is ensured and privacy leakage is mitigated. [D.2.3:3.2]

It should be emphasised that the only information being shared between hospital platforms are model parameters. This, however, does not guarantee that no private information can be retrieved from this data [Li et al. 2019]. The application developers need to make sure that the information their applications send out is unproblematic and this is also being checked again by a certification authority before the application becomes admitted to the app store. [D.2.3:3.3.3] Section 5.2.7 provides guidance for app developers in this regard.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Since the coordinator also acts as a participant, processing also takes place in that respective role regarding patient data.

7.1.4 Participant

The participant processes the data subject's raw data which typically were collected by the participant to train the local model. From a participant's perspective, the data he uses to train the local model should typically be qualified as personal data. This holds true even if he works on pseudonymous data, as long as it is still possible that the participant may assign the pseudonymised copy to the original dataset. A participant does not process raw data from other participants [D.7.2: 2.2] and regularly does not enter a joint controllership with other participants [6.1.4].

7.2 Lawfulness of Processing

The processing of personal data is subject to a prohibition with reservation of permission. Article 6 (1) GDPR and Article 9 (2) GDPR exhaustively and conclusively list the possible legal permissions for the processing of personal data. There is no hierarchical relationship between these permissive clauses but each of them is assigned an equal status (Kastelitz et al. 2018, Art 6 para 14). In the following, the theoretical foundations of the most relevant legal bases for FeatureCloud will be introduced.

7.2.1 Consent

The legal basis of consent is of central importance in data protection law and can be understood as a normative expression of the principle of informational self-determination (Buchner and Petri 2018, Art 6 para 17). The conditions for a legally valid consent are primarily defined and set out in Articles 4 (11) and 7 GDPR. Consent is therefore any freely given, specific, informed and unambiguous indication of the data subject's wishes by which they signify agreement to the processing of personal data. Thus, the data subjects shall have a real choice, i.e. they must not feel pressured to give their consent or suffer negative consequences if they do not consent. This also means that there shall not be a clear imbalance between the data subjects and the controller, such as in the case of authorities or employers who act as controllers (Article 29 Data Protection Working Party 2017a, p 6 ff). In addition, Article 7 (4) GDPR stipulates that the performance of a contract or provision of a service shall not be made conditional on consent to the processing of personal data, which is not necessary for the performance of the contract. Furthermore, consent shall be obtained for each purpose separately, and the refusal and withdrawal of consent shall be possible at any time without adverse effects for the data subjects (Article 29 Data Protection Working Party 2017a, p 8 ff.)

7.2.2 Performance of a contract

Article 6 (1) (b) GDPR seems to stipulate a matter of course: Data processing, which is necessary in the context of a (pre-)contractual obligatory relationship, must be permitted (Buchner and Petri 2018, Art 6 para 26). In the case of medical treatment, however, this legal basis is usually of no relevance, since it involves the processing of special categories of personal data and therefore in particular Article 9 (2) (h) GDPR is applicable (Buchner and Petri 2018, Art 6 para 54). Article 9 (2) (h) permits the processing of sensitive data for the purposes of preventive or occupational medicine, medical diagnosis, the provision of health or social care or treatment, or the management of health or social care systems and services. However, in order to be admissible, it must be a contract with

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

a health professional on the one hand; on the other hand, the data must be processed by professionals who are subject to professional secrecy (Buchner and Petri 2018, Art 6 para 55).

7.2.3 Further Processing

The principle of purpose limitation set out in Article 5 (1) (b) GDPR is a core component of European data protection law and is also enshrined in primary law in Article 8 (2) CFR. Purpose limitation means that personal data may only be collected for (pre)defined, explicit and legitimate purposes and may not be further processed in a manner incompatible with those purposes (Hötzendorfer/Tschohl/Kastelitz in Knyrim (Ed.), DatKomm Art 5 para 20). However, the second half-sentence of Article 5 (1) point b GDPR already mentions the possibility of (purpose-changing) further processing for compatible purposes (Kastelitz, Hötzendorfer and Tschohl in Knyrim (Ed.), DatKomm Art 6 para 58).

Article 6 (4) GDPR explicitly regulates the further processing of personal data in terms of "secondary use". In this respect, the provision of Article 6 (4) GDPR represents a normative breach of the strict purpose limitation principle and knows two constellations, namely further processing for incompatible purposes and for compatible purposes.

Further processing for incompatible purposes occurs in the case of downstream data processing if its purpose is not compatible with the original purpose of use (Kastelitz, Hötzendorfer and Tschohl in Knyrim (Ed.), DatKomm Art 6 para 60). Such further processing for incompatible purposes is permissible if it is based on data subject's consent or on Union or Member State law (which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1)).

On the other hand, further processing for compatible purposes occurs in the case of downstream data processing if its purpose is compatible with the original purpose of use (Kastelitz, Hötzendorfer and Tschohl in Knyrim (Ed.), DatKomm Art 6 para 61). For the admissibility of such further processing for compatible purposes, the controller must carry out a compatibility test, which according to Article 6 (4) point a-e GDPR includes five elements:

- any link between the purpose of the collection and the purposes of the intended further processing,
- the context in which the personal data were collected, in particular regarding the relationship between data subjects and the controller (in particular the reasonable expectations of data subjects based on their relationship with the controller as to their further use, cf. recital 50 of the GDPR),
- the nature of the personal data, in particular whether data are processed pursuant to Art 9 or Art 10,
- the possible consequences of the intended further processing for the data subjects, and
- the existence of appropriate safeguards, which may include encryption or pseudonymization.

For archiving purposes in the public interest, for scientific or historical research purposes or for statistical purposes, Article 5 (1) point b, third half-sentence GDPR, with reference to Article 89 (1), establishes the (legal) fiction (*praesumptio iuris ac de iure*) (Kotschy, *Die Zu-*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

lässigkeitsvoraussetzungen für Forschungsdatenverarbeitungen nach dem FOG – eine kritische Analyse, in Jahnel (Ed.), Datenschutzrecht. Jahrbuch 2020 (2021), 287), according to which further processing (including of "sensitive data") (Gabauer, *Die Verarbeitung personenbezogener Daten zu wissenschaftlichen Forschungszwecken* (2019), 53) for these purposes is not considered incompatible with the original purposes. Based on the clear wording of Article 5 (1) point b GDPR parts of the literature argue that a compatibility test does not have to be carried out at all (Kastelitz, Hötzendorfer and Tschohl in Knyrim (Ed.), DatKomm Art 6 para 64; Reimer in Sydow (Ed.), DS-GVO Art 5 para 27). Roßnagel points out that the facilitated change of purpose does not result from the general superiority of the four processing purposes. Rather, these specific purposes lead to the fact that the data processing does not typically relate to the person whose data is being processed (Roßnagel in Simitis, Hornung and Spiecker (Eds.), Datenschutzrecht Art 5 (1) para 104). Recital 162 GDPR states this explicitly regarding statistical purposes. For this reason, the fiction shall not apply to all procedures that use scientific, historical or statistical methods, but only to those that aim at non-personal results that are not personal (see also Article 29 Data Protection Working Party 2013, p 28). Roßnagel argues, that while there is a presumption in favour of compatibility of purposes in the case of research as a secondary purpose, a case-by-case examination of compatibility with the purpose of collection must be carried out even for scientific processing (Roßnagel in Simitis, Hornung and Spiecker (Eds.), Datenschutzrecht Art 6 (4) para 41) for it's being indicated by the formulation with the double negative "not be considered", This does not rule out the possibility of a compatibility of purpose, but it's not automatically given. Since FeatureCloud projects necessarily have to be tuned to produce models that are not considered personal, processing based on Article 6 (4) may be well within reach.

It should also be noted that such permitted further processing must take into account the safeguards for the protection of the fundamental rights and freedoms of the data subjects referred to in Article 89 (1) GDPR. Also according to Art 13 (3) and Art 14 (4) the data controller must inform the data subject about the change of purpose. This also applies to changes of purpose that are compatible with the purpose of the collection(Roßnagel in Simitis, Hornung and Spiecker (Eds.), Datenschutzrecht Art 6 (4) para 16). Since the legislator only privileges the processing purposes because it assumes pseudonymous or anonymous processing results, the view that it is necessary to examine in the individual case whether this is also the case is also convincing.

It is disputed in the literature whether the processing for these (deemed) compatible purposes requires a separate legal basis or not (In favour Herbst in Kühling/Buchner (Eds.), DS-GVO BDSG3 Art 5 para 54; other view, against a separate legal basis Roßnagel in Simitis, Hornung and Spiecker (Eds.), Datenschutzrecht Art 5 para 98 f; Kastelitz, Hötzendorfer and Tschohl in Knyrim (Ed.), DatKomm Art 6 DSGVO para 62). However, Recital 50 second sentence of the GDPR suggests that compatible further processing does not require a new or separate legal basis.

From this, the following criteria regarding further processing of data for scientific research purposes can be deduced:

If the further processing of existing data is carried out

a. for scientific research purposes,
b. by the same controller,
c. solely producing output which does not contain personal data,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

d. such purposes are always deemed compatible (see above)
e. the opinion is upheld that further processing does not require a new legal basis (see above)

then it is lawful under Article 6 (4) GDPR.

7.2.4 Research privilege and data protection

Since the scientific research in question is based on the processing of health data (special categories of personal data), Article 9 (2) (j) GDPR will be relevant. Herein is stated that in such cases the processing is permissible if it is based on Union or Member State law, whereby the respective law 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.

To the present project, the provision on the processing of personal data in the context of scientific research, as laid down in Art 5 (1) (b) and Art 89 GDPR, is of particular importance. In general, it should be noted that the freedom of science, like privacy and data protection, is a fundamental right. According to Article 13 of the EU Charter of Fundamental Rights (CFR) 'scientific research shall be free of constraint' and the 'academic freedom shall be respected'. Therefore, in the case of scientific research based on patients' health data, a balancing between the fundamental right to data protection and the freedom of science is required. For the balancing and the principle of proportionality see Art 52 (1) CFR. The rights and freedoms recognised by the CFR may be limited or restricted on the grounds set out in Article 52 (1) CFR. In addition, Article 3 (2) CFR contains restrictions on the freedom of science; these are the prohibition of eugenic practices, the prohibition of using the human body and its parts as a source of financial gain and the prohibition of reproductive cloning of human beings.

The term 'science', which appears in the title of Article 13 CFR, is recognised as the generic term for research (Jarass 2021, Art 13 para 7). However, neither the term 'science' nor the term research is legally defined in the GDPR. According to the European Data Protection Supervisor (EDPS) '[s]cientific research applies the 'scientific method' of observing phenomena, formulating, and testing a hypothesis for those phenomena, and concluding as to the validity of the hypothesis. […] The conduct of research must allow testing of hypotheses, with both the conclusion and the reasoning transparent and open to criticism. Openness and transparency help distinguish between science and pseudo-science.' (EDPS 2020, p 10).

Moreover, Recital 53 GDPR clarifies that scientific research in general may include studies conducted in the public interest in the area of public health (Recital 156 and 157 GDPR). According to Recital 159 GDPR, the definition of processing of personal data for scientific research purposes should be interpreted in a broad manner, including for example technological development and demonstration, fundamental and applied research as well as privately-funded research - Recital 159 GDPR also refers to the Union's objective under Article 179 (1) TFEU of achieving a European Research Area. In its opinion on data protection and scientific research the European Data Protection Supervisor (EDPS) states that 'not only academic researchers but also not-for-profit organisations, governmental institutions or profit-seeking commercial companies can carry out scientific research' (EDPS 2020, p 11).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The GDPR contains several provisions privileging data processing for scientific research purposes (Art 5(1)(b), Art 9 (2)(j) and Art 14(5)(b) GDPR; see also Recitals 53, 156 and 157 GDPR). In particular Art 89 (1) GDPR holds that the processing of personal data for scientific research purposes shall be subject to appropriate technical and organisational measures (like data minimisation and pseudonymisation) to safeguard the rights and freedoms of the data subject.

In addition, Art 89 (2) GDPR states that where personal data are processed for scientific research purposes, Union or Member State law may provide for derogations from the rights referred to in the Art 15 (right of access), Art 16 (right to rectification), Art 18 (right to restriction of processing) and Art 21 (right to object) of the GDPR, in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes (Rothmann et al. 2022, p 197 ff).

7.2.5 Governance Body

The Governance Body has a legitimate interest in processing the personal data of employees of participants pursuant to Art 6 (1) lit f GDPR. In the context of a balancing of interests, the Governance Body's interest in a smooth handling of the participation must be weighed against the employee's interest in the protection of his or her data. In the vast majority of cases, this consideration should be in favour of the Governance Body during an active employment relationship between the participant and its employee, since business contact data is not very sensitive and it is not obvious what legitimate interest the employee has with regards to not being able to be contacted by the Governance Body (or with that respect other participants).

7.2.6 Developer

The developer does not process any personal data within the context of this DPIA and thus does not determine a legal basis.

7.2.7 Coordinator

The coordinator may engage in a joint controllership with participants [6.1.3]. If the Coordinator trains local models in his own right, he additionally assumes the role of a participant.

The coordinator will process non-personal data in the form of the model parameters it receives from the participants in the course of federated learning workflows and will engage in a joint controllership with each individual participant. It is important to note that joint controllership does not constitute a legal basis for processing by several controllers on the one hand, nor does it require a legal basis for several controllers to join together on the other hand. Insofar as a particular controller within the scope of the joint controllership processes personal data, this particular controller requires its own legal basis for this processing operation (DSK 2018, p 1). This is also supported by the wording of Articles 6 and 9 GDPR which clearly state that a legal basis is required for the "processing" of personal data. The term "processing" is defined by Article 4 (2) GDPR as follows:

> *"...any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction."*

Notably this list – although demonstrative – only includes holding the data and actions with the data, i.e. operations with very direct impact on data, and no actions that may have only indirect effect on the data such as the determination of the purposes of the data processing actually carried out by another party. To summarise, a precise analysis leads to the conclusion that the GDPR ties the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

requirement of a legal basis to actually processing personal data and not to being in the role of (joint) controller.

A coordinator who enters into a joint controllership with a participant but does not process any personal data itself, as described above, is therefore not required by the GDPR to have a legal basis. Only each participant, for actually processing the data, needs to have a legal basis in accordance with the GDPR as described above, regardless of the presence of joint controllership. To repeat, this conclusion only holds true if the model parameters sent from the participants to the coordinator do not contain personal data (e.g., data leakage through the model parameters must be prevented; see e.g., Song et al. 2017; 7.1.3) so that the coordinator does not process personal data itself.

7.2.8 Participant

The participant engages in the processing of personal data in the course of learning the local model and in this regard needs to determine a legal basis. The entire range of legal bases of the GDPR is available to the participant. Subject to national legislative acts, the legal basis will presumably be provided by further processing for a compatible purpose (secondary use) with regard to data collected in the context of a treatment contract (7.2.3) or by obtaining consent (7.2.1).

7.3 Automated decisions (Art 22 GDPR)

Art 22 GDPR regulates the permissibility of automated decisions in individual cases, including profiling.

Art 22 GDPR reads:

Automated individual decision-making, including profiling

(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

(2) Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests;

or

(c) is based on the data subject's explicit consent.

(3) In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

(4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

The provision is structured in such a way that para 1 establishes a prohibition, para 2 includes exceptions to this prohibition and para 4 counter-exceptions, which in turn lead to the applicability of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

the prohibition. Para 3 specifies certain legal consequences for cases in which the prohibition does not apply due to certain exceptions and automated decision-making is therefore permitted.

Art 22 GDPR does not subject every automated individual decision to a legal consequence per se. An automated individual decision is only covered by Art 22 GDPR if it entails legal effects concerning the data subject or similarly significantly affects the data subject.

A distinction must therefore be made between three elements of the scenario:

- There is a decision in the sense of Art 22 GDPR
- The decision is based solely on automated processing.
- The decision has legal or other significant effects on the data subject.

7.3.1 Presence of an automated decision

Recital 71 GDPR gives the following typical examples of such decisions: "automatic refusal of an online credit application" or "e-recruiting practices without any human intervention". These are processes that are traditionally decided without automated decision-making by a human being in the form of a more or less structured decision-making process involving several decision-making factors. This also applies to medical procedures such as anamnesis and diagnosis which traditionally aren't automated. Data subjects whose data is used to train the model are not subject to an automatic decision as a result of this process. However, patients whose personal data is processed in the course of the model's application for inference may very well be subject to an automated decision, such as whether to undergo treatment or further examination. Furthermore, it could be argued that there is no decision within the meaning of Art 22 GDPR here, because an authentication process does not involve the assessment of personal aspects of the data subject referred to in Recital 71 GDPR but rather objective facts. However, this constituent element cannot be found in Art 22 GDPR. The interpretation that this constituent element must be present in all cases and that Art 22 GDPR should be teleologically reduced in this respect is conceivable (Buchner 2018, Art 22 para 19), but by no means mandatory. At least as plausible is the interpretation that Recital 71 mentions the classic scenario that Art 22 GDPR is intended to regulate with the assessment of personal aspects of the data subject, without wanting to reduce Art 22 GDPR to this scenario. This is supported by the fact that the current Art 22 GDPR still contained this element of assessment in earlier drafts - as did Art 15 GDPR - but this restriction of the provision has been removed in the final version (Buchner 2018, Art 22 para 17).

According to *Haidinger*, the term "decision" in Art 22 GDPR is to be understood broadly (Haidinger 2018, Art 22 para 18) and includes "measures" according to Recital 71 GDPR. If one considers the wording in the sense of common parlance, what constitutes an inference process is at the core of the term decision: Based on the criteria contained in the model, a decision is made as to whether the patient data entered fulfils the criteria or not.

7.3.2 Is the decision based solely on automated processing?

This question can only be answered in the context of a specific project or specific data processing. In principle, the answer will depend on whether human expertise is interposed between the decision of the model and an action or treatment based on this decision.

7.3.3 Is there a decision with legal or other significant effect?

This element requires that a decision based solely on automated processing affects the rights of a person. It may also affect the legal status of a person or their rights under a contract.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Even if a decision-making process does not affect the rights of an individual, it may still fall within the scope of Article 22 GDPR if it has such an effect or significantly affects the individual in a similar way. In other words, even if the data subject's rights or obligations do not change, they may be sufficiently affected to require the protection of this provision.
In the present context, the effect of the decision is that due to an incorrect assignment a health measure is taken or rather not taken and it can therefore have other significant effect in many cases if the system does not assign the patient an affliction although one exists and therefore no measures are taken (false negative). It is also conceivable to cite the reverse case here, where the assessment concludes the presence of a disease although there is none present and the patient and potentially risk-bearing health measures (false positive). Such a decision can have far-reaching health consequences, can therefore significantly affect the person and thus Art 22 may apply during inference in the medical field.

7.3.4 Exemptions

For the reasons stated above, all three of the above-mentioned elements of Art 22 (1) are to be considered fulfilled, at least in case of doubt. The exceptions under Art 22 (2) GDPR may be fulfilled:

- Whether Article 22 (2) (b) GDPR is fulfilled can only be examined in the specific case in accordance with the national legal system. The existence of such a legal basis is at least not immediately apparent.
- Article 22 (2) (a) GDPR may be fulfilled if the use of the model is based on a contract: Whether the processing of a treatment contract necessarily requires the model to be applied is conceivable, but can only be examined on the basis of a specific case.

- Finally, the permissibility of the decision can also be based on Article 22 (2) lit c GDPR if the data subject's express consent to automated decision-making is obtained.

According to Art 22 (3) GDPR, in the cases referred to in (2) (a) and (c), the controller must “implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision." Art 22 (4) GDPR standardises a restriction for automated decisions based on special categories of personal data within the meaning of Art 9 (1) GDPR: Decisions based on such data are only permitted if

a) the data subject has expressly consented to this processing of special categories of personal data (Art 9 (2) (a) GDPR)

or

b) such processing is necessary for reasons of substantial public interest on the basis of Union 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 data subject's fundamental rights and interests (Art 9 (2) (g) GDPR).

As for FeatureCloud, it is foreseeable that in many cases processing will be based on special categories of personal data (health data) and Art 22 (4) GDPR thus shall apply.

7.3.5 Conclusion

The applicability of Article 22 (4) GDPR can be justified if the following occurs: The model user decides to centrally include the processing of health data of a data subject during model inference. This may have significant effects for the data subject in the sense explained in more detail above. If the model user relies on this - without human intervention in relation to the individual case - and has made the decision to link model outputs to further form of treatment, it is probably impossible to come to any other conclusion in terms of the telos and, in particular, the protective purpose of Art 22 GDPR

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

than to attribute the entire process, including the processing of health data, to the controller in its entirety, at least for the assessment under Art 22 GDPR.

The two variants of Art 22 (4) GDPR are each special cases of two of the three above-mentioned exceptions to Art 22 (2) GDPR (consent, Art 22 (2) (c) GDPR or legal provision, Art 22 (2) (b) GDPR), whereby the second case of paragraph 4 leg. cit. contains significantly more specific requirements than Art 22 (2) (b) GDPR (Art 22 (2) (a) GDPR, contract is excluded). Therefore, the following conclusion on paragraph 4 also already covers paragraph 2:
As no legal provision providing for the processing of health data during model inference is apparent in relation to the use of FeatureCloud in a medical setting and such a provision that fulfils the requirements of Art 9 (2) (g) GDPR, in particular the requirement of necessity on grounds of substantial public interest, does not appear realistic *de lege ferenda*, it is recommended to obtain the explicit consent of the data subject regarding automated decision-making based on the processing of health data, which fulfils the requirements of Art 22 (2) and (4) GDPR as well as Art 7 and Art 9 (2) (a) GDPR and thus also the requirement of voluntary consent.

In addition, as explained above, appropriate measures must be taken to protect the rights and freedoms as well as the legitimate interests of the data subject according to Art 22 (2) and (4) GDPR, including at least the right to obtain the intervention of a person on the part of the controller, to express his or her point of view and to contest the decision Art 22 (2) GDPR.

8 AI-specific Regulation

FeatureCloud is at its core about applying machine learning methods on medical data. In the course of the project, due to the recent progress in AI development and the increased public attention, the issue of regulating AI by law has gained momentum, in particular on the EU level.

8.1 AI Act – general remarks

The use of AI systems can pose risks to the safety, health and fundamental rights of the affected persons. For example as regards medical research, the selection and maintenance of the data basis is one of several factors which determines whether the system is discriminatory against certain groups of persons (Lekadir et al. 2022, 20). The EU aims to address these risks with a package of measures. The proposal for the AI Act is one part of this package, its main content being product safety rules for the placing on the market, putting into service and use of high-risk AI systems.

The AI Act aims to promote the “uptake of human centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and rule of law, and the environment from harmful effects of artificial intelligence systems in the Union while supporting innovation (Art 1 para 1 as amended by the EP). By limiting the scope to a few actual high-risk applications, the aim is not to undermine the technology's innovation potential in areas such as medicine. For low-risk applications, the draft AI Act therefore contains no regulations. For some that interact with natural persons (e.g. chatbots, deepfakes), there is only a transparency requirement.

The AI Act is a horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI. It aims not to unduly constrain or hinder technological development or otherwise disproportionately increase the cost of placing AI solutions on the market. It complements existing and forthcoming EU safety regulation, following the logic of the New Legislative Framework (NLF; a common EU approach to the regulation of certain products in the form of “controlled self-regulation” through a so-called conformity assessment procedure carried out by the manufacturer; essential obligations are prescribed by law and concretized through [harmonised] standards), and incorporating the AI Act into the NLF legislation on product safety require-

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

ments as a horizontal standard with provisions on material obligations, market monitoring and surveillance and conformity assessments (Veale and Zuiderveen Borgesius 2021, 102). Additional horizontal and sectoral rules on AI and algorithmic decision making can be found in other legislative acts of the EU.

The following section first describes the core provisions of the AI Act and examines whether it is applicable to medical research in general and FeatureCloud in particular. Then, indications for a legally compliant use of artificial intelligence in the context of the research project are derived.

The AI Act is still under discussion in the EU legislative process. Unless expressly stated otherwise, the article designations and descriptions refer to the proposal of the European Commission (EC COM/2021/206). The amendments of the European Parliament (EP COM/2021/0206 – C9-0146/2021 – 2021/0106(COD)) and the Council (COM/2021/206 - 5698/22, 2021/0106(COD)) are taken into account especially as regards the definition of AI and the scope of exceptions for research and open source components.

8.2 Definition of AI

The definition of AI in Art 3 is supposedly technology-neutral. The EC proposal listed a number of techniques in Annex I that were to be part of the definition. The most recent version of the definition in the EP amendments however is aligned with the OECD's definition (OECD 2019), which is increasingly used internationally, to ensure consistency with international instruments. Accordingly, ‘‘artificial intelligence system’ (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments”. Essential for the definition is the aspect of autonomy, which is primarily intended to cover so-called black-box systems, i.e. complicated neural network models, whose results are not very transparent or explainable. However, highly complex expert systems can also be covered if they influence their environment with a certain degree of autonomy. The definition shall not cover any type of software but be limited to technology that bears a certain level of risk for the safety, health and fundamental rights of individuals (Ebers 2021, 590).

8.3 Risk categories

The AI Act introduces four risk categories which – similar to the purpose limitation principle in the GDPR - relate to the purpose of the respective AI system. The decisive factor in determining whether a system is subject to the AI Act is therefore not an abstract risk, but rather the risk in context of a specific purpose the system is intended to serve.

1. Prohibited systems: A few applications that pose unacceptable risks to society and democracy fall into the first risk category of prohibited systems (“Prohibited artificial intelligence practices”, Art 5), e.g. “the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person’s consciousness in order to materially distort a person’s behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm”.

2. High risk AI systems: The second category, high-risk AI systems, are defined in Art 6 and specifically enumerated in two annexes (Annex II and III). The classification as high risk shall be limited to those systems that have a significant harmful impact on the health, safety and fundamental rights of persons in the Union**.**

As regards Annex II, Art 6 (1) contains the following provision:

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

"Irrespective of whether an AI system is placed on the market or put into service independently from the products referred to in points (a) and (b), that AI system shall be considered high-risk where both of the following conditions are fulfilled:

(a) the AI system is intended to be used as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II;

(b) the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II."

Thus, AI systems that are intended to be used as a safety component of a product listed in Annex II or that are themselves such a product, and which must undergo third-party conformity assessment for health and safety risks under these regulations, are considered high-risk. The material provisions of the AI Act shall be considered and assessed during the conformity assessment procedure defined in the respective *lex specialis* regulation. Annex II lists 19 EU harmonizing acts that regulate high risk products or applications. It lists in its Section A the Regulation (EU) 2017/745 on Medical Devices (MDR), and Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR).

In addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall also be considered high-risk (Art 6 (2)). In the Council version an additional restriction was added if the output of the system "is purely accessory in respect of the relevant action or decision to be taken and is not therefore likely to lead to a significant risk to the health, safety or fundamental rights". Similarly, in the EP version these systems are only considered as high-risk AI systems "if they pose a significant risk of harm to the health, safety or fundamental rights of natural persons". Accordingly, Annex III lists some specific purposes from eight areas of application that are also considered to be high-risk. The material provisions of the AI Act apply however not to the respective area as a whole, but to specific purposes pursued by means of an AI system covered by one of these eight areas. In the EC proposal 21 such purposes are listed.

Medical research, medical or health applications are not listed in Annex III. An AI system in the medical area is only covered by the AI Act as a high-risk system insofar as it is "intended to be used as a safety component of a product, or is itself a product" covered by the two Medical devices regulations listed in Annex II (cited above).

Most material provisions of the AI Act only apply to high-risk AI systems.

3. Low risk AI systems (transparency requirements): Art 52 imposes transparency obligations on certain AI systems. This third risk category includes AI systems intended to interact with natural persons (natural persons shall be informed that they are interacting with an AI system unless this is obvious from the circumstances and the context of use), users of an emotion recognition system or a biometric categorisation system and users of an AI system that generates or manipulates image, audio or video content (deepfakes).

4. No or minimal risk AI systems: For all other AI systems, the EC proposal for AI Act imposes no obligations.

The EP proposes to include general principles applicable to all AI systems in the AI Act (Art 4a). These principles follow the recommendations of the EU's High-Level Expert Group on Artificial Intelligence (AI HLEG) and refer to human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness and social and environmental well-being. For high-risk AI systems, their principles are reflected in the requirements of Art 8-15, which will be discussed in the following subsection.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

8.4 Material provisions

High risk AI systems should only be placed on the Union market or put into service if they comply with certain mandatory requirements, which should ensure that they do not pose certain unacceptable risks.

Title III, Chapter 2 contains minimum product safety requirements with regard to high-risk AI systems (Art 8 – 15). They are complemented by obligations in the following chapters and titles. These obligations include, among others

- Risk management system: The establishment, implementation, documentation and maintenance of a risk management system in relation to the high-risk AI system (Art 9);
- Data and data governance: High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet certain quality criteria; among others, training, validation and testing data sets shall be relevant, representative, free of errors and complete and avoid bias (Art 10);
- Technical documentation: The technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to date; the documentation shall demonstrate that the high-risk AI system complies with the requirements of this Chapter and shall serve as a basis for the conformity assessment (Art 11 and Annex IV);
- Record keeping: High-risk AI systems shall be designed and developed with capabilities enabling the automatic recording of events ('logs') while the high-risk AI system is operating. Those logging capabilities shall conform to recognised standards or common specifications. The logs shall ensure, among others, traceability of the AI system's functioning throughout its lifecycle and the monitoring of its operation (Art 12);
- Transparency and provision of information to users: Designers and developers shall ensure that a high-risk AI system's operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately (Art 13) [see in more detail D4.8];
- Human oversight: High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use (Art 14);
- Accuracy, robustness and cybersecurity: High risk AI systems shall achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle (Art 15);
- Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with the AI Act (Art 17);
- Providers shall ensure that the high-risk AI system undergoes the relevant conformity assessment procedure, prior to its placing on the market or putting into service (Art 19);
- Providers take the necessary corrective actions, if the high-risk AI system is not in conformity with the requirements (Art 21);
- Document retention: Technical documentation and additional documentation shall be kept for a period of 10 years (Art 50);
- Registration: Providers shall register a high-risk AI system in the EU database referred to in Article 60 (Art 51)
- Providers shall report any serious incident or any malfunctioning of those systems which constitutes a breach of obligations under Union law intended to protect fundamental rights (Art 62);

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- The EP proposed to include, additionally, an obligation for deployers (users) to carry out a fundamental rights impact assessment prior to putting a (stand-alone) high-risk AI system into use (Art 29a).

8.5 Scope / Applicability

The AI Act applies to 1) providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country, 2) users (deployers) of AI systems located within the Union and 3) providers and users (deployers) of AI systems that are located in a third country, where the output produced by the system is used in the Union (Art 2). Title III Chapter 3 defines the obligations for providers, users (deployers), product manufacturers, importers, distributors and other third parties. Most material provisions refer to the providers of high-risk AI systems.

To delimit the AI systems the material provisions of the AI Act apply to, the definition of AI must be read together with the definition and delineation of high-risk systems. The substantive requirements of the AI Act in Title III ("High-Risk AI Systems," Art 6 to Art 51) apply only to the comparatively few applications listed in Annex II and III. An AI system is considered to be high risk if

- it is intended to be used as a safety component of a product, or is itself a product and the intended use falls within EU harmonisation legislation listed in Annex II and a third-party conformity assessment is required pursuant to the aforementioned EU legislation OR
- AI systems that are intended to be used in areas covered by Annex III of the AI Act are automatically classified as high risk (Karathanasis 2023).

According to Art 43 (3) of the AI Act, the providers of high-risk AI systems, to which legal acts listed in Annex II, section A, apply, shall follow the relevant conformity assessment as required under those legal acts (e.g. Art 52 MDR). The requirements set out in Chapter 2 of this Title (see above) shall apply to those high-risk AI systems and shall be part of that assessment (Points 4.3., 4.4., 4.5. and the fifth paragraph of point 4.6 of Annex VII shall also apply).

Opting out of a conformity assessment according to the legal acts cited in Annex II section A is possible only if the manufacturer has (additionally to the required standards therein) also applied harmonised standards or, where applicable, common specifications referred to in Article 41, covering the requirements set out in Chapter 2 of Title III (Art 43 (3) *sub-para* 3 AI Act). Such standards and specifications that cover the requirements of Chapter 2 of Title III are still to be defined.

8.6 Applicability of the AI Act to the area of health and medical applications

According to Recital 28 of the AI Act “AI systems could produce adverse outcomes to health and safety of persons, in particular when such systems operate as components of products. Consistently with the objectives of Union harmonisation legislation to facilitate the free movement of products in the internal market and to ensure that only safe and otherwise compliant products find their way into the market, it is important that the safety risks that may be generated by a product as a whole due to its digital components, including AI systems, are duly prevented and mitigated.”

Accordingly, Annex II lists the MDR and the IVDR, which are therefore relevant to the scope of the AI Act as regards high risk AI systems. Art 2 MDR defines a ‘medical device’ as “any instrument, apparatus, appliance, *software*, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of” certain specific medical purposes listed in this article (e.g. diagnosis, prevention, monitoring, prediction, prognosis,

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

treatment or alleviation of disease, injury or disability). Not included are devices that achieve their principal intended action by pharmacological, immunological or metabolic means, in or on the human body.

Artificial intelligence falls under the definition of software: An interpretation of the guidelines (MEDDEV 2016; MDCG 2019) and of the aims of the MDR provides a software definition that includes AI driven systems (Kiseleva 2020, 11, see also Niemiec 2020, 3 on the risk classification of AI devices). Software "intended to provide information which is used to take decisions with diagnosis or therapeutic purposes" or "to monitor physiological processes" (Annex VIII 6.3 Rule 11 MDR) is at least classified as a medical device class IIa necessitating a conformity assessment procedure including a third party, a so-called notified body. At the moment only static "frozen" AI systems without online-learning capabilities are certifiable (IG NB 2022).

The MDR however only covers medical devices and software with an intended medical purpose and specifically excludes software for general purposes, even when used in a healthcare setting, or software intended for life-style and well-being purposes (Recital 19 MDR). Thus health-related AI applications (such as those used to track medication) and administrative AI systems used by doctors in hospitals are not in scope of the MDR (Bogucki 2022, 4). Apart from AI systems falling under the scope of the MDR, the healthcare sector and other health related applications (health apps, chatbots and apps generating customized dietary recommendations) are absent from the list of high-risk areas (Kolfschooten 2022, 26 seq). This limited alignment between health-related rules and the proposed AI Act has led to demands during the committee phase of the legislative process in the EP to include certain AI systems used in the area of healthcare, but not covered by the Regulation on Medical Devices, to be considered as high risk in the AI Act. This was justified by the fact that software impacting diagnostics, treatments or medical prescriptions and access to health insurance can significantly affect health and safety (Bogucki 2022, 21). However, the proposals were not fully included in the EP amendments.

8.7 Research and Open-Source Exceptions

The EP proposed an amendment aiming to exclude research from the scope of the AI Act. The exception refers to "AI systems specifically developed for the sole purpose of scientific research and development" and aims "to ensure that the Regulation does not otherwise affect scientific research and development activity on AI systems. Under all circumstances, any research and development activity should be carried out in accordance with the Charter, Union law as well as the national law" (Recital 2f). Accordingly, the AI Act "shall not apply to research, testing and development activities regarding an AI system prior to this system being placed on the market or put into service, provided that these activities are conducted respecting fundamental rights and the applicable Union law. The testing in real world conditions shall not be covered by this exemption". The EC shall be authorized to adopt delegated acts that clarify the application of this paragraph (Art 2 (5d)). Another exception proposed by the EP refers to open-source components. The AI Act "shall not apply to AI components provided under free and open-source licences except to the extent they are placed on the market or put into service by a provider as part of a high-risk AI system or of an AI system that falls under Title II or IV. This exemption shall not apply to foundation models as defined in Art 3" (Art 2 (5e)).

8.8 Conclusion

Provided the final version of the AI Act retains these amendments and no other material changes are added in the final stage of the legislative process (which is expected to be concluded by the end of 2023), the following can be said about the scope of application of the AI Act:

- The MDR and IVDR are covered by Annex II of the AI Act, the rules on high-risk AI systems shall apply to the areas covered by these regulations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Other health related use of AI systems outside the scope of the MDR or IVDR is not considered high-risk (e.g. software for general purposes, even when used in a healthcare setting, or software intended for life-style and well-being purposes).
- AI systems specifically developed for the sole purpose of scientific research and development are excluded from the scope of the AI Act prior to such systems being placed on the market or put into service.
- The AI Act shall not apply to AI components provided under free and open-source licences except to the extent they are placed on the market or put into service by a provider as part of a high-risk AI system or of an AI system that falls under the prohibited or the low-risk categories.

In summary, FeatureCloud as a research project (or similar research projects the AI Act would be temporally applicable to) would not be covered by the AI Act. Research, testing and development with the tools provided would also not be covered. However, if an AI system is developed with the tools provided as a result of the project that falls under the MDR or IVDR and that is placed on the market or put into service, the obligations of the AI Act for high-risk AI systems shall have to be complied with.

8.9 Recommendations for a legally compliant use of artificial intelligence

Even if (research, development and testing of) an AI system does not fall under the scope of the AI Act, it is advisable to comply with certain material obligations as soon as they are finalized and adopted. Generally speaking, the use of AI systems in the field of medical research is certainly associated with high risks to the health and fundamental rights of individuals due to the use of highly sensitive health data, regardless of the applicability of the Medical Devices Regulation.

Especially with a view to research that can result in products or services being placed on the market or put into service, it is essential to not only follow a strict approach to ethical principles and accountability, but also conduct research and development with the greatest diligence and adequately take into account standards regarding data selection and data governance, risk assessment and risk mitigation, robustness and documentation. The measures implemented during the research, development and testing phase shall enable future users to carry out fundamental rights and data protection impact assessments, to inform affected individuals transparently and to monitor compliance with safety, security and fundamental rights throughout the AI system's life cycle. The documentation on data and data governance, technical aspects, intended use of the AI system, as well as the record keeping (logging) and post-market monitoring requirements should meet the highest possible standards.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9 Risk Analysis

Risk analysis has been a major aspect of the work in the FeatureCloud project from the beginning. This includes the continuous risk management process as an important part of project management (WP1) on the one hand as well as several risk assessment tasks and related work with various emphases and objectives in several work packages on the other hand. These include:

- D1.3 Report on risk assessment and details on measures to prevent misuse of research findings
- D2.1 Risk assessment methodology: Fundamentals of risk assessment, ISO 31000:2009
- D2.2 Cyber risk assessment and mitigation
- D6.2 Model for defining user rights in federated machine learning: Threats identified in chapter 5.1
- D10.1 POPD – Requirement No. 2

The following risk analysis, apart from newly identifying some specific risks, gathers and summarises this work in a structured manner and builds upon it with the primary objective of providing a compiled risk analysis document for future deployment of FeatureCloud. For further details regarding specific aspects consider the deliverables listed above, e.g. D2.1 regarding methodological fundamentals and details.

9.1 Methodology

The methodology to conduct the present impact assessment, which is described in this section, has been developed and is being continuously improved and adapted by the Research Institute - Digital Human Rights Center in several projects in research and practice, based on existing methodological work as referenced below. At its core it aims to fulfil the requirements of Article 35 GDPR in a practical way but at the same time goes significantly beyond mere fulfilment of these fundamental requirements.

As a first step, it needs to be established which fundamental rights and freedoms might be impacted or are at risk and how this might occur. This phase of the impact assessment encompasses describing possible risk scenarios and linking them to specific rights and data protection principles. From Recitals 75 and 94 GDPR it can be deduced that a risk to the fundamental rights and freedoms of natural persons is conceptualised as the possibility of an event occurring, which itself represents damage or can lead to further damage to one or more natural persons. In general, the risk analysis is about an estimation and classification of the probability and severity of risks. To estimate the severity of risk scenarios several factors must be considered. These include among others (Kernell et al. 2022, p 22 ff):

- The number of people affected
- The characteristics of the impacted groups
- The geographical and demographical reach
- The extent of adverse effects and their reversibility
- The likelihood of exacerbating existing biases, stereotypes, discrimination and inequalities
- Possible cumulative impacts
- The effort required to minimise the risk (e.g. time spent amending information, extra costs, low or sufficient capacity to remediate the impact, long-term psychological or physical ailments, etc.)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

In the case of FeatureCloud, the assessment furthermore should address AI-specific risk factors, like the dependence of potentially affected persons on AI-based decisions (Mantelero 2022, p 166) [D3.6:3.3.3].

As Recitals 84 and 90 GDPR point out, the risk assessment should consider the origin and nature as well as the scope, context and purposes of the respective data processing. The origin and nature of the risks can be distinguished by the following criteria (Bitkom 2018, p 27):

- Internal/external human source or internal/external non-human source: e.g. internal or external employee, software error or hardware defect, environmental impact (natural forces), cybercriminal (hacker/malware), state institutions (intelligence service, law enforcement), management.
- Intentional, negligent, or unintentional: e.g. the damage to the affected person can be either condoned or intended and the goal of action, or due to individual or structural errors.

Furthermore, a distinction can be made between physical, material and non-material types of damage (Recital 75 GDPR). Typical risk causes include unauthorised or unlawful processing, processing contrary to good faith, processing that is not transparent for the data subjects, unauthorised disclosure of and access to data, unintentional loss, destruction, or damage of data, denial of data subjects' rights, use of data by controllers for illegitimate purposes, not intended processing of data, processing of inaccurate data, incorrect processing (technical failures, human errors), processing beyond the retention period, processing itself when the harm lies in the performance of the processing (e.g. because it is illegitimate/lacks a legal basis) and processing contrary to the purpose limitation principle (Recitals 75 and 83 GDPR).

Based on these considerations, the following overarching questions may guide the assessment of a risk scenario (Martin et al. 2020, p 43):

- What damage can occur for data subjects based on the planned data processing?
- Which actions or circumstances can lead to the occurrence of the respective damaging events?
- Which actors or (non-human) risk sources are involved and how?
- Which remedial measures have already been implemented or are planned?

9.2 Assessment of likelihood and severity

The risk assessment is meant to be as objective as possible (Recitals 75 and 76 GDPR). This is, however, not always attainable in practice due to ambiguities about assignable likelihoods, possible types of damage, and the subjective perceptions of risk by the various stakeholders.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

According to the GDPR, risks should be assessed by combining the likelihood (probability or chance) of occurrence (or happening) and the severity (or magnitude) of consequences (Recital 76 GDPR). Therefore, the risk evaluation is an estimation of the likelihood of a threat scenario materialising and an estimation of the impact severity of each risk scenario (Vemou and Karyda 2020, p 48). These two variables are combined in the combinatorial matrix using an ordinal scaled system to estimate the overall risk level. The decision which risk level an expected impact corresponds to, lies with the evaluation of experts who possess knowledge of the system in question, of case law, literature, and the relevant legal framework.

The exact methodology for assessing the impact varies from author to author. However, most models work with locating the respective likelihood and severity on a risk matrix, using an ordinal scale like (1) negligible, (2) limited, (3) substantial and (4) maximum.

In order to systematically determine the **probability**, the following statements can be attributed to each of the risk levels (Mantelero 2022, p 56) [D3.6:3.3.4]:

Negligible	It is very improbable that the damage occurs; it hardly seems possible for the selected risk sources to materialise the threat under the given circumstances (e.g. theft of paper documents stored in a room protected by a badge reader and access code).
Limited	It is rather improbable that the damage occurs; it seems difficult for the selected risk sources to materialise the threat under the given circumstances (e.g. theft of paper documents stored in a room protected by a badge reader).
Substantial	It is rather probable that the damage occurs; it seems possible for the selected risk sources to materialise the threat under the given circumstances (e.g. theft of paper documents stored in offices that cannot be accessed without first checking in at the reception).
Maximum	It is very probable or even inevitable that the damage occurs; it seems easy for the selected risk sources to materialise the threat under the given circumstances (e.g. theft of paper documents stored in the public lobby).

The same goes for determining the level of **severity** (CNIL 2018, p 4):

Negligible	The severity of the damage is very low; affected individuals and groups may encounter a few inconveniences, which they will overcome without any problem (e.g. time spent amending information, annoyances, irritations, etc.).
Limited	The severity of the damage is rather low; affected individuals and groups may encounter inconveniences, which they will be able to overcome despite a few difficulties (e.g. extra costs, fear, lack of understanding, stress, minor physical ailments, etc.).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Substantial	The damage is rather severe; affected individuals and groups may encounter consequences, which they should be able to overcome albeit with real and serious difficulties (e.g. economic loss, property damage, worsening of health, etc.).
Maximum	The damage is very severe; affected individuals and groups may encounter serious or even irreversible consequences, which they may not overcome (e.g. long-term psychological or physical ailments, death, etc.).

The subsequent combination of these two variables results in the following risk matrix (Bitkom 2017, p 32) [D3.6:Tab. 3.2]:

Risk assessment		**Probability of occurrence**			
		negligible (1)	*limited (2)*	*substantial (3)*	*maximum (4)*
Severity of damage	*maximum (4)*	normal	high	very high	very high
	substantial (3)	low	normal	high	very high
	limited (2)	very low	low	normal	high
	negligible (1)	very low	very low	low	normal

In accordance with Art 35 (1) and Art 36 (1) GDPR, the risk acceptance level is defined as normal or below. Risk mitigation measures are, however, considered regarding all risks. According to the risk assessment prescribed by Articles 24, 25 and 32 GDPR, depending on the available technical and/or organisational measures to mitigate a particular risk among other factors, risks with a level of normal, low, or very low can be considered to be acceptable. Risk scenarios located in the range of high or very high of the matrix, on the other hand, always require further risk treatment until the respective risks are sufficiently contained. If high risks would nonetheless remain, either the data protection supervisory authority must be consulted (see Art 36 GDPR) or the processing must not be carried out at all.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.2.1 Appraisal of proportionality

Next to determining the risk level, a DPIA should include the examination of the proportionality of the infringement of the fundamental rights and freedoms in question. The principle of proportionality (and the related step of examining the necessity) can be understood as a doctrinal tool for the resolution of conflicts between two competing rights or interests (Möller 2012, p 10).

The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights (EDPS 2017, p 4). In this spirit, Article 35 (7) (b) GDPR stipulates an assessment of the necessity and proportionality of the processing operations in relation to the purpose, in case the operations interfere with other fundamental rights (Jandt 2018, Art 35 para 39 ff). Also, according to the European Data Protection Supervisor (EDPS), the appraisal of necessity and proportionality is an essential requirement with which any proposed measure that involves processing of personal data must comply (EDPS, 2019b, p 3). For the present impact assessment this appraisal of proportionality is of particular importance since the underlying methodical approach is not limited to the protection of personal data but extends to other fundamental and human rights as well.

Contrary to the several existing methodologies for risk assessments, approaches for assessing the proportionality and necessity in the context of personal data protection are rather scarce (Kloza et al. 2020, p 29). In accordance with Article 52 CFR and the relevant legal literature on the methodological implementation of the proportionality principle, the following steps or criteria can be differentiated (Möller 2012, p 711 ff; EDPS 2017, p 4 ff; EDPS 2019b, 6 ff):

- Legality: Is the legal basis for the data processing provided for by a law of a sufficient quality (e.g. clarity, accessibility, precision, foreseeability, conformity with the rule of law), and does this legal basis respect the essence of the fundamental rights and freedoms?
- Legitimacy: Does the envisaged data processing operation serve a legitimate aim or meet objectives of general interest to protect the fundamental rights and freedoms of others?
- Suitability: Is the envisaged data processing operation appropriate (even capable) to achieve the given legitimate aim?
- Necessity: Is the envisaged data processing operation necessary to achieve this legitimate aim?
- Proportionality (sensu stricto): Is the interference with the right proportionate (balanced) with regard to the protection of the competing right or interest?

In its related guidelines the EDPS points out that the examination of necessity (just as the entire DPIA) is a facts-based process, rather than a merely abstract legal notion. It must therefore be considered in light of the specific circumstances surrounding the use-case in question as well as the concrete purpose it aims to achieve. This means that the respective data protection operation should be genuinely effective to achieve the pursued objective of general interest (EDPS 2017, p 8).

In addition, the envisaged data protection operation should be the least intrusive for the fundamental right at stake. Consequently, the assessor needs to consider alternative measures which are comparably effective but with less impact on e.g. the protection of personal data or the right to respect of private life (Jandt 2018, Art 35 para 39). Only if existing or less intrusive measures are not available

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

and the envisaged data processing operation is essential and limited to what is absolutely necessary to achieve the objective of general interest, the criterion of necessity is met (EDPS 2017, p 17).

At the core of the appraisal process lies the concept of a balancing. This final step is a procedure of weighing up the intensity of the interference against the legitimacy (or importance) of the objective pursued in the given context. The balancing (of advantages/disadvantages and benefits/costs) should lead to the decision whether the data processing operation in question is proportionate or not. If the conclusion is that it is not proportionate, the assessor should make sure to take all factors which determine the appraisal as disproportionate into account and determine and introduce (if possible) safeguards which render the data processing activity proportionate (EDPS 2019b, p 11).

9.2.2 Risk treatment and mitigation

Once the potential risks and proportionality of the data protection operation in question were identified and analysed, the controller (assessor) has to decide, whether to deploy the respective data processing operation without changes, to modify the system, context or processing operation or to cancel the data processing altogether (Kloza et al. 2020, p 37).

A common procedure consists of elaborating measures to mitigate the expected adverse impact. Article 35 (7) (d) GDPR only requires mentioning these measures; the actual implementation lies outside the impact assessment and constitutes a separate process (Jandt 2018, Art 35 para 47 ff). Mitigation measures can be of a regulatory (legal), technical, organisational or behavioural nature (Kloza et al. 2020, p 30). They should particularly address the general data protection principles and obligations such as the data protection by design and by default approach (Jandt 2018, Art 35 para 47 ff.).

The treatment of risks and the implementation of measures to mitigate or control the identified risks can take several forms (EDPS 2019a, p 16 f):

- Prevention: amendments to the processing operation or technology to prevent or avoid risks from materialising
- Detection: (self-)monitoring (logging) of processing operations in order to ensure to quickly notice breaches or illicit use
- Repression: quickly ending detected breaches; procedures to correct inaccurate data; certificate revocation mechanisms to stop the use of compromised credentials
- Correction: undoing or limiting the damage after the fact; measures to restore or revert to the status/conditions that existed prior to the impact

Different types of control measures and policies as well as privacy enhancing technologies can be considered (Vemou and Karyda 2020, p 48 f) [D3.6:3.6]. For example, regarding the security of the processing Article 32 (1) GDPR refers to the following technical and organisational measures to be implemented, as appropriate:

- Pseudonymisation and encryption of personal data
- Ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services
- Ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Implementation of a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing

In Article 32 (4) GDPR reference is further made to measures of access restriction or access control. The various measures, safeguards and procedures are ultimately intended to ensure the protection of personal data (and other fundamental rights) and to demonstrate compliance with the provisions of the GDPR (Recital 90 GDPR).

The order of addressing the identified risk scenarios is guided by their previously prescribed risk level, not by their proximity (Kernell and Veiberg 2020, p 11).

As a bare minimum, the mitigation measures should reduce all risks rated as 'high' or "very high" to the point where they can be classified as 'normal'. It is not always necessary to implement additional measures; sometimes it might be more feasible to strengthen already existing measures (Martin et al. 2020, p 48).

Besides specifying the necessary mitigation measures it is also recommended to develop a response plan to prioritise the envisaged steps, name the person responsible and determine a date (or time plan) for the implementation (Kloza et al. 2020, p 35). In the context of new technologies such as the FeatureCloud system, it is crucial to consider mitigation measures early in its lifecycle. For example, addressing human rights already in the design process of a product ('human rights by design') can effectively prevent future negative impacts (Kernell and Veiberg 2020, p 23).

For those remaining cases and residuals where the risk cannot be mitigated despite the measures taken, the assessor once more has to decide how to proceed (Kloza et al. 2020, p 37). The follow options are available (Bitkom 2017, p 33):

- Risk acceptance: in case the level of risk in terms of likelihood and severity is low enough
- Risk reduction: through the implementation of further mitigation measures; defining their priority and their practical implementation
- Risk avoidance: by completely refraining from and stopping the activity if no appropriate measures can be implemented

In theory, risk transfer constitutes a fourth option. It primarily refers to the risks of companies (corporate bodies). In the case of data subjects as natural persons, this option is not always feasible. For the sake of completeness, this option is nevertheless listed.

If high risks remain, additional measures must be selected until the respective risks are sufficiently contained. Otherwise, either the data protection supervisory authority must be consulted (see Art 36 GDPR) or the processing must be stopped entirely. This final decision on how to react or treat the risk lies with the controller.

In order to carry out the risk treatment systematically, it is therefore necessary to go through the following questions [D3.6:3.6]:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- What treatment or procedure (acceptance, reduction, avoidance) was chosen to address the identified risk scenario? Justify the chosen treatment/approach.
- If a reduction of the estimated risk level is envisaged, what measures are being implemented to mitigate the risks? Describe the mitigation measures related to the specific risk scenario and how they will be implemented to address the identified risk. Design a response plan to prioritise the envisaged steps, identify responsible persons (in charge) and set a date (or timetable) for the implementation.
- Does the impact assessment indicate that the data processing will result in a high risk? Is it necessary to consult the data protection authority?

9.2.3 Revisiting and monitoring

Finally, it must be mentioned that assessing a digital activity's/product's/service's risks to data protection, privacy and other human rights is not just a one-off but continues throughout its lifecycle. For instance, it might have to be decided whether and when to perform the DPIA process again (entirely or partly) once the envisaged processing operations have been deployed (Kloza et al. 2020, p 40). According to Article 35 (11) GDPR the controller must perform a review of the DPIA process where necessary; at least when there is a change in the risk represented by the processing operations (i.e. if the nature, scope, context or purpose of the processing operations or the relevant law have changed) and hence so has the level of risk. This step serves as quality control and preventive measure to avoid future risks (Jandt 2018, Art 35 para 59 f; see also Kloza et al. 2020, p 42) [D3.6:3.7].

To that end, a monitoring process should be implemented. Key questions include (Kernell and Veiberg 2020, p 24):

- What exactly (which specific data processing operation) will be monitored?
- When will the monitoring activities begin?
- How often and at what intervals will the monitoring activities occur?
- Who will conduct the monitoring activities (e.g. internal or external persons)?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.2.4 Risk assessment template

Risk title

1) Risk identification	**Risk description**
	What scenario are we facing? What is the risk? *Add description of the respective data processing operation and explanation of the possible risk scenario (naming of actors and persons involved; type and characteristics of the processing; naming of processed data categories; naming of existing technical or organisational measures, etc.)*
	Risk source
	What elements trigger the occurrence of the damage? Is it a human or non-human risk source?
	Risk cause
	What leads to the 'realisation of the risk'?
	Possible damage for data subjects
	Physical damage: Bodily damage by incorrect medical treatment, psychological damage like anxiety or depression, in the case of breaches of confidentiality also violent crimes (incl. stalking) etc. Material damage: Economic damages, occupational damages such as loss of employment or promotion; reduction of state benefits, discrimination (e.g. in insurance contracts or apartment search), unjustified fees or fines, etc. Non-material damage: Social and societal disadvantages (e.g. damage to reputation, mobbing etc.), damage to privacy, intimidation effects (so-called chilling effects, where people refrain from exercising their rights out of fear) etc.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(1-4): *Add risk level [negligible (1), limited (2), substantial (3), maximum (4)] and explain (see Chapter 4.1.1 for guidance on determining the adequate risk level).* Comment: Optional explanation of the classification	(1-4): *Add risk level [negligible (1), limited (2), substantial (3), maximum (4)] and explain (see Chapter 4.1.1 for guidance on determining the adequate risk level).* Comment: Optional explanation of the classification	(1-16): *Estimate the level of risk on the basis of the given facts; assign the risk to a value/level in the present matrix*

3) Measures	Existing measures
	If it is envisaged to reduce/mitigate the identified risk: What regulatory, technical, organisational or behavioural measures need to be implemented to reduce the risk? *Add list of selected envisaged/future mitigation measures, including a detailed description and explanation how they address the risks and disproportionality of the processing operations identified to protect the rights and freedoms of the data subjects and to demonstrate compliance with law.*

	Probability of occurrence	Severity of damage	Risk assessment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4) Risk analysis and evaluation (after mitigation)	(1-4): *Add risk level [negligible (1), limited (2), substantial (3), maximum (4)] and explain (see Chapter 4.1.1 for guidance on determining the adequate risk level).* Comment: Optional explanation effect of the measures	(1-4): *Add risk level [negligible (1), limited (2), substantial (3), maximum (4)] and explain (see Chapter 4.1.1 for guidance on determining the adequate risk level).* Comment: Optional explanation effect of the measures	(1-16): *Estimate the level of risk on the basis of the given facts; assign the risk to a value/level in the present matrix*

5) Risk treatment / Proportionality of risk / Future measures	Proportionality of risk / Future measures / Permanent measures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3 Identified risks

9.3.1 Misidentification of risks

1) Risk identification	Risk description
	The misidentification of risks to individual rights and freedoms caused by not carrying out an appropriate risk assessment. As a consequence, an organisation cannot put in place appropriate technical and organisational measures to prevent putting individuals in danger.
	Risk source
	• Participant / Coordinator
	Risk cause
	• Failure to identify risks and place appropriate measures.
	Possible damage for data subjects
	• Failure to implement appropriate measures may (indirectly) lead to manifestation of all of the risks addressed in this DPIA.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4) Comment: Without assessing risks, chances are very high that appropriate measures won't be implemented.	**(3)** Comment: Misidentification does not have an immediate effect.	**(12)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

3) Measures	Measures
	• Risk analysis and management throughout the project. • Highest due diligence in risk identification, creative and broad approach when imagining potential risks, using established methods and experience/examples as well as "creative dystopian thinking". • 6-eyes principle on multidisciplinary senior expert level. • All project participants carried out risk identification and assessment and were participating in the overall risk management process. • The present DPIA has been carried out throughout the project, partially identifying new risks, partially collecting risks from the other risk analysis processes in the project. • Transparency of residual risks.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2) Comment: It is still conceivable that risks remain undetected or that documented measures are not fully implemented.	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.2 Lack of responsibility and possibility of intervention in a federated setting

1) Risk identification	Risk description
	Accountability refers to the obligation of organisations to take responsibility for the data they collect, process, and use. In the context of federated machine learning, where data is distributed across multiple parties, ensuring accountability becomes more complex. While the federated approach of FeatureCloud at its core decreases data protection risks, the federated setting at the same time increases some risks in terms of data protection law, as each participant is potentially a single point of failure. Participants are enabled to participate in FeatureCloud projects via token. Apart from removal from the project - by withdrawing the token - coordinators have only limited possibilities to oversee, influence and control data processing by the participants. The coordinator relies on participants to contribute as much high-quality data as possible, which is why there is a certain relationship of dependence in this direction. By this, the coordinator may be guided not to immediately withdraw system access from a participant acting in misconduct, especially if the participant provides high-quality data.
	Risk source
	• Participant / Coordinator
	Risk cause
	• Unclear/ambiguous role allocation • Coordinator has sole responsibility regarding which participants he invites and has a potential relationship of dependence to them • Failure to identify or implement adequate measures by a single actor.
	Possible damage for data subjects
	• Failure to implement appropriate measures and or to exclude participants from projects, when necessary, may (indirectly) lead to manifestation of several of the risks addressed in this DPIA.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Closed number of possible responsible parties.	**(3)** Comment: Materialisation of this risk does not have an immediate effect.	**(9)**

3) Measures	Measures
	● Closed circle of recipients, invitation by token. ● The coordinator is in the relationship of joint controllership with every participant (see section 6.1.) and has to conclude an adequate joint controllership agreement with every participant. ● The Blockchain-based mechanism for logging and auditing of data usage developed in WP6 is able to detect some forms of misconduct of participants (see section 5.1.6). Both components of the blockchain-based auditing mechanism developed in WP6 must be actually utilised, i.e. (1) participants must be contractually obliged to use the logging mechanism, that logs which data has been used for which purpose by whom and (2) actual audits must be carried out.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.3 Risks originating from (sub-)processors

1) Risk identification	Risk description
	(Sub)-processors may compromise data privacy and security on a technical level or may not be aware of or compliant with all the relevant regulations and standards for data protection. E.g. (sub-)processors may process data for their own purposes without having a legal basis to do so. In particular, it seems conceivable that participants may use (sub-)processors to calculate local FeatureCloud models.
	Risk source
	• Participant / Coordinator that make use of a processor to calculate and emanate local models
	Risk cause
	• Properties of or relationship with the processor (e.g. inadequate security measures, level of protection in the country where the processor is based or unfavourable contractual relationship).
	Possible damage for data subjects
	• Bodily damage: Individuals may be exposed to incorrect decisions (e.g. due to data poisoning by Sub-Processor) • Material damage: Restrictions on the conclusion of contracts when health data is disclosed (e.g. Adverse effects in insurance contracts) • Non-material damage: Social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Attack vector is restricted to certain parties.	**(4)** Comment: Damage to the body is possible. There is also a risk of becoming too dependent on a (sub-) processor, making it more difficult to switch if e.g. the processor refuses to work in a legally compliant manner, which solidifies the damage.	**(12)**

3) Measures	Measures
	• Minimise the need/involvement of (sub-)processors: The FeatureCloud system is designed in order to run locally at any participant involved and that the local components can be provided to each participant by FeatureCloud. Therefore, no (sub-)processors need to be involved. • The project consortium provides information on how to set up projects locally. • Highest scrutiny in selection of (sub-)processors must be applied. • Ensure and document clear contractual measures that stipulate who has access to the training data, training code, and deployment code, and when they have access. • Document clear audit trails of how personal data is moved and stored from one location to another during the training and testing phase. • FeatureCloud uses Docker as a virtualization technique. Docker offers the appropriate level of isolation, preventing Internet and file access if not explicitly granted, and sandboxing to limit the usage of compute and memory resources if necessary. These isolated running environments (containers) are created from pre-defined images, which are the federated apps in our case. • The training results must always be critically reviewed by a natural person before they are applied to patients.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2) Comment: The measures above significantly decrease the likelihood that (sub-)processors are involved at all.	**(3)** Comment: Human review of results prevents immediate application of results to humans and therefore prevents severe damage.	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.4 Dilution of data protection awareness

1) Risk identification	Risk description
	The promise of privacy by architecture and privacy enhancing techniques might lead to relaxed self-assessment of privacy concerns by patients and participants. Firstly, this might be exploited by attackers with regard to participants. Secondly it might be exploited by participants regarding data subjects when collecting consent.
	Risk source
	• Participant / Coordinator
	Risk cause
	• Unchecked confidence in architectural or technical security. E.g.: While SMPC provides input privacy and allows protecting the privacy of intermediate results, it reveals the final result - the output of the function. In federated learning, the output can be also potentially sensitive and vulnerable to inference attacks. Another technique needs to be applied to ensure output privacy as a complement to SMPC, e.g., Differential Privacy.
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. Adverse effects in insurance contracts based on unlawful disclosure). • Non-material damage: Social and societal disadvantages (e.g. damage to reputation based on unlawful disclosure).

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Risk cause is restricted to certain parties.	**(3)** Comment: Materialisation of this risk does not have an immediate effect.	**(9)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

3) Measures	Measures
	● Data protection and privacy are taken very seriously in FeatureCloud irrespective of the fact that the architecture is already privacy-enhancing and several different measures are taken in this regard, in particular ○ WP2 - Cyber risk assessment and mitigation ○ the present DPIA, ○ the app certification mechanism, ○ the logging and auditing mechanism, ○ very specific data protection and security measures on the implementation level and the ○ measures and recommendations in the deployment manual for the different stakeholders in Annex I. ● Transparency with regard to limitations of privacy-enhancing architecture and techniques towards participants and towards data subjects ● Open-source implementation ● Logging and auditability of data usage

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.5 Failure to comply with individual rights

1) Risk identification	Risk description
	The failure to respond adequately to rights requests (Articles 15-22 GDPR) is caused by a lack of awareness that data subject rights apply throughout the lifecycle of an AI system wherever personal data is used. Peculiar problems arise as it may be technically challenging to identify specific personal data in a model to meet requests. This is especially necessary within the scope of the right to access, right to erasure and right of rectification. In particular it might destroy a model if specific data need to be deleted from the model (Edwards and Veale 2017, p 67 ff). However, since the aim is for the model to contain no personal data anyway, the latter should not be a problem in the context of FeatureCloud.
	Risk source
	• Participant / Coordinator
	Risk cause
	• Lack of awareness or adequate processes to fulfil data subject rights • Impossibility to identify data subject's data in a model • Lack of resources to process requests.
	Possible damage for data subjects
	• Non-material damage: Individuals are denied informational self-determination (lose control over how their personal data is used).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Identification of data is technologically demanding. It is conceivable that, for example, an increased number of requests is received due to a media report.	**(2)** Comment: No material / bodily damage	**(6)**

3) Measures	Measures
	● FeatureCloud avoids this risk by aiming at producing only models which do not contain any personal data; this is ensured by ○ the implemented privacy-enhancing techniques (SMPC and DP) and ○ the app certification process. ● If identification of specific personal data in a model is not (directly) possible, Art 11 GDPR stipulates that no additional data must be processed solely in order to comply with data subject's rights. ● The standard processes for complying with data subject's rights lie with the participants which originally collected the data ● Joint controllers shall clearly assign responsibility for information and transparency.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(2)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.6 Unlawful processing

1) Risk identification	Risk description
	Failing to choose an appropriate lawful basis causes the unlawful collection of personal data. The excessive and irrelevant collection of personal data can be caused by a default approach to collect as much data as possible to produce accurate models (Zarsky 2017, p1006 ff.). As a consequence, individuals suffer from unlawful and unfair processing of their personal data. If processing is based on consent, coordinators may include data as a basis for training where consent has expired or was already revoked. On the other hand, unlawful processing can also be caused by implementation of ever new FeatureCloud workflows using already collected data without checking for a legal basis or not defining what purpose a particular use of data shall fulfil. On a wider scale, individuals lose control over how their data is used, become uninformed and lose trust in the organisation handling their personal data or in the medical system as a whole.
	Risk source
	● Participant / Coordinator
	Risk cause
	● Processing more data than is strictly necessary, not determining a legal basis or basing processing on expired consent.
	Possible damage for data subjects
	Unlawful processing of personal data, in particular if acting lawfully would have prevented storing and processing the data at all, may lead to the manifestation in particular of: ● Non-material damage: Social and societal disadvantages (e.g. damage to reputation based on specific diagnosis). Data subjects lose trust over how their data is used, suffer from unfair processing and restrictions on informational self-determination.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

	• Material damage: Restrictions on the conclusion of contracts (e.g. Adverse effects in insurance contracts).

2) Risk analysis and evaluation (before mitigation)	**Probability of occurrence**	**Severity of damage**	**Risk assessment**
	(3) Comment: An additional risk only arises when acting lawfully would have prevented storing and processing the data at all.	**(3)**	**(16)**

3) Measures	**Measures**

	• The key measure implemented to prevent this risk is the blockchain-based mechanism for logging and auditing of data usage [5.2.6]
	• In addition the following recommendations shall be followed, which are contained in the Deployment Manual in Annex I:
	  ○ Consultation with domain experts to ensure that the data to be included in a project is appropriate and adequate.
	  ○ Documentation of purpose(s) for using personal data at each stage of the processing lifecycle. Assessment whether they are compatible with the originally defined purpose, and schedule reviews for reassessment.
	  ○ Documentation of the data collected to train the system. Assessment whether it is accurate, adequate, relevant, and limited to the specified purpose(s).
	  ○ Reassessment and documentation of what data is necessary, adequate, and relevant for training and testing the system. Consideration of the trade-off between data minimisation and statistical accuracy.

4) Risk analysis and evaluation (after mitigation)	**Probability of occurrence**	**Severity of damage**	**Risk assessment**
	(2)	**(3)**	**(6)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.7 Lack of transparency

1) Risk identification	Risk description
	A lack of transparency, interpretability and/or explainability may be caused by the application ecosystem of FeatureCloud. The coordinator sets up a research project and chooses which applications shall be relevant. This may lead to data subjects having a lack of understanding about how their data is being used and how FeatureCloud affects them, for the specific means processing are not constant but may be dynamically adapted by the coordinator. Thus there can be no uniform information given on the workings of research via FeatureCloud. Rather the coordinator (and/or participant) has to provide specific information suitable to the project. If no adequate information is provided, individuals cannot exercise their rights and may feel disempowered to object to decisions or processing.
	Risk source
	• Participant / coordinator
	Risk cause
	• Insufficient information and/or understating of data use by coordinator and/or participants.
	Possible damage for data subjects
	• Non-material damage: Individuals are denied informational self-determination (lose understanding over how their personal data is used).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3)	**(2)** Comment: No material / Bodily damage	**(6)**

3) Measures	Measures
	● The following recommendations shall be followed, which are contained in the Deployment Manual in Annex I: ○ Participants must be informed about their duty to provide precise information on means and purposes of processing. ○ Joint controller agreements shall clearly state which party is responsible for providing information. Most importantly, the coordinator as the party being in control of the means of processing must enable the participants to inform the data subjects about the processing.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(2)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.8 Pressure regarding consent/consent revocation

1) Risk identification	Risk description
	Given the medical context, a data subject could likely be in a very desperate situation due to its medical condition and consent obtained from this data subject could be influenced by this/not freely given/not given in an informed manner, in particular if the data subject has the impression that consenting is necessary to receive optimal treatment or at least optimal attention by the doctor. Patients could feel a pressure to give their consent and/or not to withdraw it later in order to improve their relationship with their treating physician, on whom their life may depend.
	Risk source
	• Participant
	Risk cause
	• Actual or felt pressure to consent or not to withdraw consent in order to maintain the best chance for getting cured.
	Possible damage for data subjects
	• Non-material damage: Processing of personal data against the will of the data subject. Due to limited, insufficient or missing voluntariness of consent, unlawful data processing occurs. Opening up the potential for one of the other risks described.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3)	**(2)** Comment: Degree of the infringement of fundamental rights is not very high, as research purposes	**(6)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

		are pursued and results are planned to be anonymous.	

3) Measures	Measures
	• Information, in particular that consenting or not does not influence the treatment in any way. • Revocation unobservability: The attending physician/doctor treating the data subject or the doctor who recruited the data subject for a study should not be able to know that the data subject withdrew consent. This in turn also mitigates the harm of pressure to consent in the first place as consent can truly be withdrawn freely. The blockchain-based consent management solution is designed in order to provide revocation unobservability [D6.5:6.2.1], but it also depends on the individual implementation in the hospital, in particular if the consent management is paper-based there.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(2)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.9 Breach of integrity/availability of the model

1) Risk identification	Risk description
	An attacker manipulates training data (e.g. editing, inserting or removing data instances) to change the model's behaviour, e.g. dropping the model's accuracy or achieving a particular misclassification. One can distinguish the following types of security attacks on machine learning models [D2.1:8.3]: Poisoning attacks can be related to the data tempering category and also can cause a denial of a service, when the machine learning model, for instance, gets corrupted and gives (primarily) false predictions. Evasion attacks include scenarios when an attacker feeds the network with adversarial input to reach the goal of e.g. (targeted) misclassification or confidence reduction. Applying certain perturbation (e. g. specific pixels to sample images) to the input can cause the network to misclassify. An insider attacker (participant or coordinator) participates in the federated learning process and has access to the models during training which is why coordinators and participants pose an increased potential to cause harm. An outsider attacker has basically access only to the final model after the federated learning process is finished. Depending on whether the consequences of the attack are discovered, the outcome could either be that the model is useless and the learning effort is frustrated (breach of availability), or that the compromised model is applied, leading to wrong predictions (breach of integrity). The following analysis focuses on the latter, as this affects individuals.
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Poisoning attacks [D2.1: 8.3] • Evasion Attacks (e. g. adding specific pixels to sample images) [D2.1: 8.3]
	Possible damage for data subjects

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

	• Physical damage: Bodily damage by incorrect medical treatment due to an unavailable model • Non-material damage: Social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis)

2) Risk analysis and evaluation (before mitigation)	**Probability of occurrence**	**Severity of damage**	**Risk assessment**
	(3) Comment: Neural networks are especially vulnerable to this type of attacks, as it is harder to interpret these types of models, and due to their behaviour to overfit, they are further more likely to learn the backdoor pattern. Even when an adversary can manipulate only one participant of the training, the attack still can be successful. Yet access to training data has to be obtained prior to the relevant training stage.	**(3)** Comment: Materialisation of this risk does not have an immediate effect thus can be overcome, yet indirect effect may be severe (e.g. wrong treatment).	**(9)**

3) Measures	**Measures**
	• Filtering methods for the input (Y. Liu, Xie, and Srivastava 2017) [D2.1:9.2.1] • Pruning the network (K. Liu, Dolan-Gavitt, and Garg 2018) [D2.1:9.2.1] • Modifying training samples, model structure or combining the model with other models (Papernot et al. 2016) [D2.1:9.2.1] • Use of simpler machine learning methods less susceptible to such attacks and/or make such attacks being recognized more easily. • Coordinator can technically exclude participants by revoking their token.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	(3)	(6)

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.10 Membership inference attacks

1) Risk identification	Risk description
	The membership inference attack (Shokri et al. 2017) refers to the scenario when an adversary has a sample record of a form of training set data, and a "black-box" access to the model. The attacker can then infer if this record was in the training set of the model, or not, which can reveal certain meta-data about the individual, e.g. if the training set was on a study about a certain disease, the membership inference attack could reveal that the individual has this disease. Membership inference can be categorised (using LINDDUN notation) as detectability threat or disclosure of information threat. [D2.1:8.4]
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Successful membership inference attack [D.2.1: 8.4]
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts). • Non-material damage: Mental health problems because of data disclosure, Social and societal disadvantages (e.g. damage to reputation based on data disclosure).

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4)	**(3)**	**(12)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

3) Measures	Measures
	● The key measure to prevent this risk is the certification process for apps [5.2.3]. App certification ensures on a case-by-case basis that effective measures are in place to prevent this. ● Such measures can, inter alia, be the following, which are contained in the implementation manual: ○ Use of privacy enhancing techniques (Differential privacy or other noise addition methods) [D2.1:9.2.2], [D2.5:5.4.3] ○ SMPC (mitigate against coordinator as attacker) [D2.1:9.2.2], [D2.5:5.4.3] ○ Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7] ○ Data anonymisation [D2.5:5.4.1] ○ Black-box only access to model [D2.1:8.4] ○ Reduction of information content of model output ○ Adversarial Regularization, Early Stopping (Tang et. al (2021) ○ Use of a simpler model

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.11 Model Inversion attacks

1) Risk identification	Risk description
	In the model inversion attack (Fredrikson, Jha and Ristenpart 2015), an adversary tries to recreate data samples that represent the underlying original objects. This has been shown to work in very specific settings, such as in the case of recreating pictures of the people to be identified by a facial recognition system. It is more difficult to achieve in other settings, where an individual does not correspond to one of the classes distinguished by the machine learning system. Model inversion is related to the identifiability threat in LINDDUN. [D2.1:8.4]
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Successful Model inversion attack [D.2.1:8.4]
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts). • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4)	**(3)**	**(12)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

3) Measures	Measures
	● In the present setting, one individual does not correspond to one of the classes distinguished (which would facilitate this attack). ● The key measure to prevent this risk is the certification process for apps [5.2.3]. App certification ensures on a case-by-case basis that effective measures are in place to prevent this. ● Such measures can, inter alia, be the following, which are contained in the Deployment Manual in Annex I: ○ Use of privacy enhancing techniques (Differential privacy or other noise addition methods) [D2.1:9.2.2], [D2.5:5.4.3] ○ SMPC (mitigate against coordinator as attacker) [D2.1:9.2.2], [D2.5:5.4.3] ○ Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7] ○ Data anonymisation [D2.5:5.4.1] ○ Black-box only access to model [D2.1:8.4] ○ Reduction of information content of model output

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.12 Property inference attacks

1) Risk identification	Risk description
	Property inference attacks can be performed on machine learning models to infer information about training data. Specifically, global properties of the training data can be inferred from the model (Ganju et al. 2018). In (Ateniese et al. 2015), the authors showed how to infer statistical properties of the training data, by comparing the difference of the model before and after training on this data. Machine learning models also can leak users' private information when the adversary has access to their public data. (Weinsberg et al. 2012) showed how to infer the gender of a user from a recommendation system, based on ratings which the user has given. Using LINDDUN categorisation the attack can cause threats like detectability, disclosure of information or identifiability. [D2.1:8.4]
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Successful Property inference attacks [D.2.1: 8.4]
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts). • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4)	**(3)**	**(12)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

3) Measures	Measures
	● The key measure to prevent this risk is the certification process for apps [5.2.3]. App certification ensures on a case-by-case basis that effective measures are in place to prevent this. ● Such measures can, inter alia, be the following, which are contained in the Deployment Manual in Annex I: ○ Use of privacy enhancing techniques (Differential privacy or other noise addition methods) [D2.1:9.2.2], [D2.5:5.4.3] ○ SMPC (mitigate against coordinator as attacker) [D2.1:9.2.2], [D2.5:5.4.3] ○ Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7] ○ Data anonymisation [D2.5:5.4.1] ○ Black-box only access to model [D2.1:8.4] ○ Reduction of information content of model output

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.14 Data exfiltration

1) Risk identification	Risk description
	Data exfiltration via machine learning (ML) models refers to embedding information in the models or model updates, as described e.g. by (Song et al, 2017). It is vital that ML models trained on sensitive inputs (e.g., personal images or documents) not leak (too much) information about the training data. An operator of a machine learning model who supplies model-training code to the data holder, does not observe the training. An adversarial operator might then obtain white- or black-box access to the resulting model. If the algorithm is designed in such a way that it "memorizes" information about the training dataset, the operator can extract that information from the model. This attack is in some way similar to the model inversion attack, just with the differentiation that in this setting, the attacker is able to influence the amount and type of information embedded in the model. The attacker's goal is to let the model appear unsuspicious, i.e. train it to be as accurate and predictive as a conventionally trained model. Data exfiltration attacks therefore can cause a number of privacy threats e.g. disclosure of information or identifiability within LINDDUN categories. According to CVSS frameworks, the metric values are more restricted than previously mentioned attacks, as the attacker in this case is an operator of the model - the attacker requires some privileges and potentially also user interaction, therefore the score would be lower, with values of 5 being reasonable. [D2.1:8.4]
	Risk source
	• Coordinator • Outside attacker
	Risk cause
	• Successful data exfiltration attack [D.2.1: 8.4]
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts). • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4)	**(3)**	**(12)**

3) Measures	Measures
	● The key measure to prevent this risk is the certification process for apps [5.2.3]. App certification ensures on a case-by-case basis that effective measures are in place to prevent this. ● Such measures can, inter alia, be the following, which are contained in the Deployment Manual in Annex I: ○ Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7] ○ Data anonymisation [D2.5:5.4.1] ○ Black-box only access to model [D2.1:8.4] ○ Reduction of information content of model output ○ Use of a simpler model ○ LSB sanitization [D2.5:5.4.4] ○ Sign modification [D2.5:5.4.4] ○ Activation Based Neuron Pruning [D2.5:5.4.4]

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.15 Models leaking information about their training data in another way

1) Risk identification	Risk description
	Privacy risks in federated learning are mostly connected to models leaking information about their training data. If the shared model updates, i.e. the data leaving the institution, are not sufficiently anonymous, attackers may be able to link the updates back to individual users, thereby compromising their privacy [D2.1:8.4]. Specific threats in this regard have already been described in the risks 9.3.10 to 9.3.13 above. In the following, all other and possibly unknown ways to exfiltrate personal data from a local or the global model are covered. Such attacks can originate from outsiders but in particular a malicious coordinator is able to perform more targeted attacks against particular nodes (including attacks like membership inference, model inversion, attribute inference and others) (Orekondy et al. 2019). [D2.1:8.4.1]
	Risk source
	● Participant / Coordinator ● Outside attacker
	Risk cause
	● Successful attack on model (other than those mentioned the 9.3.10 to 9.3.13 above) and thereby caused exfiltration of personal data.
	Possible damage for data subjects
	● Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts) ● Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4)	(3)	(12)

3) Measures	Measures
	● The key measure to prevent this risk is the certification process for apps [5.2.3]. App certification ensures on a case-by-case basis that effective measures are in place to prevent this. ● Such measures can, inter alia, be the following, which are contained in the Deployment Manual in Annex I: ○ Use of privacy enhancing techniques (Differential privacy or other noise addition methods) [D2.1:9.2.2], [D2.5:5.4.3] ○ SMPC (mitigate against coordinator as attacker) [D2.1:9.2.2], [D2.5:5.4.3] ○ Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7] ○ Data anonymisation [D2.5:5.4.1] ○ Black-box only access to model [D2.1:8.4] ○ Reduction of information content of model output ○ Use of a simpler model

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	(3)	(6)

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.16 Differential privacy breaches

1) Risk identification	Risk description
	Differential privacy techniques may be employed to preserve user privacy. However, if these techniques are not applied correctly or effectively, the desired privacy level might not be achieved, resulting in potential data exposure. Exemplarily the choice of privacy parameter Epsilon (ε) controls the level of noise added to the data to protect individual privacy. With larger epsilon values less noise is added to the data, resulting in weaker privacy protection but higher utility, as the data remains more accurate and useful for analysis. The choice of an actual privacy level by a data steward in regard to her business requirements is a non-trivial task (Lee et al. 2011). If epsilon is too large, the level of noise added may not be sufficient to protect privacy.
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Execution of malicious applications
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts) • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: The risk is only relevant in a setting where differential privacy is already in place as a measure.	**(3)**	**(9)**

3) Measures	Measures
	• Optimal choice of privacy parameter Epsilon (ε) [D.2.1:9.2.2] Guidance in this regard is provided by the FeatureCloud governance body (https://featurecloud.ai/assets/developer_documentation/privacy_preserving_techniques.html#parameter-guide-anchor) • Empirically perform attacks to test the setting

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.17 Data leakage in distributed systems

1) Risk identification	Risk description
	As the federated FeatureCloud system is a form of distributed system, the associated attacks (which may not be specific to federated learning) may materialise. High-level threats include attacks such as eavesdropping or masquerading. One specific threat in the FeatureCloud approach is that of an attacker that manages to have a malicious application executed at the remote sites. [D2.1:7]
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Eavesdropping (attack that tries to listen to private communication of other parties, without their consent) • Masquerading (attacker pretends to be an authorised user of a system to gain access to it. It can use stolen passwords and logins to gain unauthorised access through a legitimate access identification)
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts) • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(4)	**(3)**	**(12)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

3) Measures	Measures
	• Strong authentication and authorisation mechanisms • Encryption of data storage and communication • Minimisation of data exchange and use of privacy enhancing techniques (DP, SMPC) as described in D2.1:9 • Architectural Measures: FeatureCloud is designed in a way that neither the coordinator nor a participant can access personal data of other participants. • New participants must be actively informed that neither participants nor the coordinator need to obtain access to raw data at any stage and that asking for raw data or local models outside the predefined communication channels of the FeatureCloud platform is to be considered fraudulent (anti-phishing training).

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.18 Denial of Service in distributed systems

1) Risk identification	Risk description
	Denial of service attacks (DoS) aim to reduce the availability of a system or other network resource and are designed to make a machine or network resource unavailable to its intended users (Hansman and Hunt 2005). Different targets can be distinguished. E.g. an individual user might be addressed, by deliberately entering a wrong password repeatedly to cause the victims account to be locked. Further, whole systems might be the target of the attack, trying to overload the capabilities of a machine or network to answer requests and thus to block all users at once. Attacks from a single source can relatively easily be identified and defended against, by e.g. blocking that source. Especially powerful are distributed denial of service attacks, where the attack comes from a larger number of attackers, and it is thus more difficult to handle all attackers. Such an attack is often performed using botnets, or by attacks that fool innocent systems into sending traffic to the target. [D2.1:7]
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Successful denial of service attack
	Possible damage for data subjects
	• Non-material damage: Data subjects lose transparency and trust over how their data is used, suffer from unfair processing and restrictions on informational self-determination, in particular when DOS leads to (temporary) inability to withdraw consent.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3)	**(2)** Comment: Materialisation of this risk does not have an immediate effect thus can be overcome, and can only occur indirectly in extreme cases.	**(6)**

3) Measures	Measures
	The following measures shall be implemented and have been documented in the deployment manual in Annex I: ● The damage can be overcome, in particular the data subject can use other means of communication to withdraw consent, which must be provided anyway (Art 13 (1) (a) GDPR).

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(2)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.19 Data leakage through malicious app

1) Risk identification	Risk description
	Apps may be uploaded to the FeatureCloud "App Store". These Apps are used by coordinators to set up projects. Bad actors may upload malicious apps. Such a malicious app could either exfiltrate personal data bluntly in the form of a covert channel (cf. Zander, Armitage and Branch 2007), or could more subtly leak some personal data as part of its legitimate output which might be difficult to grasp. This would mean to hide the information to be transmitted along with the lawful communication about the model updates, and being able to extract that information afterwards. This can be seen as a form of steganography, which is a form of information hiding that conceals the existence of the secret data hidden in a cover medium (the model updates). [D2.1:7]
	Risk source
	• Outside attacker (app developer) • Coordinator
	Risk cause
	• Execution of malicious applications which exfiltrate or leak data
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts) • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: While anyone can upload an app to the FeatureCloud "App store", for the risk to manifest, a coordinator has to make use of the malicious app image.	**(3)** Comment: Mere upload of a malicious app does not have an immediate effect thus can be overcome, yet the indirect effect may be severe.	**(9)**

3) Measures	Measures
	• Certification process of apps and clear labelling of certified apps [5.2.3]. • Enforcement of certification as a strict requirement for practical use. • Multiple reviews per app. • Apply least significant bit pruning. • Automatic monitoring of the amount of data communicated over the network connection (O-Notation; sub-linear exchange quota) [D2.2:3.3.3], meeting KPI 3 [D2.5:5.2]. • Execution of FeatureCloud-Apps within a docker container (virtualisation).

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.20 Data leakage at the local site (participant)

1) Risk identification	**Risk description**
	An attacker could find another way (than through uploading a malicious app to FeatureCloud "App Store", see above) to execute malicious code at the remote sites or is otherwise able to obtain unauthorised access to sensitive information there and is able to leak personal data.
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Execution of malicious code • Identity disclosure, also known as re-identification, is typically acknowledged as the most potent form of disclosure. This entails an attacker's ability to link an individual directly to a specific record. • Attribute disclosure enables an attacker to discover (precisely or approximately) the value of one or more attributes associated with an individual present in the targeted dataset. For instance, with some background knowledge on the individual, an attacker could learn details such as the medical diagnosis or salary of a person in the dataset. • Using membership disclosure, an attacker could, for instance, deduce whether an individual is part of a dataset by linking data from multiple sources.
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts) • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3)	(3)	(9)

3) Measures	Measures
	• Execution of FeatureCloud-Apps within a docker container (virtualisation). In addition the following recommendations shall be followed, which are contained in the Deployment Manual in Annex I: • Strong authentication and authorisation mechanisms, encryption of data storage and communication, and minimisation of data exchange. [D2.1:9] • Data fingerprinting [D2.5:5.4.2] • General IT security best practises and four eyes principle has to be adhered to within the IT of the participant. As a guidance while eliciting and to achieve a large coverage of potential threats, we employ NIST cybersecurity framework (https://www.nist.gov/cyberframework/framework), OWASP (https://owasp.org/www-project-top-ten/#), LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1). • Use of privacy enhancing techniques (Input Differential privacy or other noise addition methods to the original data) [D2.5:5.4.6] • Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7] • Data anonymisation techniques [D2.5:5.4.1]

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	(3)	(6)

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.21 Risks emanating from blockchain technologies

1) Risk identification	Risk description
	An underlying blockchain solution for managing patient consent has to ensure that after an audit-log record is recorded in the blockchain, it cannot be tampered with, be altered, or removed from the blockchain data structure anymore. At the same time it must be assured that no personal data is stored on the blockchain, otherwise the health status can be inferred. Designs, system characteristics, and assumptions that are indicative of unproven approaches for integrating blockchain technologies are not necessarily vulnerable, however there exists considerably less research and experience, requiring additional diligence and careful protocol analysis to ensure correctness and security. [D6.1]
	Risk source
	• Participant / Coordinator • Outside attacker
	Risk cause
	• Vulnerabilities in the smart contracts • Inadequate security measures (weak access controls) • Lack of testing and quality assurance during development (config files, endorsement policies) • Human error (intentional or unintentional), or majority attack
	Possible damage for data subjects
	• Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts) • Non-material damage: Mental health problems because of data disclosure, social and societal disadvantages (e.g. damage to reputation based on data disclosure); dysfunctional auditing mechanism may restrict data subject's informational self-determination.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3)	**(3)** Comment: Creating/updating/revoking patient consents would not be possible unless the private key of a patient is stolen.	**(9)**

3) Measures	Measures
	• The blockchain mechanism is intentionally designed in a way not to store personal data on the blockchain • Strong authentication, proof and authorisation mechanisms • Rigorous testing and validation of the blockchain solution, smart contracts and policies • Robust governance and increased number of peers and orderers

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(2)	**(3)**	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.22 Incorrect or inaccurate model due to differential privacy

1) Risk identification	**Risk description**
	Use of differential privacy in deep learning may disproportionately affect the accuracy of the model, increasing the chance of false predictions when using the model. Exemplarily the accuracy of a model trained using differential privacy tends to decrease more on these classes and subgroups vs. the original, non-private model (Bagdasaryan and Shmatikov 2019).
	Risk source
	• Coordinator
	Risk cause
	• Incorrect application of differential privacy techniques • Application of differential privacy not appropriate for the machine learning method or incorrect addition of noise.
	Possible damage for data subjects
	• Physical damage: Bodily damage by incorrect medical treatment due to application of incorrect or inaccurate model • Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts based on wrong diagnosis) • Non-material damage: Mental health problems because of wrong diagnosis; social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Damage occurs only if an incorrect or inaccurate model is applied to a patient in the treatment context.	**(4)** Comment: Damage to the body is possible. Comment: Damage to body or high material damage not possible.	**(12)**

3) Measures	Measures
	The following measures shall be implemented and have been documented in the deployment manual in Annex I: ● Measurement (testing) of utility (accuracy) on some validation set. ● Application of empirical risk minimization (ERM) algorithms, using them to "search" the space of privacy levels to find the empirically strongest one that meets the accuracy constraint (e.g. Ligett et al. 2017). ● Results shall not be applied without human verification (by a doctor). ● In case a traditional examination method shall be replaced by a predictive model, be particularly aware of the sensitivity of the model (rate of false negatives). Calculate the absolute number of potential false negatives, put it into context and consider what this number means, how many cases the model will overlook and how this can be overcome. ● Explainable AI / Human in the loop (Evaluation applications are consistently being uploaded to the FeatureCloud AppStore https://featurecloud.ai/). ● The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient's safety provisions therein, what should be considered before use.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(1) Comment: In the medical domain with its regulations, strict procedures and involvement of doctors in every treatment decision, professional ethics	**(4)**	**(4)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

	as well as guidelines that must be put in place are able to reduce the risk of unverified application of a result to a minimum. In addition, even the effects of incorrect medical treatment do not necessarily lead to very severe damage and can usually be overcome by countermeasures.		

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.23 Unintended bias

1) Risk identification	Risk description
	Due to the decentralised nature of federated machine learning, certain participants may contribute lower quality or more biased data than others, resulting in biased or skewed results (e.g. Chang and Shokri 2023). This could potentially compromise the accuracy of the global model. Unintended model bias in machine learning refers to situations where a model systematically over- or under-estimates outcomes for certain groups based on certain characteristics or features.
	Risk source
	● Participant / Coordinator ● Developer ● Outside Attacker (injecting data)
	Risk cause

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- The following categories of biases, while not exhaustive, constitute prominent risks and vulnerabilities to consider when designing, developing, deploying, evaluating, using, or auditing AI applications (Schwartz et al. 2022 p6 ff):
    - Systemic Bias: This arises from procedures and practices within institutions that result in certain social groups being favoured or disadvantaged. It may not necessarily involve conscious prejudice, but rather the majority following existing norms. Examples include institutional racism and sexism, as well as limitations for individuals with disabilities due to infrastructures not designed with universal accessibility in mind.
    - Statistical and Computational Bias: These biases result from errors when the sample used is not representative of the entire population. They arise from systematic, non-random errors and can occur without any intentional discrimination. In AI systems, this type of bias can be found in datasets and algorithmic processes, especially when algorithms are trained on limited or specific types of data.
    - Human Bias: This stems from systematic errors in human thinking based on simplified judgmental principles. These biases are often implicit and can affect how individuals or groups interpret information, make decisions, or fill in missing data. They are pervasive in decision-making processes across the AI lifecycle. Examples include cognitive and perceptual biases, which are fundamental aspects of human thinking. These biases can be both helpful (as mental shortcuts) and problematic (leading to cognitive biases).
- Deliberate input of biased data: Arbitrarily selected or fake data gets considered or eligible data is intentionally excluded to influence the results.
- In addition, the fact that federated learning provides only indirect access to the data complicates the discovery of bias and makes it more likely that low-quality data worsens the quality of the global model.

Possible damage for data subjects

- Physical damage: Bodily damage by incorrect medical treatment due to application of incorrect or inaccurate model
- Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts based on wrong diagnosis)
- Non-material damage: Mental health problems because of wrong diagnosis; social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Damage occurs only if an incorrect or inaccurate model is applied to a patient in the treatment context.	**(4)**	**(12)**

3) Measures	Measures
	The following measures shall be implemented and have been documented in the deployment manual: ● Measurement (testing) of utility (accuracy) on some validation set. ● Application of empirical risk minimization (ERM) algorithms, using them to "search" the space of privacy levels to find the empirically strongest one that meets the accuracy constraint (e.g. Ligett et al. 2017). ● Results shall not be applied without human verification (by a doctor). ● In case a traditional examination method shall be replaced by a predictive model, be particularly aware of the sensitivity of the model (rate of false negatives). Calculate the absolute number of potential false negatives, put it into context and consider what this number means, how many cases the model will overlook and how this can be overcome. ● Explainable AI / Human in the loop (Evaluation applications are consistently being uploaded to the FeatureCloud AppStore https://featurecloud.ai/) ● The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient's safety provisions therein, what should be considered before use. ● Definition of mandatory time cycles for evaluation and address biases in the training data and model predictions. ● General IT security best practises and four eyes principle has to be adhered to within the IT of the participant. As a guidance while eliciting and to achieve a large coverage of potential threats, we employ NIST cybersecurity framework (https://www.nist.gov/cyberframework/framework), OWASP (https://owasp.org/www-project-top-ten/#), LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1). ● Harmonization of the examination setup for data collection in cooperation with the coordinator. ● Assessment of what data is needed to ensure a representative, reliable and relevant training dataset and perform training on that data. ● Require checks of the local data on the local level.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(1) Comment: In the medical domain with its regulations, strict procedures and involvement of doctors in every treatment decision, professional ethics as well as guidelines that must be put in place are able to reduce the risk of unverified application of a result to a minimum. In addition, even the effects of incorrect medical treatment do not necessarily lead to very severe damage and can usually be overcome by countermeasures.	**(4)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk can be achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.24 Incorrect model due to malicious apps

1) Risk identification	**Risk description**
	Applications may be uploaded to the FeatureCloud "App Store". These Apps are used by coordinators to set up projects. Bad actors may upload malicious apps. Such malicious apps may intentionally or by mistake negatively affect the accuracy of the model, increasing the chance of false predictions when using the model. For example the coordinator can craft a model that will always yield the result that the coordinator wants, regardless of the input or only aggregates selected results of the federated learning.
	Risk source
	• Outside attacker (application developer) • Coordinator
	Risk cause
	• Execution of described malicious applications
	Possible damage for data subjects
	• Physical damage: Bodily damage by incorrect medical treatment due to application of incorrect or inaccurate model • Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts based on wrong diagnosis) • Non-material damage: Mental health problems because of wrong diagnosis; social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: While anyone can upload an app to the FeatureCloud "App store", for the risk to manifest a coordinator has to make use of the malicious app image. In addition, damage occurs only if an incorrect or inaccurate model is applied to a patient in the treatment context.	**(4)**	**(12)**

3) Measures	Measures
	The following measures shall be implemented and have been documented in the deployment manual: ● Certification process of apps and clear labelling of certified apps [5.2.3] and preferred choice of such apps during project setup. ● Measurement (testing) of utility (accuracy) on some validation set. ● Application of empirical risk minimization (ERM) algorithms, using them to "search" the space of privacy levels to find the empirically strongest one that meets the accuracy constraint (e.g. Ligett et al. 2017). ● Results shall not be applied without human verification (by a doctor). ● In case a traditional examination method shall be replaced by a predictive model, be particularly aware of the sensitivity of the model (rate of false negatives). Calculate the absolute number of potential false negatives, put it into context and consider what this number means, how many cases the model will overlook and how this can be overcome. ● Explainable AI / Human in the loop (Evaluation applications are consistently being uploaded to the FeatureCloud AppStore https://featurecloud.ai/) ● The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient's safety provisions therein, what should be considered before use. ● Definition of mandatory time cycles for evaluation. ● General IT security best practises and four eyes principle has to be adhered to within the IT of the participant. As a guidance while eliciting and to achieve

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

	a large coverage of potential threats, we employ NIST cybersecurity framework (https://www.nist.gov/cyberframework/framework), OWASP (https://owasp.org/www-project-top-ten/#), LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1). • Harmonization of the examination setup for data collection in cooperation with the coordinator. • Assessment of what data is needed to ensure a representative, reliable and relevant training dataset and perform training on that data. • Require checks of the local data on the local level.

4) Risk analysis and evaluation (after mitigation)	**Probability of occurrence**	**Severity of damage**	**Risk assessment**
	(1) Comment: In the medical domain with its regulations, strict procedures and involvement of doctors in every treatment decision, professional ethics as well as guidelines that must be put in place are able to reduce the risk of unverified application of a result to a minimum. In addition, even the effects of incorrect medical treatment do not necessarily lead to very severe damage and can usually be overcome by countermeasures.	**(4)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.25 Model drift

1) Risk identification	Risk description
	The fundamental assumption in machine learning is that the patterns the model learns from the historical data (training data) will hold true for the new, unseen data (test or future data). However, this assumption may not hold true in a dynamic environment where patterns can shift and evolve over time. This can lead to model drift, impacting the overall performance and learning system. It poses particular challenges for federated learning, because drifts arise staggered in time and space (across clients) (Jothimurugesan et al. 2022).
	Risk source
	• Participant / Coordinator
	Risk cause
	• Training data no longer being relevant or adequate. • Undetected model drift is caused by irregular system testing or by Non-meaningful human review due to a lack of training for human reviewers to interpret and challenge outputs made by an AI system.
	Possible damage for data subjects
	• Physical damage: Bodily damage by incorrect medical treatment due to application of incorrect or inaccurate model • Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts based on wrong diagnosis) • Non-material damage: Mental health problems because of wrong diagnosis; social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Damage occurs only if an incorrect or inaccurate model is applied to a patient in the treatment context.	**(4)**	**(12)**

3) Measures	Measures
	The following measures shall be implemented and have been documented in the deployment manual: ● Document and define a testing regime to occur at regular intervals to detect and correct model drift in appropriate timeframes. ● Document and define measures to ensure human review remains meaningful (e.g. periodically test whether a human reviewer identifies an intentionally inaccurate decision). ● Measurement (testing) of utility (accuracy) on some validation set. ● Application of empirical risk minimization (ERM) algorithms, using them to "search" the space of privacy levels to find the empirically strongest one that meets the accuracy constraint (e.g. Ligett et al. 2017). ● Results shall not be applied without human verification (by a doctor). ● In case a traditional examination method shall be replaced by a predictive model, be particularly aware of the sensitivity of the model (rate of false negatives). Calculate the absolute number of potential false negatives, put it into context and consider what this number means, how many cases the model will overlook and how this can be overcome. ● Explainable AI / Human in the loop (Evaluation applications are consistently being uploaded to the FeatureCloud AppStore https://featurecloud.ai/) ● The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient's safety provisions therein, what should be considered before use. ● Definition of mandatory time cycles for evaluation and address biases in the training data and model predictions. ● General IT security best practises and four eyes principle has to be adhered to within the IT of the participant. ● Harmonization of the examination setup for data collection in cooperation with the coordinator. ● Assessment of what data is needed to ensure a representative, reliable and relevant training dataset and perform training on that data. ● Require checks of the local data on the local level.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4) Risk analysis and evaluation (after mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(1) Comment: In the medical domain with its regulations, strict procedures and involvement of doctors in every treatment decision, professional ethics as well as guidelines that must be put in place are able to reduce the risk of unverified application of a result to a minimum. In addition, even the effects of incorrect medical treatment do not necessarily lead to very severe damage and can usually be overcome by countermeasures.	**(4)**	**(4)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9.3.26 Wrongful application or interpretation of outputs

1) Risk identification	Risk description
	The application of federated learning models can present certain risks if not carried out appropriately. These risks often arise due to a misunderstanding of the model's capabilities, its limitations, or the context in which it should be applied. In particular these risks arise when the model is applied to cases for which it was not developed (e.g. model was trained to detect diabetes type 2 and is applied to detect diabetes type 1).
	Risk source
	• Participant / Coordinator • User of a Model (who applies the model for diagnostic purposes)
	Risk cause
	• In federated learning, different nodes could have different amounts of data, and this data may vary in quality and relevance. If these imbalances are not addressed, they can introduce bias and inaccuracies in the global model. • The model trained in federated learning is a global model that generalises across all local data. If it is directly used for specific local predictions without considering the unique characteristics of local data, it can lead to poor performance or misleading results. • Federated learning models, like other machine learning models, can be complex and difficult to interpret. If they're applied in high-stakes areas without a clear understanding of how they make predictions, this can lead to inappropriate decisions and outcomes. • All these limitations - if not well understood - may be the cause of damage.
	Possible damage for data subjects
	• Physical damage: Bodily damage by incorrect medical treatment due to application of incorrect or inaccurate model. • Material damage: Restrictions on the conclusion of contracts (e.g. adverse effects in insurance contracts based on wrong diagnosis) • Non-material damage: Mental health problems because of wrong diagnosis; social and societal disadvantages (e.g. damage to reputation based on wrong diagnosis)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2) Risk analysis and evaluation (before mitigation)	Probability of occurrence	Severity of damage	Risk assessment
	(3) Comment: Damage occurs only if an incorrect or inaccurate model is applied to a patient in the treatment context.	**(4)** Comment: Damage to body possible	**(12)**

3) Measures	Measures
	The following measures shall be implemented and have been documented in the deployment manual: ● Document and define measures to ensure human review remains meaningful (e.g. periodically test whether a human reviewer identifies an intentionally inaccurate decision). ● Testing of model limitation. ● Training of users, stakeholders, and decision-makers on the limitations of the model and the potential risks associated with misinterpretation. ● Assessment of the model's robustness to variations in input data. ● Results shall not be applied without human verification (by a doctor). ● In case a traditional examination method shall be replaced by a predictive model, be particularly aware of the sensitivity of the model (rate of false negatives). Calculate the absolute number of potential false negatives, put it into context and consider what this number means, how many cases the model will overlook and how this can be overcome. ● Explainable AI / Human in the loop (Evaluation applications are consistently being uploaded to the FeatureCloud AppStore https://featurecloud.ai/) ● The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient's safety provisions therein, what should be considered before use. ● Definition of mandatory time cycles for evaluation of model limitations. ● General IT security best practises and four eyes principle has to be adhered to within the IT of the participant. As a guidance while eliciting and to achieve a large coverage of potential threats, we employ NIST cybersecurity framework (https://www.nist.gov/cyberframework/framework), OWASP (https://owasp.org/www-project-top-ten/#), LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

	• Harmonization of the examination setup for data collection in cooperation with the coordinator. • Assessment of what data is needed to ensure a representative, reliable and relevant training dataset and perform training on that data. • Require checks of the local data on the local level. • Provision of uncertainty estimates along with model predictions to convey the level of confidence in the output.

4) Risk analysis and evaluation (after mitigation)	**Probability of occurrence**	**Severity of damage**	**Risk assessment**
	(2)	**(3)** Comment: Human oversight can mitigate most severe damage.	**(6)**

5) Risk treatment / Proportionality of risk / Future measures	Through implementing the measures documented above an acceptable level of residual risk was achieved. The risk management process must be continued throughout further development and deployment of the FeatureCloud system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

10 Open issues

As emphasized in the methodology section above, the present report is methodologically structured on the basis of a data protection impact assessment in accordance with Article 35 of the GDPR. Since the circumstances of the use of the developed systems are not known in detail a specific DPIA on the basis of this report has to be conducted before actual use of FeatureCloud by the respective controllers.

This is especially relevant for choosing a legal basis and provision for data subject rights where actual legal requirements here may particularly vary on national level and have to be analysed before actual deployment.

11 Conclusion

This report has accomplished several objectives and serves several purposes crucial for the deployment and use of the FeatureCloud platform and app store beyond the official conclusion of the FeatureCloud H2020 research project by the end of 2023 and is meant to serve as the major guidance document for the different stakeholders deploying and using FeatureCloud as follows:

It describes the FeatureCloud system extensively in a way stakeholders with various backgrounds should be able to understand. On this basis, this report analyses the results of applying data protection law and the proposed AI Act to this system. Roles of different stakeholders are defined and legal admissibility is analysed. What follows is the centrepiece of this report, the risk analysis including the identified mitigation measures. This is the result of a process of identifying and analysing risks and identifying appropriate mitigation measures which has been carried out throughout the FeatureCloud research project from the beginning. Finally, in Annex I of this report, those risk mitigation measures and respective recommendations are compiled which by their nature go beyond what was possible to implement already during the development stage following a privacy-by-design approach but can only be put into practice by the respective stakeholders during the different phases of an actual study.

Thus, the present document, compiling a lot of knowledge developed in the FeatureCloud project in the form of a DPIA report compliant with Article 35 GDPR, contains everything, which could be collected and analysed until the end of the FeatureCloud project, i.e. end of 2023, that is necessary for conducting a DPIA for the actual use of FeatureCloud in a particular use case. By extending it in relation to the circumstances of the individual case a well-founded DPIA for the specific use case can be conducted relatively quickly.

The major finding documented in this deliverable is two-fold: It could be demonstrated that federated learning leads to the claimed privacy and security gain and that additional privacy and security risks which are not mitigated by the federated approach have been dealt with or can be dealt with appropriately. According to the risk analysis, when the results of the FeatureCloud project are applied properly, no unmitigated high risks remain: The identified risk mitigation measures, which are either already implemented in FeatureCloud as far as this has been possible by their nature or are otherwise included in the FeatureCloud deployment manual, are able to reduce all identified risks to a level below the critical threshold (“high” according to Article 35 GDPR). Therefore, it can be concluded that the federated FeatureCloud approach is able to achieve the gain in privacy and security which is the reason why it has been chosen.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

12 References

AEPD-EDPS. (2022) Joint Paper - 10 Misunderstandings about Machine Learning.

Ateniese, Felici, Mancini, Spognardi, Villani and Vitali. (2015). “Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers.” Int. J. Secur. Netw. 10 (3): 137–150. https://doi.org/10.1504/IJSN.2015.071829.

Article 29 Data Protection Working Party. (2007) Opinion 4/2007 on the concept of personal data.

Article 29 Data Protection Working Party. (2013) Opinion 03/2013 on purpose limitation

Article 29 Data Protection Working Party. (2010) Opinion 1/2010 on the concepts of "controller" and "processor".

Article 29 Data Protection Working Party. (2017). Guidelines on consent under Regulation 2016/679, WP 259 rev.01., 17/EN (2017a); https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030.

Bagdasaryan and Shmatikov. (2019) Differential Privacy Has Disparate Impact on Model Accuracy. Cornell University.

Bertino and Ferrari. 2018. Big data security and privacy. In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years (pp. 425-439). Springer, Cham.

Bonawitz, Keith, Ivanov, Kreuter, Marcedone, McMahan, Patel, Ramage, Segal, and Seth. (2017) “Practical Secure Aggregation for Privacy-Preserving Machine Learning.” In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191. (CCS ’17). New York, NY, USA: ACM. https://doi.org/10.1145/3133956.3133982.

Bourtoule, Chandrasekaran, Choquette-Choo, Jia, Travers, Zhang, Lie and Papernot. (2020) Machine Unlearning. n 42nd IEEE Symposium of Security and Privacy, University of Toronto, Vector Institute, University of Wisconsin-Madison.

Benhamouda, Herranz, Joye, Libert, (2017) Efficient Cryptosystems From 2^k-th Power Residue Symbols. J. Cryptol. 30, 519–549. https://doi.org/10.1007/s00145-016-9229-5

Bitkom. (2017) Risk Assessment & Datenschutz-Folgenabschätzung – Leitfaden (Bitkom e. V. 2017) https://www.bitkom.org/sites/default/files/file/import/FirstSpirit-1496129138918170529-LF-Risk-Assessment-online.pdf.

Bobek. (2018) Fashion ID v Verbraucherzentrale ECLI:EU:C:2018:1039.

Buchner and Petri. (2018) In Kühling and Buchner (Eds.), DS-GVO/BDSG, Kommentar (C.H. Beck 2018).

Buchner (2018) In Kühling and Buchner (Eds.), DS-GVO/BDSG, Kommentar (C.H. Beck 2018).

Buchner. (2019) Grundsätze und Rechtmäßigkeit der Datenverarbeitung unter der DS-GVO. Datenschutz und Datensicherheit – DuD 155–161;

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Bogucki, Engler, Perarnaud and Renda. (2022) The AI Act and emerging EU digital acquis (CEPS In-Depth Analysis), 2022 – 2, https://www.ceps.eu/ceps-publications/the-ai-act-and-emerging-eu-digital-acquis/.

C-25/17 Tietosuojavaltuutettu v Jehovan todistajat ECLI:EU:C:2018:551.

C-40/17 Fashion ID v Verbraucherzentrale ECLI:EU:C:2019:629.

C-582/14 Patrick Breyer v Bundesrepublik Deutschland ECLI:EU:C:2016:779.

Carlini, Liu, Erlingsson, Kos, and D. Song. (2019) "The secret sharer: Evaluating and testing unintended memorization in neural networks," in Proceedings of the 28th USENIX Conference on Security Symposium. USENIX Association.

Chang and Shokri. (2023) Bias propagation in federated learning, School of Computing, National University of Singapore, ICLR 2023.

Commission Nationale de l'Informatique et des Libertés (CNIL). (2018) 'Privacy Impact Assessment (PIA): Knowledge Base' (CNIL 2018); https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-3-en-knowledgebases.pdf.

Datenschutzkonferenz (DSK). (2018) Kurzpapier Nr. 16 Gemeinsam für die Verarbeitung Verantwortliche, Art. 26 DS- GVO.

Ebers, Hoch, Rosenkranz, Ruschmeier and Steinrötter. (2021) The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS), [J-MDPI] J 2021, 4(4), 589-603; https://doi.org/10.3390/j4040043.

EDPB. (2020) Guidelines 07/2020 on the concepts of controller and processor in the GDPR.

ENISA. (2021) Guideline on security measures under the EECC 4th Edition, https://www.enisa.europa.eu/publications/guideline-on-security-measures-under-the-eecc/@@download/fullReport.

Ennöckl. (2014) *Der Schutz der Privatsphäre in der elektronischen Datenverarbeitung. In: Forschungen aus Staat und Recht – Band 174, Verlag Österreich* (2014).

European Data Protection Supervisor (EDPS). (2020) A Preliminary Opinion on data protection and scientific research.

European Data Protection Supervisor (EDPS). (2017) Assessing the necessity of measures that limit the fundamental right to the protection of personal data: A Toolkit (EDPS 2017); https://edps.europa.eu/sites/edp/files/publication/17-04-11_necessity_toolkit_en_0.pdf.

European Data Protection Supervisor (EDPS). (2019) Accountability on the ground Part II: Data protection Impact Assessments & Prior Consultation (EDPS 2019a); https://edps.europa.eu/sites/edp/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf.

European Data Protection Supervisor (EDPS). (2019) Guidelines on assessing the proportionality of measures that limit the fundamental rights to privacy and to the protection of personal data (EDPS 2019b); https://edps.europa.eu/sites/edp/files/publication/19-12-19_edps_proportionality_guidelines2_en.pdf.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Edwards and Veale. (2017) Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For, Duke Law & Technology Review 18-84.

Esayas. (2015) "The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the 'all or nothing' approach." European Journal of Law and Technology, Vol 6, No 2.

Evans, Kolesnikov, Rosulek. (2018) A Pragmatic Introduction to Secure Multi-Party Computation. Found. Trends® Priv. Secur. 2, 70–246. https://doi.org/10.1561/3300000019

Fredrikson, Jha, and Ristenpart. (2015) “Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures.” In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS). Denver, Colorado, USA: ACM. https://doi.org/10.1145/2810103.2813677.

Gabauer, *Die Verarbeitung personenbezogener Daten zu wissenschaftlichen Forschungszwecken* in *Neue Juristische Monografien*, NWV im Verlag Österreich GmbH (2019).

Ganju, Wang, Yang, Gunter and Borisov. (2018) “Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant Representations.” In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 619–633. CCS ’18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3243734.3243834.

Gentry. (2009) Fully homomorphic encryption using ideal lattices, in: Proceedings of the 41st Annual ACM Symposium on Symposium on Theory of Computing - STOC ’09. Presented at the 41st annual ACM symposium, ACM Press, Bethesda, MD, USA, p. 169. https://doi.org/10.1145/1536414.1536440

ICO. (2022) AI and Data Protection Risk Toolkit https://ico.org.uk/media/for-organisations/documents/4020151/ai-and-dp-risk-toolkit-v1_0.xlsx

IG NB (2022) Questionnaire „Artificial Intelligence (AI) in medical devices“, Version 4; https://www.ig-nb.de/index.php?eID=dumpFile&t=f&f=2618&token=010db38d577b0bfa3c909d6f1d74b19485e86975.

Jarass. (2021) *Charta der Grundrechte der Europäischen Union: GRCh* (C.H. BECK 2021).

Jandt. (2018) In Kühling and Buchner (Eds.), DS-GVO/BDSG (C.H. BECK 2018).

Jothimurugesan, Hsieh, Wang, Joshi and Gibbons. (2022) Federated Learning under Distributed Concept Drift. In AISTATS 2023

Kühling and Buchner (Eds.). (2018) DS-GVO/BDSG.

Kastelitz, Hötzendorfer and Tschohl. (2018) In Knyrim (Ed.) DatKomm.

Haidinger. (2018) In Knyrim (Ed.) DatKomm.

Kloza, Calvi, Casiraghi, Vazquez Maymir, Ioannidis, Tanas, Van Dijk. (2020) ‘Data protection impact assessment in the European Union: developing a template for a report from the assessment process’ (Brussels Laboratory for Data Protection & Privacy Impact Assessments, Policy Brief 1/2020, VUB 2020); https://doi.org/10.31228/osf.io/7qrfp.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Karg. (2015) '*Anonymität, Pseudonymität und Personenbezug revisited' Datenschutz und Datensicherheit* – DuD 520–526; https://doi.org/10.1007/s11623-015-0463-z.

Kotschy, *Die Zulässigkeitsvoraussetzungen für Forschungsdatenverarbeitungen nach dem FOG – eine kritische Analyse, in Jahnel (Ed.), Datenschutzrecht. Jahrbuch* 2020 (2021) 287

Lee and Clifton. (2011) How much is enough? choosing ε for differential privacy. In International Conference on Information Security, pages 325–340, Springer.

Jin, Chen, Hsu, Yu, Chen. (2021) CAFE: Catastrophic Data Leakage in Vertical Federated Learning. 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Lekadir, Quaglio, Garmendia and Gallin, Artificial intelligence in healthcare - Applications, risks, and ethical and societal impacts, Panel for the Future of Science and Technology (STOA), European Parliament, 2022, https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2022)729512.

Ligett, Neel, Roth, Waggoner, and Wu. (2017) Accuracy first: Selecting a differential privacy level for accuracy constrained erm. In Advances in Neural Information Processing Systems, pages 2563–2573.

Lindblad, Kernell, Lindblad and Bloch Veiberg. (2020) 'Guidance on human rights impact assessment of digital activities' (The Danish Institute for Human Rights 2020); https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities.

Millard, Kuner, Cate, Lynskey, Loideain and Svantesson. (2019) "At This Rate, Everyone Will Be a [Joint] Controller of Personal Data!". In International Data Privacy Law (9) 2019, 217-219.

Martin, Friedewald, Schiering, Mester, Hallinan and Jensen. (2020) *Die Datenschutz-Folgenabschätzung nach Art. 35 DSGVO - Ein Handbuch für die Praxis (Fraunhofer-Institute für System und Innovationsforschung (ISI)*, Fraunhofer Verlag 2020); https://publica.fraunhofer.de/handle/publica/300193.

Mantelero. (2022) Beyond Data - Human Rights, Ethical and Social Impact Assessment in AI (T.M.C. Asser Press The Hague 2022); https://doi.org/10.1007/978-94-6265-531-7.

MDCG (2019) Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR, MDCG 2019-11; https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2019_11_guidance_qualification_classification_software_en_0.pdf.

MEDDEV (2016) Guidelines on the Qualification and Classification of Stand-Alone Software used in Healthcare within the regulatory framework of medical devices, 2.1/6; https://ec.europa.eu/docsroom/documents/17921.

Möller. (2012) 'Proportionality: Challenging the Critics' (2012) 10(3) International Journal of Constitutional Law 709–731; https://doi.org/10.1093/icon/mos024.

Mugunthan, Polychroniadou, Byrd and Balch (2019) SMPAI: Secure Multi-Party Computation for Federated Learning, In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Nowok, Raab, Dibben. (2016) synthpop: Bespoke Creation of Synthetic Data in R. J. Stat. Softw. Artic. 74.

OECD. (2019) Recommendation of the Council on Artificial Intelligence, https://oecd.ai/en/ai-principles (last accessed 07.11.23).

Office of the United Nations High Commissioner for Human Rights. (2006) 'Frequently asked questions on a human rights-based approach to development cooperation' HR/PUB/06/8.

Orekondy, Tribhuvanesh, Schiele and Fritz. (2019) "Knockoff Nets: Stealing Functionality of Black-Box Models." In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Patki, Wedge and Veeramachanen. (2016) The Synthetic Data Vault, In 2016 IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA). https://doi.org/10.1109/DSAA.2016.49

Schwartz, Vassilev, Greene, Perine, Burt and Hall. (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST Special Publication 127. https://doi.org/10.6028/NIST.SP.1270.

Shokri, Reza, Marco Stronati, Congzheng Song and Shmatikov. (2017) "Membership Inference Attacks Against Machine Learning Models." In 2017 IEEE Symposium on Security and Privacy (SP), 3–18. https://doi.org/10.1109/SP.2017.41.

Simitis, Hornung and Spiecker gen. Döhmann (Eds.). (2019) Datenschutzrecht.

Song, Ristenpart and Shmatikov. (2017) Machine Learning Models that Remember Too Much. In CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications SecurityOctober, Pages 587–601, https://doi.org/10.1145/3133956.3134077.

Tang, Mahloujifar, Song, Shejwalkar, Nasr, Houmansadr and Mittal. (2021) Mitigating Membership Inference Attacks by Self-Distillation Through a Novel Ensemble Architecture, https://doi.org/10.48550/arXiv.2110.08324.

Rothmann, Kastelitz and Rothmund-Burgwall. (2021) *'Archive als ‚öffentliches Gedächtnis' personenbezogener Patientendaten?' in Jahnel (Ed.), Datenschutzrecht. Jahrbuch 2021* (NWV 2022).

Yang, Liu, Chen, and Tong. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

Veale and Zuiderveen Borgesius. (2021) Demystifying the Draft EU Artificial Intelligence Act, Computer Law Review International 22(4) 97-112.

Vemou and Karyda. (2020) 'Evaluating privacy impact assessment methods: guidelines and best practices' (2020) 28(1) Information & Computer Security, 35–53.

Z. Wang, M. Song, Zhang, Y. Song, Q. Wang and Hairong Qi. (2019) "Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning." IEEE INFOCOM 2019 - IEEE Conference on Computer Communications

Weinsberg, Bhagat, Ioannidis and Taft. (2012) "BlurMe: Inferring and Obfuscating User Gender Based on Ratings." In Proceedings of the Sixth ACM Conference on Recommender Systems, 195–202. RecSys '12. New York, NY, USA: ACM. https://doi.org/10.1145/2365952.2365989.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Wood, Najarani and Kahrobaei. (2020) Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics, University of York, https://eprints.whiterose.ac.uk/151333/.

Zander, Armitage and Branch. (2007) "A Survey of Covert Channels and Countermeasures in Computer Network Protocols." IEEE Communications Surveys & Tutorials 9 (3): 44–57. https://doi.org/10.1109/COMST.2007.4317620.

Zarsky. (2017) Incompatible: The GDPR in the Age of Big Data, Seton Hall Law Review, University of Haifa - Faculty of LawVol. 47, No. 4(2), 2017.

13 Other supporting documents

Selected EU regulation

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016).

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (OJ L 117, 5.5.2017).

Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (OJ L 117, 5.5.2017).

Selected public FeatureCloud Deliverable Reports

Deliverable Report D2.1 – Risk assessment methodology

Deliverable Report D2.2 – KPIs and metrics for local execution platforms

Deliverable Report D2.3 – Working PAML-Layer with low distortion

Deliverable Report D3.6 - Manuscript on risk management

Deliverable Report D6.1 – Local blockchain mechanism

Deliverable Report D7.2 – App store ready and extendible by developers

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Annex I: FeatureCloud Deployment Manual

The FeatureCloud Deployment Manual is a collection of measures and recommendations developed throughout the FeatureCloud research project for different stakeholders who will use FeatureCloud in practice after the official end of the research project. These measures and recommendations by their nature go beyond what was possible to implement already during the development stage following a privacy-by-design approach. They can only be put into practice by the respective stakeholders in the respective phase of a study. The risk assessment above is based on the premise that the different stakeholders involved follow these recommendations and implement these measures.

The Deployment Manual is structured along the different roles of stakeholders using FeatureCloud in practice and along the different phases of a study. Measures relevant to more than one stakeholder role are intentionally listed in full for each respective role in order to provide all relevant information for a particular role together in one chapter. The role of the app developer is not related to deployment in the same way as the roles of participant, coordinator and model user described in the following. Guidance for app developers is therefore kept separate and can be found in chapter 5.2.7 and in particular at https://featurecloud.ai/assets/developer_documentation/, including detailed instructions for using privacy-preserving techniques.

The measures and recommendations are in this manual deliberately kept as short as possible. A comprehensive body of information for the deployment of FeatureCloud from the functioning of the system and its different elements to legal admissibility and in particular potential risks and mitigation measures, among other aspects, is the present Report on Data Protection Impact Assessment as a whole. Further details can be found in the FeatureCloud deliverables referenced in square brackets in the following format [<deliverable number>:<section number>].

We highly recommend comprehensively reading and following the FeatureCloud deliverables on quality management [D3.2], software life cycle [D3.4], risk management [D3.6], and usability [D3.8]. These guidelines introduce methodologies on which participants can adhere to different parts of development and deployment in a standardised manner.

Technical deployment will be explained in the final section below, after the following crucial measures and requirements for preventing potential risks that could emerge from deployment and use of FeatureCloud.

1. Participant

1.1 Before participation tokens are sent out

1.1.1 Information duties

- Participants must provide precise information on means and purposes of processing to the data subjects and the other mandatory points of Art 13 GDPR. In case of consent, data subjects must be informed about alternative means of communication to withdraw consent.
- Joint controller agreements shall clearly state which party is responsible for providing information.
- Whenever an application is planned to get placed in a medical context, the participants should be fully aware of the intended use of the application. The intended use definition should include the medical context, primary user audience, patient group, and use environment as specific as possible [D3.6:3.3.2][D3.8:4.3].
- Participants must be aware that neither participants nor the coordinator need to obtain access to raw data (of other participants) at any stage and that asking for raw data or local models

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

outside the predefined communication channels of the FeatureCloud platform is to be considered fraudulent.

1.1.2 Contractual duties

- An adequate joint controllership agreement (Art 26 GDPR) must have been concluded between the participant and the coordinator.

1.1.3 Lawfulness and purpose limitation

In order to adhere to the data protection principles of lawfulness and purpose limitation, the following measures need to be taken (ICO 2022):

- Consultation with experts to ensure that the data to be included in a project is appropriate and adequate.
- Assessment and specification of legal basis for data processing.
- Documentation of purpose(s) for using personal data at each stage of the processing lifecycle. Assessment whether they are compatible with the originally defined purpose, and schedule reviews for reassessment.
- Documentation of the data collected to train the system. Assessment whether it is accurate, adequate, relevant, and limited to the specified purpose(s).
- Reassessment and documentation of what data is necessary, adequate, and relevant for training and testing the system. Consideration of the trade-off between data minimisation and statistical accuracy.

1.1.4 General IT security measures

- General IT security best practises and four eyes principle has to be adhered to within the IT of the hospital. As a guidance while eliciting and to achieve a large coverage of potential threats, we recommend NIST cybersecurity framework (https://www.nist.gov/cyberframework/framework), OWASP (https://owasp.org/www-project-top-ten/#), LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1 and D2.5).
- Minimise the need/involvement of (sub-)processors: The FeatureCloud system is designed in order to run locally at any participant involved and that the local components can be provided to each participant by FeatureCloud. Therefore, no (sub-)processors need to be involved.

1.2 Before training starts

1.2.1 Attack prevention

In order to prevent attacks on the machine learning model [D2.1:8.3], in particular poisoning attacks and evasion attacks, the following recommendations shall be followed:

- Filtering methods for the input (Y. Liu, Xie, and Srivastava 2017) [D2.1:9.2.1]
- Pruning the network (K. Liu, Dolan-Gavitt, and Garg 2018) [D2.1:9.2.1]
- Modifying training samples, model structure or combining the model with other models (Papernot et al. 2016) [D2.1:9.2.1]

To prevent membership inference attacks, the following recommendations shall be followed:

- Use of privacy enhancing techniques (Differential privacy or other noise addition methods) [D2.1:9.2.2]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Optimal choice of privacy parameter Epsilon (ε) [D.2.1:9.2.2] Guidance in this regard is provided by FeatureCloud (https://featurecloud.ai/assets/developer_documentation/privacy_preserving_techniques.html#parameter-guide-anchor)
- SMPC (mitigate against coordinator as attacker) [D2.1:9.2.2]
- Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7]
- Black-box only access to model [D2.1:8.4]
- Model output contains less information
- Adversarial Regularization, Early Stopping (Tang et. al (2021))

General Measures to prevent attacks:

- Automatic monitoring of the amount of data communicated over the network connection (O-Notation; sub-linear exchange quota) [D2.2:3.3.3]
- Strong authentication and authorisation mechanisms, encryption of data storage and communication, and minimisation of data exchange. [D2.1:9]
- Data anonymisation [D2.5:5.4.1]
- LSB sanitization [D2.5:5.4.4]
- Sign modification [D2.5:5.4.4]
- Activation Based Neuron Pruning [D2.5:5.4.4]
- Data fingerprinting [D2.5:5.4.2]

1.2.2 Use of logging mechanism

- The blockchain-based auditing mechanism [D6.1 - D6.5] or an equivalent mechanism must be used, that logs which data has been used for which purpose by whom.
- Participants must be contractually obliged to use the logging mechanism, that logs which data has been used for which purpose by whom and actual audits must be carried out.

1.2.3 Prevention of AI-related risks

- Harmonization of the examination setup for data collection in cooperation with the coordinator
- Assessment of what data is needed to ensure a representative, reliable and relevant training dataset and perform training on that data.
- Checks of the local data on the local level
- Definition of mandatory time cycles for evaluation

1.3 Continuously

1.3.1 Use of logging mechanism and performance of audits

- The blockchain-based auditing mechanism [D6.1 - D6.5] or an equivalent mechanism must be used, that logs which data has been used for which purpose by whom.
- Participants must be contractually obliged to use the logging mechanism, that logs which data has been used for which purpose by whom and actual audits must be carried out.

1.3.2 Revocation unobservability

- The attending physician/doctor treating the data subject or the doctor who recruited the data subject for a study should not be able to know that the data subject withdrew consent regarding the use of her data for training. This in turn also mitigates the harm of pressure to consent in the first place as consent can truly be withdrawn freely. The blockchain-based consent management solution is designed in order to provide revocation unobservability [D6.5:6.2.1],

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

but it also depends on the individual implementation in the hospital, in particular if the consent management is paper-based there.

2. Coordinator

2.1 Before apps are selected

2.1.1 Attack prevention

In order to prevent attacks on the machine learning model [D2.1:8.3], in particular poisoning attacks and evasion attacks, the following recommendations shall be followed:

- Use of simpler machine learning methods less susceptible to such attacks and/or make such attacks being recognized more easily.
- Filtering methods for the input (Y. Liu, Xie, and Srivastava 2017) [D2.1:9.2.1]
- Pruning the network (K. Liu, Dolan-Gavitt, and Garg 2018) [D2.1:9.2.1]
- Modifying training samples, model structure or combining the model with other models (Papernot et al. 2016) [D2.1:9.2.1]

To prevent membership inference attacks, the following recommendations shall be followed:

- Use of privacy enhancing techniques (Differential privacy or other noise addition methods) [D2.1:9.2.2]
- Optimal choice of privacy parameter Epsilon (ε) [D.2.1:9.2.2] Guidance in this regard is provided by FeatureCloud (https://featurecloud.ai/assets/developer_documentation/privacy_preserving_techniques.html#parameter-guide-anchor)
- SMPC (mitigate against coordinator as attacker) [D2.1:9.2.2]
- Synthetic Data Generation (e.g. Nowok et al. 2016; Patki et al. 2016) [D2.3:2, 7]
- Black-box only access to model [D2.1:8.4]
- Model output contains less information
- Adversarial Regularization, Early Stopping (Tang et. al (2021))
- Use of a simpler model
- Data anonymisation [D2.5:5.4.1]
- LSB sanitization [D2.5:5.4.4]
- Sign modification [D2.5:5.4.4]
- Activation Based Neuron Pruning [D2.5:5.4.4]

General Measures to prevent attacks:

- Automatic monitoring of the amount of data communicated over the network connection (O-Notation; sub-linear exchange quota) [D2.2:3.3.3] (and other KPIs?)
- As a guidance while eliciting and to achieve a large coverage of potential threats, we employ LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1).
- Minimise the need/involvement of (sub-)processors: The FeatureCloud system is designed in order to run locally. Therefore, no (sub-)processors need to be involved.

2.1.2 Use of certified apps only

- Apps which are not certified must not be used in practice.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2.2 Before participation tokens are sent out

2.2.1 Information duties

- Participants must be informed about their duty to provide precise information on means and purposes of processing to the data subjects.
- Joint controller agreements shall clearly state which party is responsible for providing information. Most importantly, the coordinator as the party being in control of the means of processing must enable the participants to inform the data subjects about the processing.
- New participants must be actively informed that neither participants nor the coordinator need to obtain access to raw data (of other participants) at any stage and that asking for raw data or local models outside the predefined communication channels of the FeatureCloud platform is to be considered fraudulent (anti-phishing training).

2.2.2Contractual duties

- An adequate joint controllership agreement (Art 26 GDPR) must have been concluded between the participant and the coordinator.

2.2.3 Use of logging mechanism and performance of audits

- The blockchain-based auditing mechanism [D6.1 - D6.5] or an equivalent mechanism must be used, that logs which data has been used for which purpose by whom.
- Participants must be contractually obliged to use the logging mechanism, that logs which data has been used for which purpose by whom and actual audits must be carried out.

2.3 Before training starts

2.3.1 Lawfulness and purpose limitation

In order to adhere to the data protection principles of lawfulness and purpose limitation, the coordinator shall check whether participants have implemented the following measures (ICO 2022):

- Consultation with experts to ensure that the data to be included in a project is appropriate and adequate.
- Documentation of purpose(s) for using personal data at each stage of the processing lifecycle. Assessment whether they are compatible with the originally defined purpose, and schedule reviews for reassessment.
- Documentation of the data collected to train the system. Assessment whether it is accurate, adequate, relevant, and limited to the specified purpose(s).
- Reassessment and documentation of what data is necessary, adequate, and relevant for training and testing the system. Consideration of the trade-off between data minimisation and statistical accuracy. Consult with domain experts to ensure that the data you intend on collecting is appropriate and adequate.

2.3.2 Data bias prevention

- General IT security best practices and four eyes principle have to be adhered to within the IT of the hospital.
- Harmonization of the examination setup for data collection in cooperation with the coordinator.
- Document and define a testing regime to occur at regular intervals to detect and correct model drift in appropriate timeframes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Document and define measures to ensure human review remains meaningful (e.g. periodically test whether a human reviewer identifies an intentionally inaccurate decision).
- Always test: Measure utility (accuracy) on some validation set.

2.4 Before inference

2.4.1 Prevention of AI-related risks

- Assessment of what data is needed to ensure a representative, reliable and relevant training dataset and perform training on that data.
- Require checks of the local data on the local level.
- Certification process of apps and clear labelling of certified apps [5.2.3] and preferred choice of such apps during project setup.
- Measurement (testing) of utility (accuracy) of the model on some validation set.
- Application of empirical risk minimization (ERM) algorithms, using them to “search” the space of privacy levels to find the empirically strongest one that meets the accuracy constraint (e.g. Ligett et al. 2017).
- Use of Explainable AI / Human in the loop applications which are consistently being uploaded to the FeatureCloud AppStore (https://featurecloud.ai/)
- Definition of mandatory time cycles for evaluation and address biases in the training data and model predictions.
- Testing of model limitation
- Assessment of the model's robustness to variations in input data.
- Provision of uncertainty estimates along with model predictions to convey the level of confidence in the output.
- The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient’s safety provisions therein, what should be considered before use.

2.5 Continuously

2.5.1 Prevention of AI-related risks

- Document and define a testing regime to occur at regular intervals to detect and correct model drift in appropriate timeframes.
- Document and define measures to ensure human review remains meaningful (e.g. periodically test whether a human reviewer identifies an intentionally inaccurate decision).
- Training of users, stakeholders, and decision-makers on the limitations of the model and the potential risks associated with misinterpretation.

3. Model User

3.1 Before inference

- Check whether models and associated applications underlie the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR). In this case, it is required to fulfil regulatory requirements that can be

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

achieved by applying, for example, ISO 13485 for quality management, IEC 62304 for software life cycle, ISO 14971 for risk management, and ISO 62366 for usability engineering. Otherwise, it's recommended to follow our guideline manuscripts [D3.2, D3.4, D3.6, D3.8],

3.2 Continuously

3.2.1 Prevention of attacks

- General IT security best practises and four eyes principle has to be adhered to within the IT of the hospital. As a guidance while eliciting and to achieve a large coverage of potential threats, we recommend NIST cybersecurity framework (https://www.nist.gov/cyberframework/framework), OWASP (https://owasp.org/www-project-top-ten/#), LINDDUN (https://linddun.org/), STRIDE (https://owasp.org/www-community/Threat_Modeling_Process) and ENISA (ENISA 2021) guidance documents (see in detail D2.1 and D2.5).

3.2.2 Prevention of AI-related risks

- Results shall not be applied without human verification (by a doctor), see also Article 22 of the GDPR.
- In case a traditional examination method shall be replaced by a predictive model, be particularly aware of the sensitivity of the model (rate of false negatives). Calculate the absolute number of potential false negatives, put it into context and consider what this number means, how many cases the model will overlook and how this can be overcome.
- The application of models trained through FeatureCloud to actual patients for treatment purposes most likely underlies the Regulation (EU) 2017/745 on Medical Devices (MDR) or the Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR) and is therefore restricted by the patient's safety provisions therein, what should be considered before use.
- Document and define a testing regime to occur at regular intervals to detect and correct model drift in appropriate timeframes.
- Document and define measures to ensure human review remains meaningful (e.g. periodically test whether a human reviewer identifies an intentionally inaccurate decision).
- Training of users, stakeholders, and decision-makers on the limitations of the model and the potential risks associated with misinterpretation.

4. Deployment at the participant

The following guide for technical deployment of FeatureCloud in the clinic is a result of the requirement engineering process carried out in WP8 and is part of D8.6, sections 6.1.5.2 and 6.1.6. There, also a sequence diagram of a proposed workflow can be found.

4.1 System requirements

In the data holder's infrastructure, the following requirements need to be fulfilled for FeatureCloud components to run:

- Docker environment,
- Outbound access to internet for controller: web ports 80 and 443 towards the global backend,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Inbound and outbound access for the relay server on port 9140 for ensuring communication towards other controllers through the Global Relay Server, or inbound and outbound access to a port of your choosing when implementing a custom relay server.
- Outbound access for the frontend component (featurecloud.ai) on standard web ports 80 and 443,
- Local network access for the frontend component to the controller.

4.2 Preconditions

4.2.1 Hardware

Requirements for running the controller:

- 4 GB RAM minimum, 8 GB optimal,
- 64-bit kernel and CPU support for virtualization.

4.2.2 Software

The Docker environment should be set up on the system that will run the controller.

4.2.3 Data types

As a general recommendation, access to the production database should be restricted, and analysis should be performed on either a production data replica or an exported database containing relevant data. In special cases it can also be performed on a csv export when the data size is under 200 MB. After exporting the relevant data from the database, the data should be anonymized if possible.

4.2.4 Data pre-processing

The data must be curated and sanitized to avoid data inconsistency. These inconsistencies typically appear in historical data, where data types are not enforced.

This step should be done by a researcher with knowledge about all of the apps in the planned work-flow.

The data pre-processor app will read the anonymized dataset and prepare the data for the analysis application.

4.3 Deployment scenario

The deployment scenario is a guideline on how FeatureCloud could be deployed and used.

4.3.1 Communication protocol relevant to federated execution

The Global Relay Server component will ensure communication between controllers deployed in the two locations.

The controllers will communicate with the frontend and the global backend on standard HTTP ports.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4.3.2 Installation

The controller start script will be downloaded from featurecloud.ai or can be accessed via the FeatureCloud Python package. The controller should be started in the environment where the prepared anonymized dataset is located and where the Docker environment is available. The controller runs in a Docker container and will start other containers as part of the planned workflow. The frontend component will be run from a workstation that has access to the internet and to the controller via the local network.

4.3.3 Maintenance analysis

The only component deployed physically in Pro-Vitam and Pro-Medical environments is the controller. Presuming that the environment remains the same, the system's maintenance resumes to updating the controller to the latest version. This is handled automatically when starting the controller. Running the start script closes the already-running controller, pulls the newest image from the FeatureCloud Docker registry and starts the latest controller version.

If changes in the environment might affect controller connectivity to the database and internet, an analysis of the new system is recommended.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.