FeatureCloud

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
D6.2 "Model for defining user rights in federated machine learning"

Work Package
WP6 "Blockchains and user right management"

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) - Established by the European Commission, Unit HaDEA.A.3 – Health Research			
Project Coordinator	Jan BAUMBACH, UNIVERSITY OF HAMBURG (UHAM)			
Deliverable	D6.2 “Model for defining user rights in federated machine learning”			
Work Package	WP6 “Blockchains and user right management”			
Date of Delivery	**Contractual**	M32 (31.08.2021)	**Actual**	M33 (06.09.2021)
Nature	Report	**Dissemination Level**	Public	
Lead Beneficiary	05 SBA Research			
Responsible Author(s)	Fenghong Zhang (SBA), Aljosha Judmayer (SBA),Walid Fdhila (SBA), Nicholas Stifter (SBA)			
Keywords	Federated machine learning, blockchain, DLT, consent management, user rights			

Table of Content

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 objective of this deliverable based on the description of action is Objective 2, which aims at defining a model for consent management in federated machine learning: “To introduce user-rights management into blockchain mechanisms, where patients can define, which of their sensitive information might be used in federated learning (Task 2)”

The corresponding task in this work package is Task 2: “User-rights management in Blockchains (SBA, TUM, MUG). This task integrates user rights mechanisms into blockchain mechanisms, notably the one designed in Task 1. This will be done by analysing the actual information and meta-information required (MUG, TUM) and designing mechanisms for introducing user-rights into smart contracts based on our blockchain mechanism (SBA). This also includes research on the required and available information granularity (MUG) and, a study on usability with typical participants, including expert partners in hospitals, as well as patients (TUM). For the evaluation, SBA will use artificial data in order to be able to draw comparisons without violating patient rights.”

2 Executive Summary

Deliverable (D6.2) focuses on identifying and discussing the **threat model** related to federated machine learning of healthcare data. This goes by identifying possible malicious (or compromised) actors and the corresponding threats they could pose to the system and how to eventually mitigate them. The threat model represents an important step upon which a model for managing user rights / consents will be defined, which ensures that only eligible and consented data are used, and makes it difficult for the system actors to behave maliciously. D6.2 builds upon the requirements specification, architecture, roles and workflow model defined in D6.1 and further investigates security and privacy threats within the system along with possible mitigation mechanisms. The threat modelling follows a mixture of the threat modelling defined in the Microsoft SDL and attack trees, where threats are modelled as attack trees with attack goals as roots and alternative ways to achieve that goal as tree branches. Finally, mitigation mechanisms are proposed and validated. D6.2 identified two main attack goals consisting of (i) leaking healthcare data, and (ii) manipulating the outcome of an ML study. Furthermore, four attack surfaces were determined and grouped by the corresponding responsible actors. In this deliverable, the threats related to data manipulation such as using non-consented data, or excluding consented and eligible data from a study are particularly important in defining the consent model that minimizes the attack surface, e.g., using cryptographic material and digitized consents. The second part of the deliverable D6.2 first discusses two different proposals to manage consents; i) basic solution, and (ii) advanced solution, then presents a consent model and a consent management method that uses blockchain technology and smart contracts. The advanced solution requires clear definition of both participant and patient identities that prevents data correlation and patient identity deanonymization.

3 Introduction (Challenge)

In the healthcare domain, data is often distributed over multiple locations (e.g., hospitals or pharmaceutical companies) and follow different and diverse representations. Therefore, conducting machine learning studies usually would require such data to be homogenized and gathered in the same location, e.g., central cloud. This, however, raises concerns over data privacy and security as it creates a single point of attack and requires data to be relocated, and possibly even moved outside its jurisdiction, thereby involving regulatory and compliance challenges. In order to avoid such problems, the FeatureCloud project employs federated machine learning where locally learned

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

models are aggregated instead of the actual data, thus ensuring that all data remains both legally and technically within the data provider infrastructures. This, in turn, relies on the assumption that all actors involved in the federated machine learning are trusted and behave honestly, meaning that the collaborating parties are required to provide correct results and use eligible data. However, previous and recent data leaks have proven that such strong trust assumptions are unlikely to hold in practice, and consequently, a clear threat model that accounts for privacy and security threats needs to be elaborated. Furthermore, technical failures or user error can manifest itself in system behavior that mimics the actions of a malicious entity. It is hence not only prudent but necessary to evaluate the possible risks and threats the FeatureCloud project could face to identify the potential impact of such failures on the correctness and security of the overall system. In this regard, possible threats should be analyzed and identified, and mitigation mechanisms have to be put in place, either by enforcing specific policies such as using trusted computing infrastructures, or by employing active monitoring of study executions, i.e., runtime monitoring or post audits. Such threats include, but are not limited to, hospitals claiming data they do not have (e.g., to be considered for a study), or data providers manipulating output results by using non-consented or incomplete data.

Because data is only kept locally and it cannot be verified by external parties that hospitals have solely used data for which patients have given their consent, the FeatureCloud project particularly relies on a post auditing process, where an auditor is needed to check the integrity of the results and that no data manipulation was carried out during the learning process. Some of the challenges that need to be addressed are summarized below:

i) How to tell if the hospital has executed the ML model faithfully or has manipulated the execution and its results
ii) How to tell if the hospital has generated fake data (or patients) to bias a study.
iii) How to tell if the hospital has omitted data of patients which would have given their consent, but the data was not used because the hospital wants to bias the study.

To better detect data manipulation by hospitals, patients have to provide their consents in a digital form to also ensure that the data that was used by the hospital was correct. This, in turn, requires a clear definition of a consent model as well as a thorough understanding of how consents are initiated, managed or revoked.

The remainder of this document is structured as follows. Section 4 presents the employed methodology to achieve the desired objectives. Section 5 lays out the results of this deliverable. In particular, it defines a threat model by identifying possible security and privacy threats that could influence future design decisions and impact how consents are defined and managed. Additionally, it proposes an architectural design for managing consents using blockchain technology and smart contracts. Finally, Section 6 states open questions.

4 Methodology

In D6.1, we conducted multiple iterations to gather and analyze the requirements of WP6. We also conducted a feasibility study in order to evaluate how FeatureCloud could benefit from blockchain technology properties to conduct federated machine learning of healthcare data. We identified several technical challenges, but also opportunities for improving audits and managing user rights, thereby giving more control to patients.

In this deliverable D6.2, we build upon those requirements and employ a proactive strategy to evaluate risks. Therefore, a threat model was defined, which identifies potential threats, and develops and tests a set of mitigation mechanisms. The threat modelling follows a mixture of the

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

threat modelling defined in the Microsoft SDL[1] and attack trees[2], where threats are modeled as attack trees with attack goals as roots and alternative ways to achieve that goal as tree branches. Procedures to detect and respond to those threats are also developed. Possible threats are categorized and grouped by the corresponding authors. The main attack goals identified are as follows.

- Healthcare data of individual patients is leaked from a hospital or a FeatureCloud study
- The outcome of a FeatureCloud study execution is manipulated

Based on the outcome of the threat model, we elaborated two design solutions and evaluated how they can respond to each of the threats.
Taking into consideration this threat model, we developed an artifact for managing consents. This includes a basic model for consent creation, update and revocation. Additional iterations over this basic model will be elaborated as future work in order to improve it and address additional technical challenges. The current model is important to test the applicability and feasibility of our approach in the context of the FeatureCloud project.

5 Results

5.1 Threat Model in FeatureCloud

This threat model focuses on threats that influence basic design decisions in the overall FeatureCloud security architecture, with a special focus on user/rights management. This threat model does not include general IT-Security threats like insufficient input validation etc. For some threats for which there are classical IT security best practices or mitigation techniques, the suggested mitigation approach is listed directly after a short description of the threat. This is also the case for mitigation techniques that are not directly related to the scope of FeatureCloud. The other threats are dealt with through the overall design of the user/rights management in FeatureCloud.

5.1.1 Actors

- **Auditor** (e.g., Health ministry) *(Trusted)*
  - Checks the integrity of the studies
  - Checks that docker containers for studies are not malicious and exfiltrate data.
- **Coordinator**
  - Runs the overall study
  - Creates the workflow and invites the participants
- **Patient**
  - Physical person from which biomedical data was obtained
  - Gives, updates and revokes consents
- **Participant** (e.g., hospital)
  - Participates in studies (e.g., hospital)
  - Provides the data and patient consents
- **Local project manager (LPM)**
  - Manages the IT infrastructure of a participant (e.g., hospital) including patient data
  - Digitizes paper consents from patients

[1] https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling
[2] https://www.schneier.com/academic/archives/1999/12/attack_trees.html

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

- Fetches eligible data from patients and prepares it for the ML studies
- Controls the local FeatureCloud instance
- *Trusted to not leak patient data they have access to directly*

Additionally the **advanced solution concept,** which will be described in a later section, requires a trusted-third-party (TTP) which verifies and digitally signs that a certain cryptographic keying material belongs to a patient. For patients which are not able to manage their own cryptographic keying material with their devices, the TTP manages these keys on their behalf.

- **Trusted-Third-Party (TTP)** *(Trusted)*
  - Verifies the identity of a patient and digitally signs that a certain cryptographic keying material belongs to a patient.
  - If the patient desires, the TTP also manages the cryptographic keying material on behalf of the patient.

5.1.2 Attack Goals/ worst case scenarios:

The main goals, or the worst scenarios of an attack on the FeatureCloud ecosystem can be summarized as follows:

- Healthcare data of individual patients is leaked from a hospital or a FeatureCloud study
- The outcome of a FeatureCloud study execution is manipulated

5.1.3 Threats related to the Coordinator:

a. Coordinator provides a malicious ML algorithm to Leak data

The coordinator provides a malicious ML algorithm which exfiltrates sensitive patient data. This can either happen through leaking the input patient data directly, or through the model which leaks information about the input data in the result. Then the Local project manager (probably even unwillingly, by accident) leaks data (e.g., if the docker container is malicious).

> **Mitigation:** The FeatureCloud study runs in a sandbox. This execution environment limits the actions the docker container can perform, e.g., open connections to exfiltrate data. Apps are reviewed before they appear in the AI Store as certified, all participants are warned before they run an uncertified app.

b. Coordinator provides a malicious ML algorithm to manipulate the outcome

On the other hand, the coordinator can craft a model that will always yield the result that the coordinator wants, regardless of the input.

> **Mitigation:** The auditor has to check the provided docker containers and algorithms before they are allowed into the FeatureCloud AI Store. Therefore, it should be ensured that only meaningful studies are performed in which the outcome is not rigged.

c. Coordinator manipulates the aggregation of the model

The Coordinator only aggregates selected results of the federated learning.

> **Mitigation:** Since the local project manager commits an aggregated version of the result (in the best case a hash) to the blockchain, the auditor can later reperform the aggregation of the different ML models (results) of the FeatureCloud study to check if the outcome is consistent with the outcome produced by the coordinator. Since the results of the FeatureCloud study (the ML models) are not confidential, this step can theoretically also be performed by other parties.

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

5.1.4. Threats related to the Local project manager:

i) Local project manager manipulates input data

a. Local project manager selects non consented data, or data with expired/revoked consent.

The local project manager (LPM) is responsible for fetching and preparing the data that will be included in a study. It is therefore possible for the local project manager to include data that is not consented. Moreover, the local project manager might use data, which has expired or where the patients' consent was already revoked at the time of conducting the studies.

b. Local project manager uses arbitrary or fake data as input to the ML algorithm

The local project manager might include arbitrarily selected or fake data to get considered for a specific study, or to generally influence the results of studies. For example, if the LPM is aware of a study that compares which vaccine is more efficient towards a Covid-19 variant, an LPM could influence the outcome of the study by using fake data in order to bias the outcome toward a particular vaccine.

c. Local project manager uses incomplete data as input to the ML algorithm, i.e., excludes eligible data

Similarly to b), the local project manager might also intentionally exclude eligible data from a study to influence the results. For example, to bias a study on the effectiveness of a particular vaccine, an LPM could influence the outcome by intentionally excluding data items that show unfavourable results.

d. Local project manager deletes or excludes available data before a study could have been run that uses this data.

The LPM can selectively delete data or decide not to digitize it before any FeatureCloud study could have been announced which would like to use this data e.g., delete all medical records of patients that have been vaccinated with a certain vaccine or pre-emptively exclude data from digitization that could be interpreted in an unfavourable manner.

Mitigation: General IT security best practises and four eyes principle has to be adhered to within the IT of the hospital. See also the discussion at the end of this section.

ii) Local project manager manipulates consents

a. The local project manager issues fake consent or manipulates the scope of consents.

The local project manager might issue fake consent forms (on paper), or manipulate the scope of the consent (on paper).

b. Tricking the patient into giving excessive consent.

The local project manager tricks patients into giving consent for a study that the patient does not intend to give consent to.

iii) Local project manager manipulates audit entries

a. Local project manager deleted data, so that it is no longer available for audit.

The local project manager might intentionally delete data that was used for a specific study. Alternatively, non-malicious failures such as data loss due to hardware issues on-site at the local project manager can lead to similar effects as if the local project manager acted maliciously.

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

Mitigation: Technical failures should not influence an ML study, this has to be ensured by general IT security practices at the hospital. See also the discussion at the end of this section.

b. Local project manager does not create an audit log entry.

The local project manager might not create an audit log entry for running the ML study.

Mitigation: All participants and the FeatureCloud backend make sure that there is a valid log entry for the WF run before starting to execute it.

iv) Local project manager leaks output data / ML models

a. The local project manager can leak the ML model (result).

The local project manager can leak the result of the docker container execution (the ML model).

Mitigation: This is mitigated by the general design of FeatureCloud, the result should not give away sensitive information.

b. Local project manager can leak the ML algorithm

The local project manager can leak the docker container including the ML algorithm.

Mitigation: ML algorithms must not contain any sensitive data, so that leaking them would not infringe privacy or security. Also, most apps are open source anyways since FeatureCloud rejects security by obscurity and fosters open security and security by design.

Discussion:

Due to the possibility of technical failures in the processes assigned to the local project manager, it is possible that failures exhibit the same outcome as a malicious project manager would. This is in particular the case for omission failures, e.g. data loss of digitized consent forms, but can also occur during the digitization process, e.g. flipping of digits in patient data, erroneous OCR etc. Hence, even if we assume a strong system model where the local project manager is completely trustworthy it is necessary to contemplate the possibility of (random) technical failures and their potential impact to the correctness and security of the conducted ML studies.

5.1.5. Threats related to the Auditor:

a. Auditor leaks data

The auditor leaks data when they are reproducing the results of the docker container execution with the correct input data.

Mitigation: The Auditor is trusted to not leak patient data.

5.1.6. Threats related to the Patient:

a. Patient is hacked, or unwillingly gives/revokes consent.

If the consent of patients is handled in a digital way, i.e., through cryptographic signatures, then it is possible that the private key of the patient gets compromised.

Mitigation: The patient must be able to revoke consent, and define a new key pair, through physical presence at the hospital or a point of service.

b. Patient requests data deletion after a study

One of the legal rights of patients with respect to GDPR is to have their data deleted if requested. This, in general, does not represent a threat, but may influence audit results where the actual data is compared to the hashes (audit records). Therefore it becomes impossible to check whether the

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

corresponding hash stored for auditing purposes corresponds to the data used for a study. Besides, this may influence a possible replay of a study on the input data as some of the data were deleted. The deletion request might be motivated by privacy reasons at the patient side, but could also be triggered by a malicious local project manager who created fake patient data or even by a set of malicious patients in order to make an integrity check impossible.

Mitigation: After a data deletion request, a timeframe should be defined before the request gets executed. During this time interval an audit can be carried out.

5.2 Overview of basic solution concept: Paper consent

The following steps describe a basic solution approach which relies on paper consent forms that are signed by a patient using a pen.

1. **Patients give consent**
   In a first step the consent is given in paper form by the patient. The paper with the written signature of the patient also has to contain some high level information (e.g., consent to all data, consent to use cancer treatments only, consent to use all data but only for breast cancer studies) on the underlying data for which the consent is given. Thereby, an auditor can later use this information to spot large deviations in the data that was used in the study compared to the data for which consent was given to.
2. **The local project manager executes the study**
   When the LPM participates in a FeatureCloud study, the docker images are downloaded and run locally with the input data from the patients. For this, the LPM collects consent forms and the associated data. Hereby special care has to be taken that docker images are only downloaded from trusted sources, i.e., the FeatureCloud AI Store.
   a. After the study has been executed locally, a hash of the output (ML model) is submitted to the FeatureCloud blockchain, together with a hash of the input (patient data).
   b. Moreover, the result of the study has to be submitted to the coordinator for further processing.
3. **The FeatureCloud study is over and the results can be aggregated**
   The FeatureCloud study is over and all results have been aggregated by the coordinator. Since the results (ML models) have been hashed and submitted to the blockchain, it is possible for an auditor to later redo the aggregation step and compare the result with the result of the coordinator. This should prevent the coordinator from excluding certain results. Since the results are not confidential, sharing this data with the auditor does not create any additional data privacy risk.
4. **Randomly sample local project managers and their used data items for audit**
   To mitigate manipulation attempts from LPMs and also act as a deterrent, a subset of LPMs and their utilized data items are sampled so that they can be audited by the auditor. This approach avoids the necessity for full audits in which the auditor could gain complete knowledge of all utilized data (which would mitigate the advantages of FeatureCloud) with the trade-off of only providing probabilistic guarantees in regard to manipulation detection. To ensure that the randomness of the sample is not biased, methods of distributed random number generation can be used.
5. **The randomly sampled local project managers are audited**
   The Auditor visits randomly sampled hospitals and requests to see all the input data and the associated consent forms of input data.

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

a. After spot checking the consents, the auditor reruns the study on their own hardware and checks if the results are within bounds to the original results. In the best case the ML study is deterministic and the result is exactly the same.
b. If the Local project manager(s) who have been audited only used data for which consent exists, and if the result did match, then everything was in order.
c. If this is not the case the local project manager faces legal repercussions.

6. **The LPM can delete retained input and output data of the study.**
   All hospitals can now delete any data that was specifically retained which belongs to the study, including the input data of the patients and the output (ML model).
   This prevents compliance difficulties with corner cases of patients revoking their consent later on as no input data has to be kept specifically for FeatureCloud studies. If the ML model is kept by the coordinator this is not a problem as the model should not leak sensitive data anyway.

***Both** the advanced and the basic concept need either an **auditor** who checks the result, or **trusted computing** (e.g. SGX or on-premise sealed computers) to verify that the computation was done correctly and with the correct input data of patients (for which also consent was given by the patients).*

Without trusted computing version:

- *Auditor comes and demands access to the input data used for this FC study.*
- *Auditor checks that the input data is consistent with the hospital's patient data and has not been tampered with*
- *Auditor verifies that consent has been given for all the data that was used in the study.*
- *Auditor recomputes FC study and checks if the result is within bounds of the original result (on his own uncompromised hardware). In the best case the study is deterministic and the result is exactly the same.*

Trusted computing version:

- *TPM computes FC study and provides an attestation that the FC was executed correctly and a signature over the used input and the output (Matetic et al., 2019).*
- *Thereby, the auditor only has to check if the input that was used by the TPM was correct and that there is a valid consent for all the used input data. The recomputation of the study can be omitted since the TPM ensures that the computation was done correctly.*

5.3 Overview of advanced solution concept: Patients (or TTP on behalf of the patient) can use cryptographic techniques

In this advanced solution concept, the patients (or a trusted-third-party (TTP) on behalf of the patients) can use cryptographic techniques and manage the associated cryptographic keying material. This could for example be a smart phone application.

The main difference with this approach is that the consent is given in a digital form and can thus be checked automatically.
The main steps of the basic solution concept are similar also in this case.

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

Advantages of the advanced solution/concept:

+ This makes the task of the auditor easier because they can computationally verify that all the consents for the data a LPM claims to have are correct by checking the signature of the consent and the underlying data.
+ LPMs cannot create fake patients, if the TTP takes care of creating them.
+ The advanced concept also allows patients and the auditor to detect if the LPM has manipulated their data. By signing the underlying medical data (hash of the data) in their consent the patient ensures that they really had the described medical treatment / disease. This double checks that the data the LPM can use in FC studies is correct.

Additional challenges for advanced concept:

- **Patients have to be verified by some TTP**
  For the beginning we can assume that the hospital is doing this and this is secure. In the end this has to be done by another party to prevent LPMs from creating fake patients with fake medical records.
- **Patients have to be verified by the hospital**
  It has to be verified by the hospital that the patients actually have been treated for a given reason by this hospital.
- **A patient (or a TTP on behalf of the patient) has to provide a cryptographically signed consent**
  The consent must sufficiently describe the scope for which it is given, without leaking too much information on medical records or patient data. Moreover it must not be possible to identify the patient with the provided information of the scope alone.
- **The hospital has to be able to link the given consents with the actual data of the patients that still reside in the hospital.**
  The actual plain text data of the respective patient has to be used to perform the study:
  - ➔ This can happen either locally at the hospital by participating in a federated machine learning study, as in FeatureCloud.
  - ➔ *Outside the context of FeatureCloud, this can also happen centrally, when hospitals transmit the data on which patients have given their consent to - to the entity which performs the study. The entity which performs the study can then check if the data provided by the hospital is the same data to which the patients have given consent to.*
- **The revocation of identities has to be handled by the System**
  It should not be possible to use such revocation data to identify patients. Once consent is revoked this revocation should be reflected by the system in a timely manner. Revocation of consent should not lead to a situation where the correctness of already performed studies that rely on data for which consent was later withdrawn can no longer be validated.

High level architecture requirements for the advanced concept:

- There needs to be a TTP (issuer) that verifies that this (d)id belongs to an actual person.
  - This must not be the LPM itself, because otherwise the hospitals can create fake patients.
- LPMs have to issue a certificate for a patient/(d)id which verifies that this credential belongs to a real patient of this hospital and the data item which is stored at the hospital and belongs

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

to this patient. This can be done when the patient is still in the hospital.

- If a patient desires, the TTP can act on behalf of the patient and create, manage and use cryptographic schemes on behalf of the patient.
    - This assumption can be lifted in the future as more patients are able to handle apps/programs which create, manage cryptographic credentials and signatures themselves.

How the patient data is stored at the hospital is not in scope of FeatureCloud and its consent management. The hospital has to ensure that The patient data is always stored encrypted at the local hospital of the patient, but provides various access policies, and only serves the data to an authenticated user., If the patient loses their keys for the consent management, data is still there and not lost. Also, in case of emergency, it has to be ensured that patient data is always accessible.

- All consents can then be checked automatically by the auditor in an audit by checking if there is a valid signature from a valid person who has been a patient in the hospital.

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

5.4 Threat overview

Threat	Addressed in basic solution concept	Addressed in advanced solution concept	Comment
Coordinator provides a malicious ML algorithm to Leak data	yes	yes	This threat is addressed in the general FeatureCloud environment by providing a sandbox which is not able to leak data directly to servers on the internet
Coordinator provides a malicious ML algorithm to manipulate the outcome	yes	yes	The auditor is responsible for only allowing checked FeatureCloud docker containers into the FeatureCloud app store
Coordinator manipulates the aggregation of the model	yes	yes	The auditor or any other party with access to the committed intermediate results in the blockchain can redo the aggregation.
Local project manager selects non consented data, or data with expired/revoked consent.	yes	yes	Due to the required commitment of the input data into the blockchain, the use of non consented data can be detected during an audit.
Local project manager uses arbitrary selected or fake data as input to the ML algorithm	partially* * Only detectable if the auditor is capable to link the consent to the actual data that has been used in the study	yes	To address this issue the advanced concept requires that the digitally signed consent encompasses the treatment or medical record of that patient. Thereby, it is harder to argue a relevance for a certain study if there is none.

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

Threat	Addressed in basic solution concept	Addressed in advanced solution concept	Comment
Local project manager uses selective / incomplete data as input to the ML algorithm, i.e., excludes eligible data	no	no	This has to be addressed by the design of FeatureCloud studies themselves. It must not be possible to derive the intended purpose of a study before the study has been finished.
Local project manager deletes available data before a study could have been run that uses this data.	no	no	General IT security best practises and 4 eyes principle have to be adhered to at the hospital.
The local project manager issues fake consent or manipulates the scope of consents.	partially* * Only detectable if the auditor is capable to recognize fake paper consent forms of non-existent patients	yes	To address this issue the advanced concepts require that only data from real patients can be used. This is ensured by a TTP that signs the cryptographic keying material of a patient.
Tricking the client into giving consent for some other thing.	no	no	The patient should be able to revoke already given consent.
Local project manager deleted data, so that it is no longer available for audit.	yes	yes	As soon as the data is committed to the blockchain, any deletion of the data will lead to legal consequences if detected in an audit.
Data Loss due to hardware issues on-site at the local project manager.	yes	yes	As soon as the data is committed to the blockchain, any deletion of the data will lead to legal consequences if detected in an audit. General IT security best practises have to be adhered to at the hospital.

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

Threat	Addressed in basic solution concept	Addressed in advanced solution concept	Comment
Local project manager does not create an audit log entry.	yes	yes	If the LPM does not create an audit log entry, the results he produced are not eligible for the aggregation within the FeatureCloud study.
The local project manager can leak the ML model (result).	yes	yes	This should not be a problem for the current design of FeatureCloud, as the ML model (result) should not contain sensitive data
Local project manager can leak the ML algorithm	yes	yes	This should not be a problem for the current design of FeatureCloud, as the ML algorithm (docker container with code) should not contain sensitive data
Patient is hacked, and unwillingly gives/revokes consent.	yes	yes	It has to be ensured that the patient can revoke consent. Since all data that was used as input to a docker container to produce the ML model (result) can be deleted as soon as some hospitals have been selected for an audit, no records are required to be kept about the execution of the study.
Patient requests data deletion after a study	yes	yes	This threat is mitigated by auditing FeatureCloud studies right after they have been executed. This way, all associated data can be deleted as soon as the audit of a random sample of participants has finished. So if there is a permissible delay between a patient's request for data deletion and the actual deletion that is within the duration of a FeatureCloud study everything can be fulfilled.

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

5.5 Preliminary Architecture for Consent Management

In this section, a preliminary design for consent management based on blockchain technology and smart contracts will be presented. In particular, the network architecture for consent management including access rights will be defined and a basic consent model will be elaborated, which will enable auditors to check whether or not a study result was based on consented data.

In contrast to using traditional solutions or local IT infrastructures for managing consents, e.g, databases, blockchain will serve as an audit trail that prevents participants from **tampering with** the data or consents, thereby storing the data hashes as well as consents on blockchain. This will also prevent a participant from creating fake consents, and at the same time provide a **transparent** and **secure** way for patients to track the use of their data (e.g., in which studies they have been included).

5.5.1 Actor Description

In the following sections, we present the architectural design of the FeatureCloud blockchain network system. As aforementioned, the main actors in the FeatureCloud system are the coordinators, participants, auditors, and patients.

The **coordinator** initiates FeatureCloud studies, defines the workflow, and provides the machine learning algorithm.

The **participants** are the data providers, e.g., a hospital, pharmaceutical company. They are responsible for local data security, perform data querying and preparation for ML studies. They manage patient's identities and consents

The **auditor** initiates and conducts the **auditing** processes for the studies.

Finally, the **patient** is the actual data owner, whose digital consent (and thus is identity within FeatureCloud) might be managed by the data provider (e.g., hospital) or another trusted-third-party (TTP). Patients give and revoke consents for using their data in ML studies. This happens either in paper form through a handwritten signature which is translated into a digital consent by the hospital, or directly in digital form via interacting with the FeatureCloud consent infrastructure. If the consent is given in paper form, the correct translation has to be checked on-site as well by the auditor during an audit.

5.5.2 Network Architecture

The FeatureCloud blockchain network is a communication network that relies on a public key infrastructure (PKI) to ensure that the communication between different network actors, i.e., nodes, is secure and guarantees that correctly published messages can be authenticated by other nodes. Therefore, every joined node in the system keeps a white list of trusted nodes, which means that any medical institutions who want to join the FeatureCloud network must first be approved by the FeatureCloud authority (Root CA). This authority may be a consortium, which includes the health ministry, auditors, participants, but for now, we abstract on these details and we will address the challenges on how this authority is created and managed, and how stewards (i.e., new members) are added or removed from this authority for the next deliverable.

As shown in Figure 5.1.1. Every actor presents themselves as a node within the FeatureCloud network and keeps a synchronized copy of the FeatureCloud's information in their node’s database. What kind of information is saved in the database is described in the next section. Moreover, once a node propagates a new transaction to the FC network, the "majority endorsement" policy is used. That means the majority of the nodes should execute the transaction as defined in a smart contract and validate it.

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

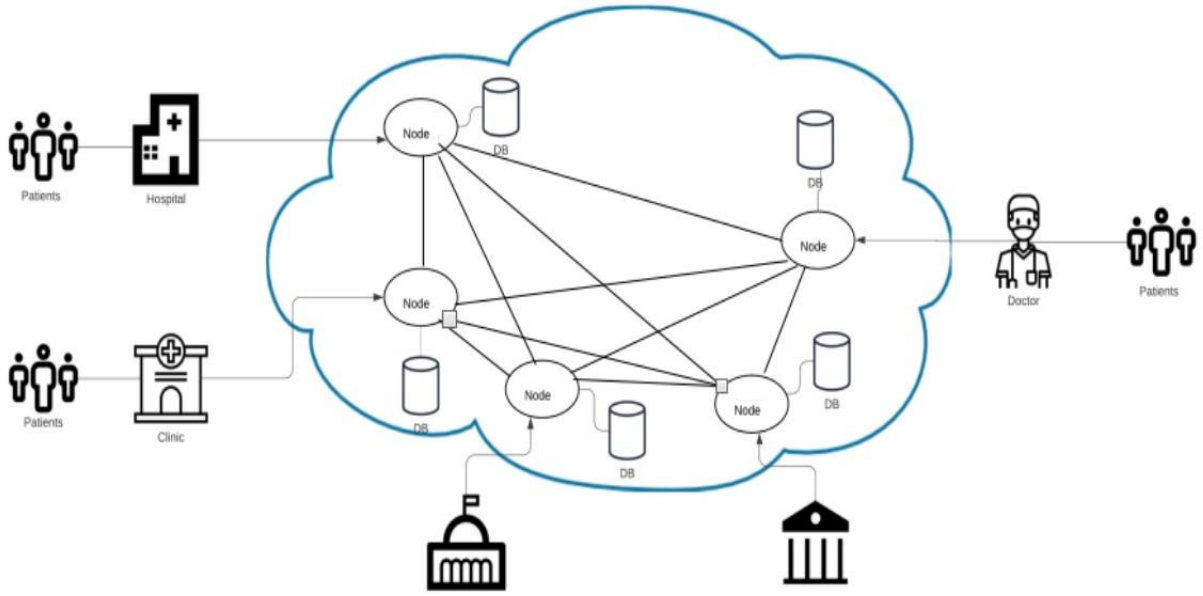

Figure 5.1.1 FeatureCloud Blockchain Network Architecture

5.5.3 FeatureCloud Entities

A. Consent

In the FeatureCloud network, the actual healthcare data is not saved on blockchain (or node copy of the ledger), but locally at the participant storage. Therefore, a blockchain-based mechanism for managing data consents and recording what data has been used ML study executions is required. Every "Consent" represents an agreement to use healthcare data records of a patient for a specific or all studies. The "Consent" is issued by the patient (digitally or using a handwritten signature) that provides their medical data. It is also possible that a trusted-third-party issues a digitally signed consent on behalf of the patient.

A "Consent" has three states, namely, "granted," "revoked", and "expired". Once a "Consent" is issued, as default, it gets the "granted" state.

Moreover, every "Consent" has an expiration date. After reaching the expiration date, a Consent is set to "expired". At the same time, patients are allowed to revoke their consent and require the data provider to update the consent’s state in the FC network. In this way, the "Consent" is set to the state "revoked". Of course, it is also possible for a patient to renew/update their consent in the system.

As shown in Figure 5.5.3.1, we illustrate the possible state transformations of a "Consent". However, if the patient is able to interact with the FC network directly is not yet decided. In this deliverable, we assume that patients can only manage their consent offline through their data provider.

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

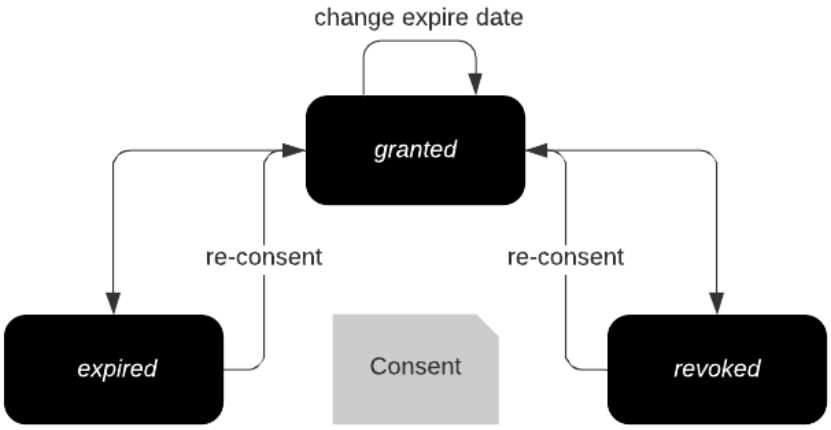

Figure 5.5.3.1 Consent’s states

The following Figure 5.5.3.2 shows the entity "Consent" of the FC network in detail. Besides the signature over the transaction and all including data from a legitimate key, It includes the following information:

1. a unique ID
2. organization (entity who gave the digital consent on behalf of the patient)
3. issue date
4. expiry date
5. type of the medical data
6. granularity (e.g., for all studies, for selected studies)
7. state (granted, revoked, expired)
8. data hash

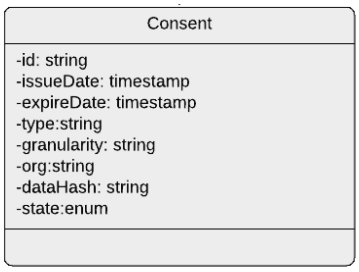

Figure 5.5.3.2 Consent’s attributes

Data hash saves the patient's identity and its real medical data hash. This attribute ensures data’s integrity and thus guarantees the trustful audit process later.

B. ML Study

The coordinators of FeatureCloud publish the machine learning study. They initiate and coordinate the process and provide a machine learning application that the participants can execute locally. All

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

data providers could decide whether or not to join the study.

In the FC network, a machine learning study has a unique ID. The other information about this study, like study description, docker container of the ML application, etc., is held somewhere else (E.g., FeatureCloud homepage, Docker Hub).

C. ML Study Result

Once the data provider chooses to join an ML study, they first need to perform the data discovery to check and compute the data which are relevant to this study. It is mandatory to use only the data that has valid consent in the FC network. It is also necessary to record the corresponding patient's consent of the used medical data for the study and submit the execution information to the FeatureCloud network. In other words, the submitted execution details should contain a list of the consent that refers to the patient's local medical data. As described above, every such consent is saved in the distributed ledger. All participant nodes in the network keep an identical copy of it. The execution detail contains the following information (See Figure 5.3.3.3):

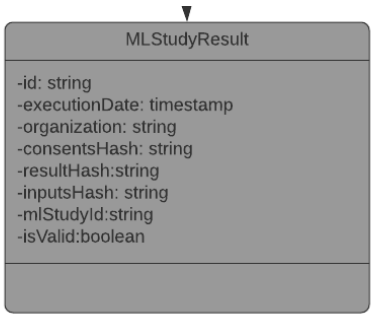

Figure 5.3.3.3 MLStudyResult’s attributes

1. a unique ID
2. machine-learning study ID
3. execution date
4. organization
5. result hash
6. a hash of Consents' ID
7. a hash of study inputs
8. validity (this attribute is set by the auditor)

5.5.4 Smart Contracts

In section 5.5.3.1, we present the architecture design of the FC network. Every node in the FC blockchain network has its own database, which holds the current state and history of the entities (Consent and MLStudyResult) of FC. A smart contract here defines the executable rules and logic as a program.

Those smart contracts are installed in every node in the network. The execution of a method in a smart contract will either trigger a transaction or return a previous or current state of a system entity.

A. ConsentContract

A ConsentContract is an executable program installed in the FC nodes. It manages the patient's consent of medical data. In Figure 5.5.4.1., we describe the functions defined in this smart contract.

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

By calling "issue" the data provider propagates a new consent to the FC network. Suppose the majority of network nodes have approved the message. Then, a new consent is written in the node's database. By calling "revoke" an existing consent is going to be revoked. This transaction is going to be recorded in all nodes' databases. Besides, "re-consent" helps to give consent again on an expired or revoked consent, and "changeExpireDate" makes changing the expiry date in a consent possible. All those functions require data ownership, which means only the data provider of the data is allowed to modify the consent. The functions that require ownership are marked with green color in Figure 5.5.4.1..

Furthermore, "getAll", "getByTypeAndOrg" and "getById" help the data provider or ML study participant to quickly filter the data and keep a synchronized view of consent both locally and remotely in the FC network. Suppose a hospital wants to take part in a ML study. This hospital calls the function "getByTypeAndOrg" of ConsentContract. In return, it gets a list of the valid consents of a certain type of medical data, which can be used for this ML study.

Last but not least, we need to take the audit process into account. The function "isValidOn" is for this purpose. During auditing, the auditor takes the date of the ML study execution and the consent ID as parameters. The same contract returns the validity of this consent at the ML study execution date. Thus, by retrieving all consents from the ML study the auditor can verify if a submitted result is valid.

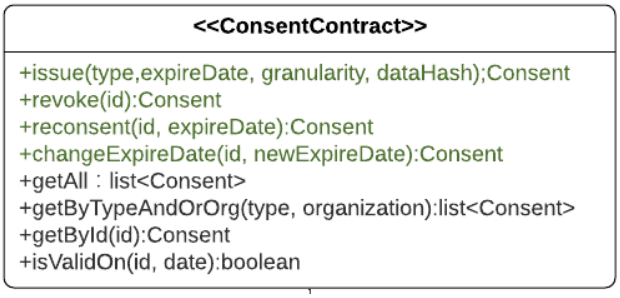

Figure 5.5.4.1. ConsentContract

B. MLStudyContract

Whereas the ConsentContract handles the consent management, MLStudyContract manages the machine learning study result submitted by participants. This smart contract provides two functions that trigger the write action in the node's database and three read functions that deliver the ML Study Results. "submitResult" is called by machine learning study participants. By calling it, a participant submits the execution result of a study, following parameters must be submitted at the same time:

1. execution time
2. machine-learning study id
3. a hash of consents' ID
4. a hash of the execution result

The hash of consents' ID of the corresponding medical data and the execution time facilitate the audit process by the auditor. According to the transaction record of consent in the network, one can easily recognize whether consent is valid at a certain timestamp and thus verify if the used medical data for a study has valid consent at the study execution time point.

The execution result of the study won't be kept in the FC blockchain network but somewhere else, like locally on the data provider side. However, there is a hash of the execution result, which assures

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

the integrity of the result. The consent ID of used data is also saved locally on the participant’s database. In the FC network, we save a hash of the id list to ensure that locally saved ids haven’t been tampered with.

Furthermore, this MLStudyContract provides the functions to quickly search all submitted ML Study Results by ML study ID. As health ministries, they can update the validity of a result by calling "setValidityOfStudyResult".

Figure 5.5.5.2 MLStudyContract

5.5.5 User Activities

In this section, we illustrate the user activities as an activity diagram in Figure 5.5.5. As shown in the diagram, there are three main workflows for interacting with the FC blockchain network, and two kinds of stakeholders are involved.

Three workflows are presented, which include consent management, machine learning study submission, and auditing. The participant is connected with consent management. In this scenario, we can either call them data provider or data owner. Additionally, they participate in the submission of the machine learning study result. Those who have the permission, like the health ministry, are involved in the auditing workflow.

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

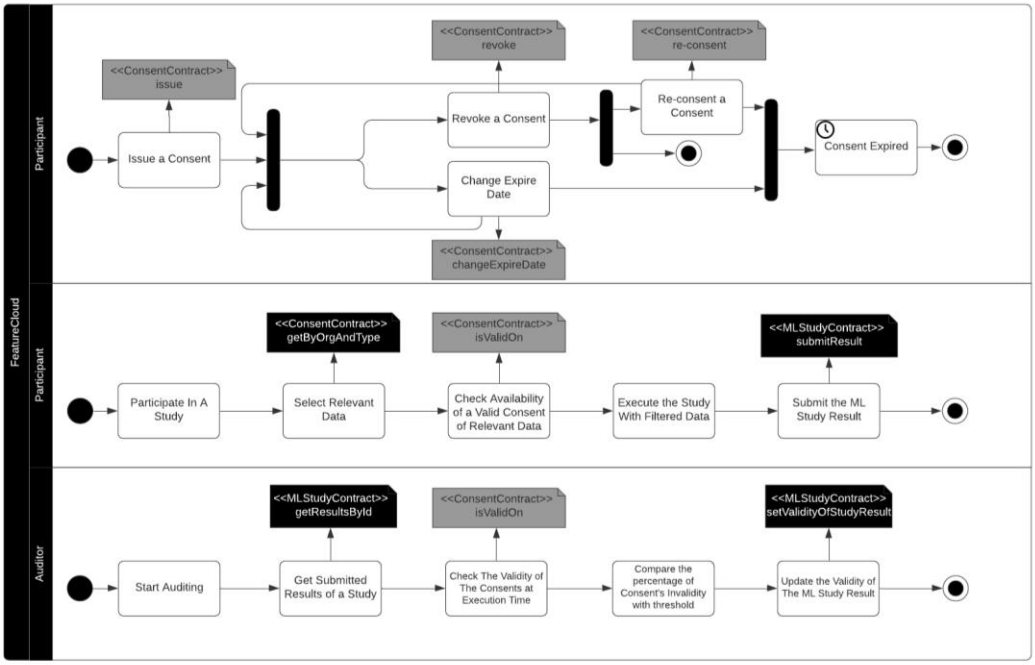

Figure 5.5.5. FeatureCloud Activities

A. Consent Management

As a data provider who originates the medical data, they are eligible to issue a new consent and reference it to the corresponding local medical data. The complete medical data context of a patient is authenticated by some hashing algorithm and submitted at issuance. Furthermore, expiration date, type of the medical data, and the granularity of the consent (for all studies, for selected studies, etc.) must also be considered.

Participants are registered organizations or private entities (doctor or clinic) in FeatureCloud. The FeatureCloud Certificate Authority issues the digital certificate to those parties. They authenticate to each other with the Public Key Infrastructure. From this point of view, a data provider broadcasts the issuance of consent to all registered nodes in the FC network. This message or let’s call it transaction is written into the FC blockchain network and thus immutable.

After the issuance, consent is valid before its expiration date. Although, the data owner (patient) can still revoke the consent offline from the data provider or change the expiration date. It is also possible to re-consent a revoked consent. A successful update on consent in the network is not anymore changeable due to the blockchain’s inherent properties. This characteristic makes the auditing of the machine learning study execution of the participant feasible. Based on those transaction records, the auditor can easily verify if the consent of used data is valid at the study execution time.

In Figure 5.5.5., we show that every action on consent is triggered by calling a function in the installed smart contract in the FC network on its node. For creating a new consent, the data provider calls the “issue” function in the node, its node then broadcasts this transaction to all active nodes in the network. This transaction is propagated to all node’s databases in the network once this transaction is approved by the majority of the system’s nodes.

B. ML Study Result Submission

The FeatureCloud system includes the global API where project details are stored and the AI store where FeatureCloud apps can be fetched to provide the comprehensive detail about a FeatureCloud

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

study. Besides, every study has an identical ID.

If a data provider decides to join a ML study, then eligible data is fetched. This includes checking whether selected data items have valid consents in the FC network. Only the data that has valid consent is going to be selected for the ML study execution. The participant must keep a record of the consent of used medical data. This record is a hash of all consents' IDs. The submission requires the execution time of the study, a hash of the consents’ IDs, a hash of the execution result, and the id of the ML study.

C. Auditing

In Deliverable 6.1, we discussed the issue of non-determinism in ML studies conducted through FeatureCloud, and how this poses challenges in regard to auditing and verifying that the execution was correct. Yet this problem remains as a critical open issue. In the following, we only consider the auditing of data consent for a study.

The auditor queries all related submitted records in the network. Our network returns a list of the ML Study Result related to this study. In addition, they retrieve the consent of used data for this study and check its validity at execution time by calling the smart contract function "isValidOn" of ConsentContract. If the digital consents have been issued on behalf of the patient, the underlying handwritten signature on a consent form has to be verified through spot-checks on-site at the hospital during an audit.

Therefore, a random sample of participants as well as data sets is drawn and audited in detail. Hereby, also the correct execution of the study is checked and the input as well as output hashes are compared to the timestamped results submitted to the blockchain.

As soon as the audit is finished the auditor updates the validity of the ML Study results and concludes the auditing process. The associated study results and metadata which have been kept to facilitate the verifiability and thus the correct execution of the study at the participants can now be deleted.

6 Open issues

Identities

Consent and healthcare data management currently relies on the assumption that identities of all actors involved in FeatureCloud are well defined, and that there is an established public key infrastructure PKI for using and managing cryptographic keys. In this deliverable, for example, the basic consent management model, where consents are given in paper format, assumes that participants (e.g., hospital) act on behalf of patients and correctly digitizes paper consents and manages them according to patient paper requests in a trusted manner. An audit can always check the paper version against the digitized version on chain (although a participant may not execute a revocation request and hide the revocation paper). Using common patient identifiers (e.g., insurance number) clearly puts in jeopardy one of the privacy objectives that FeatureCloud tries to solve by enabling cross correlation of patient data not only across participants, but also by the identity issuing entity (e.g., health insurance company).

Additionally, relying solely on participants to issue identities enables them to create fake identities and associate them with fake data and consents, thereby manipulating the outcome of the studies.

Therefore, it is of utmost importance to have a clear and secure identity model that prevents misuse and correlation of identities. As such, we would want to investigate further solutions that give patients control over their identities. On possible research direction could be the use of decentralized

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

identifiers (DiDs)[3], a standard proposal by the W3C, which are new types of identifiers that enable verifiable, decentralized digital identity. Future work will then investigate different DiD methods and check their suitability and applicability in the context of FeatureCloud (Coelho et al., 2018; Fdhila et al., 2021; Ghesmati et al., 2021; Lesavre et al., 2019).

Another research prospect is the use of verifiable credentials[4], i.e., a decentralized and privacy-aware type of credentials that may be associated with DiDs, and which enable decentralized verification of claims/attestations (Camenisch et al., 2011; Lagutin et al., 2019; Schanzenbach et al., 2019). Using verifiable credentials, a verifier does not need to communicate with the issuer in order to check the validity and integrity of a credential. An investigation on how to use verifiable credentials for modeling consents will enable privacy-aware and decentralized management and verification of consents.

7 Conclusion

Federated machine learning not only prevents healthcare data from moving outside of data provider local storages, but also minimizes the risk of creating single points of attack by keeping the data distributed. From a legal point of view, it also helps avoiding complicated compliance checks related to GDPR and transfer of data outside of certain jurisdictions. Despite these benefits, federated machine learning does not stop the actors involved in a study from behaving maliciously.

In deliverable D6.2, we have elaborated a threat model that identifies there, attack goals, actors and threats, and provides mitigation mechanisms on how to respond to such threats. The aim of this threat model is to minimize the attack surfaces in the FeatureCloud system, by providing means for identifying and handling threats. The second part of this deliverable focuses on a basic consent management model to prove the solution feasibility, and applicability of blockchain technology and smart contracts in this context. Future work will consist of elaborating over and improving the consent management model, providing a clear definition of identities and how they are managed, defining a clear governance process for the blockchain network, and implementing the smart contract on a blockchain solution.

[3] https://www.w3.org/TR/did-core/
[4] https://www.w3.org/TR/vc-data-model/

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

8 References

Camenisch, J., Kohlweiss, M., Sommer, D., 2011. Pseudonyms and Private Credentials, in: Camenisch, J., Leenes, R., Sommer, D. (Eds.), Digital Privacy: PRIME - Privacy and Identity Management for Europe, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 289–308. https://doi.org/10.1007/978-3-642-19050-6_10

Coelho, P., Zúquete, A., Gomes, H., 2018. Federation of Attribute Providers for User Self-Sovereign Identity. J. Inf. Syst. Eng. Manag. 3, 32.

Fdhila, W., Stifter, N., Kostal, K., Saglam, C., Sabadello, M., 2021. Methods for Decentralized Identities: Evaluation and Insights, in: González Enríquez, J., Debois, S., Fettke, P., Plebani, P., van de Weerd, I., Weber, I. (Eds.), Business Process Management: Blockchain and Robotic Process Automation Forum, Lecture Notes in Business Information Processing. Springer International Publishing, Cham, pp. 119–135. https://doi.org/10.1007/978-3-030-85867-4_9

Ghesmati, S., Fdhila, W., Weippl, E., 2021. Studying Bitcoin Privacy Attacks and Their Impact on Bitcoin-Based Identity Methods, in: González Enríquez, J., Debois, S., Fettke, P., Plebani, P., van de Weerd, I., Weber, I. (Eds.), Business Process Management: Blockchain and Robotic Process Automation Forum, Lecture Notes in Business Information Processing. Springer International Publishing, Cham, pp. 85–101. https://doi.org/10.1007/978-3-030-85867-4_7

Lagutin, D., Kortesniemi, Y., Fotiou, N., Siris, V.A., 2019. Enabling Decentralised Identifiers and Verifiable Credentials for Constrained IoT Devices using OAuth-based Delegation, in: Proceedings 2019 Workshop on Decentralized IoT Systems and Security. Presented at the Workshop on Decentralized IoT Systems and Security, Internet Society, San Diego, CA. https://doi.org/10.14722/diss.2019.23005

Lesavre, L., Varin, P., Mell, P., Davidson, M., Shook, J., 2019. A Taxonomic Approach to Understanding Emerging Blockchain Identity Management Systems. ArXiv190800929 Cs. https://doi.org/10.6028/NIST.CSWP.07092019-draft

Matetic, S., Wüst, K., Schneider, M., Kostiainen, K., Karame, G., Capkun, S., 2019. BITE: bitcoin lightweight client privacy using trusted execution, in: Proceedings of the 28th USENIX Conference on Security Symposium, SEC’19. USENIX Association, USA, pp. 783–800.

Schanzenbach, M., Kilian, T., Schütte, J., Banse, C., 2019. ZKLaims: Privacy-preserving Attribute-based Credentials using Non-interactive Zero-knowledge Techniques. Presented at the ICETE 2019 - Proceedings of the 16th International Joint Conference on e-Business and Telecommunications, pp. 325–332.

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

9 Table of acronyms and definitions

BFT	Byzantine Fault Tolerance
concentris	concentris research management GmbH
DiD	Decentralized Identifier
GND	Gnome Design SRL
LPM	Local Project Manager
ML	Machine Learning
MS	Milestone
MUG	Medizinische Universitaet Graz
Patients	In this deliverable, we use the term “patients” for all research subjects. In FeatureCloud, 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”.
PKI	Public Key Infrastructure
PoS	Proof of stake
PoW	Proof of Work
RI	Research Institute AG & Co. KG
SBA	SBA Research Gemeinnützige GmbH
SDU	Syddansk Universitet
TTP	Trusted Third Party
TUM	Technische Universitaet Muenchen
UHAM	University of Hamburg
UM	Universiteit Maastricht
UMR	Universitaet Marburg
VC	Verifiable Credential
W3C	World Wide Web recommendation
WP	Work package

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