D4.8

Specification of protocol and APIs for research health data sharing - V1

ABSTRACT

This deliverable provides specifications for the Research Data Sharing (RDS) Protocol that governs the process of collecting health data from citizens' smart health data, contained on their mobile devices, for the purposes of cross-border medical research. The deliverable defines the scope of the protocol within the entire process of setting up and preparing a research study. It also defines the actors involved, the process, as well as the underlying programming interfaces, high-level system components, and data models.

Delivery Date	February 26th, 2021
Work Package	WP4
Task	T4.3
Dissemination Level	Public
Type of Deliverable	Report
Lead partner	UNITN

InteropEHRate project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106.

This document has been produced in the context of the InteropEHRate Project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106. All information provided in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.

CONTRIBUTORS

	Name	Partner
Contributors	Gábor Bella	UNITN
Contributors	Simone Bocca	UNITN
Contributors	Stefano Dalmiani	FTGM
Contributors	Francesco Torelli	ENG
Contributors	Marcel Klötgen	FRAU
Contributors	Salima Houta	FRAU
Contributors	Sofianna Menesidou	UBITECH
Contributors	Chrysostomos Symvoulidis	BYTE
Contributors	Stella Dimopoulou	BYTE
Contributors	Vincent Keunen	A7
Contributors	Lucie Keunen	A7
Contributors	Martin Marot	A7
Reviewers	Nicu Jalba	SIMAVI
Reviewers	Francesco Torelli	ENG

LOG TABLE

Version	Date	Change	Author	Partner
0.1	2020-02-13	TOC and first draft created	Gábor Bella	UNITN
0.2	2020-04-10	Integrated observations from all partners, published new diagram and corresponding text	Gábor Bella	UNITN
0.3	2020-04-17	Further progress, started low-level protocol description	Gábor Bella	UNITN
0.4	2020-05-10	Added monitoring phase, improved protocol description	Gábor Bella	UNITN

		according to partner input		
0.5	2020-07-17	Added interface descriptions, data models, as well as sections on security and anonymization	several partners	several partners
0.6	2020-07-26	Prepared version for internal review	Gábor Bella	UNITN
0.65	2020-07-30	Finished data model section	Marcel Klötgen, Gábor Bella	FRAU, UNITN
0.7	2020-09-08	Internal review and improvements	Francesco Torelli, Gábor Bella	ENG
0.75	2020-09-23	Revision of data model descriptions, transferring content to D2.8 and merging the corresponding section with section 3	Marcel Klötgen, Salima Houta, Gábor Bella	FRAU, UNITN
0.8	2020-11-20	Updated data flow diagram by new version provided by Francesco Torelli	Gábor Bella, Francesco Torelli	UNITN, ENG
0.9	2020-12-16	Updated the sections on security and privacy with new requirements and specifications, extended the general component architecture and APIs with components related to security and privacy	Gábor Bella, Sofianna Menesidou, Thanassis Giannetsos, Stella Dimopoulou	UNITN, UBITECH, BYTE
0.95	2021-01-26	Last edits on security and privacy, version ready for 2nd internal review	Gábor Bella, Sofianna Menesidou, Thanassis Giannetsos, Stella Dimopoulou	UNITN, UBITECH, BYTE
1.0	2021-02-22	Integrated improvements suggested by internal reviewers.	Gábor Bella	UNITN
Vfinal	2021-02-25	Quality check and submission	Laura Pucci	ENG

ACRONYMS AND TERMS

Acronym	Term	Definition
CA	Certificate Authority	An institution that issues digital certificates.
CN	Central Node	A node of the Research Network (a server) that stores published research studies and provides a central access point to S-EHR Apps for retrieving the descriptions of research studies.
-	Citizen	Any person potentially participating in a research study and having the minimal technical means to do so, i.e. the S-EHR App installed on their smartphone.
-	Client	The (public or private) legal entity who has ordered the research study and is paying for it.
CRC	Coordinating Research Centre	A medical research centre that initiates a particular research study and is in charge of defining it and carrying it out.
PI of the Research Centre	Principal Investigator of a Research Centre	The researcher (person) in charge of the citizens enrolled for a specific study at a RC.
PI of the Study	Principal Investigator of the Study	The researcher (person) in charge of a specific study at the CRC.
PP	Pseudonym Provider	An institution that generates and provides pseudonyms as a service.
RDD	Research Definition Document	A document written in a formal, computer-processable language that describes the research datasets to be retrieved from citizens' EHRs, enrolment and exit criteria, as well as related metadata.
RDDI	RDD Interface	Application Programming Interface allowing the exchange of RDDs between the Central Node and the S-EHR App.
RDS	Research Data Sharing	Acronym of the Research Data Sharing Protocol, the protocol covered by this deliverable
RDSI	RDS Interface	Application Programming Interface allowing the exchange of consent and health data between the S-EHR App and Research Centres.
RN	Research Network	The network of research centres and technical nodes that implement the Protocol.
RRC	Reference Research Centre (of a citizen)	A research centre participating in a given study that is a reference point for a specific citizen. The citizen sends health data to it for the duration of the study and the reference research centre is responsible for monitoring the citizen during the study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Scope of the Document

The overarching goal of the Research Data Sharing Protocol (in the following: *the Protocol*) is to specify a set of remote APIs and constraints on their usage that provide the technical means to citizens for the sharing of their health data for the purposes of cross-border medical research, in a cross-border interoperable manner. The particularity of this protocol, as opposed to current practices in medical research, is that it puts the citizens in full control of the sharing of their data: after explicit consent, data are retrieved directly from a citizen's mobile device in an anonymized manner.

From a technical point of view, the Protocol is almost peer to peer and decentralised, in particular the citizen shares the health data only with specific research centres. Only the metadata that describes the ongoing research studies are centralised. Moreover, the Protocol does not tie the citizens and research centres to specific software products, but specifies just the APIs and constraints that the interacting software systems must support and satisfy.

The Protocol addresses the giving and revocation of consent, the enrolment into specific research studies, the verification of enrolment criteria, as well as the retrieval and transfer of relevant health data. The Protocol defines the human and automated actors, the operations, and the communication channels and interfaces involved in these processes.

1.2. Intended Audience

This deliverable is intended primarily for a technical audience, interested in implementing the Protocol described in the document, or in understanding how data collection for cross-border medical research can be carried out with a direct involvement of citizens and their mobile-based health records. A certain familiarity of the medical research preparation process and of the challenges of cross-border data collection is useful for the understanding of this deliverable. Furthermore, the reader is expected to be familiar with certain other standards, formats, and specifications designed or used within the InteropEHRate project, such as the FHIR standard [FHIR], the InteropEHRate Architecture [D2.5] and the profiles for EHR interoperability [D2.8].

1.3. Structure of the Document

Section 2 provides a high-level overview of the Protocol. Section 3 defines the high-level architecture, interfaces of the typical Research Network that implements the Protocol, and the principal datatypes they use. Section 4 describes the rationale and approaches used to anonymize and pseudonymize the health data shared by citizens. Section 5 describes the security aspects of the Protocol. Section 6 provides process definitions that clarify the interactions of the system components and human actors, in the form of activity and sequence diagrams. Section 7 describes the related work. Section 8, finally, provides conclusions.

1.4. Differences with respect to previous versions of the deliverable

Not applicable, this is the first version.

2. PROTOCOL OVERVIEW

This section provides an informal, high-level description of the goals, scope, participants, and processes covered by the Protocol. For a general high-level storytelling of how the Protocol is used to collect data for research studies[1], the reader is referred to the relevant section of deliverable [D2.2].

2.1. Goals and Scope

The Research Data Sharing Protocol addresses the general problem of collecting health data for medical research directly from citizens, possibly involving citizens from multiple European countries. The motivations underlying the solution presented here are (1) to give more control to citizens over the use of their health data for research purposes; (2) to allow citizens to participate in research studies also remotely through their smartphones; enable cross-border data collection in a way that involves citizens more directly in the decisions regarding the sharing of their data. This is achieved through a novel approach that retrieves data directly from the electronic health records stored on citizens' mobile devices. Citizens have complete control over their data as they can give or decline consent for data sharing on a per-study basis, and be informed of precisely what data are used by a given study.

In order to respond to the numerous technical challenges underlying such an approach, the Protocol brings novel solutions as well as relying on existing results from inside and outside the InteropEHRate project. It deals with the heterogeneity of cross-border data through relying on interoperable data representations, such as the *Interoperability Profile* defined by the InteropEHRate project [D2.8]. It automates data queries and the checking of eligibility criteria inside the mobile device. It addresses privacy constraints by in-device data anonymization. It ensures the security of data transmission between mobile devices and research centres by relying on state-of-the-art encryption techniques. It provides a formal framework for consensual data sharing through digital signatures.

For simplicity, in the rest of this document we will assume that the mobile device of the citizen is a smartphone, but the protocol actually applies to any mobile device of a citizen able to run a suitable mobile application supporting smart health data, such as the *S-EHR App* proposed by the InteropEHRate project.

[1] Note that in D2.2 a research study or its description are also called "research protocol". To avoid ambiguities, in the present document, the term "protocol" is used only to refer to the communication protocol for research health data sharing.

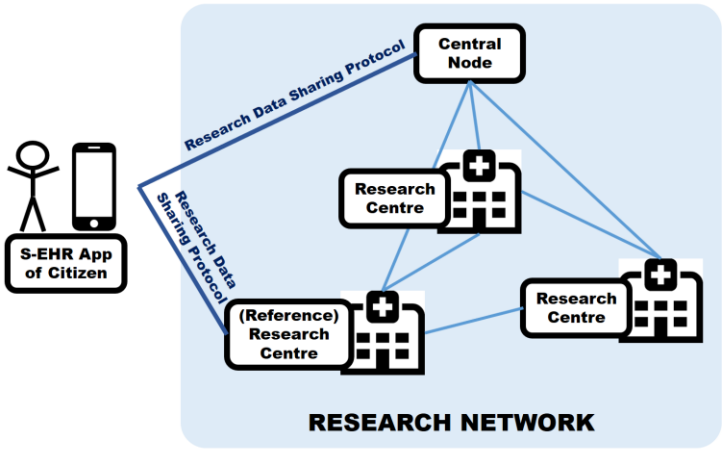

Figure 1 - High-level overview of the entities involved in the Research Data Sharing Protocol

Figure 1 shows a simple schematic diagram of the main components of a research data sharing scenario as assumed by the Protocol. The setup consists of (a) patients in possession of electronic health records stored on their mobile devices (the S-EHR App in the picture); and (b) a *Research Network* that consists of interconnected *Research Centres* as well as a *Central Node*. In the case of a so-called *multi-centric research study*, multiple research centres may simultaneously collect data for the same study, each citizen being formally "attached" to a single research centre that becomes his or her *Reference Research Centre* (RRC). The role of the Central Node is to store the formal, machine-interpretable definitions of research studies in the form of *Research Definition Documents* (RDD), and to provide these documents for download by mobile devices holding electronic health records. The S-EHR App then interprets the RDD and, in case the citizen is eligible and is willing to participate in the research study, retrieves relevant data from the citizen's health records and transmits them to the RRC in a fully secure and privacy-aware manner. The Research Data Sharing Protocol specifies the communication modalities between the S-EHR App and, on the one hand, the Central Node and, on the other hand, the citizen's Reference Research Centre.

2.2. Actors and Systems

The only human actors explicitly involved in the Protocol are *citizens.* A **citizen** is any person potentially participating in a research study with his/her health data, and having the minimal technical means to do so, i.e. a S-EHR App installed on his/her smartphone (or any other personal mobile device).

While not directly involved in any of the interactions in the scope of the Protocol, the following actors are participants of the overall research definition and data collection process and the semantics of some of the exchanged data is related to them. Moreover they are responsible for some of the systems involved in the protocol.

- **Principal Investigator (PI) of the Study:** the researcher (person) in charge of a specific study, including its formal definition. The PI of the Study produces the Research Definition Document and has it published on the Central Node of the Research Network.

- **Principal Investigator (PI) of a Research Centre (RC):** the researcher (person) in charge of the patients enrolled for a specific study at a RC. The PI of the RC monitors the process of patient enrolment and the retrieval of their data.
- **Central Node Administrator:** a single person in charge of overseeing at the Central Node the publishing of new research studies on the Research Network.

These actors intervene through the following systems:

- **S-EHR App.** An application installed on the Citizen's smartphone (or any other personal mobile device) that stores and manages the Citizen's health records, and is in charge of executing the Protocol implementation, that is, the library on the phone. It must fulfil the constraints specified by deliverable [D3.1].
- **Central Node (CN).** A node of the Research Network (a server) that stores published Research Definition Documents as defined by their respective PIs, and provides a central access point to S-EHR Apps for retrieving them.
- **Research Centre Information System.** The information system of a RC participating in a given study. It collects data shared by a set of citizens who are officially attached to this centre for the duration of the study.

The aforementioned actors also interacts with the following security-related systems:

- **Certificate Authority (CA).** A trusted organisation that offers credential management services by issuing, certifying and removing digital certificates and the corresponding public keys linked to the long-term identity of their owners.
- **eIDAS Node.** An implementation of the eID eIDAS Profile provided by the European Commission for electronic identification, authentication and trust services. One node per country is needed, to support services capable of identifying citizens and businesses from other Member States.
- **Pseudonym Provider (PP).** A trusted organisation that is responsible for the pseudonym management of the short-term anonymous credentials [1609.2-2016].

2.3. Data Exchanged

The following are the main kinds of data whose exchange is covered by the Protocol:

- **Research Definition Documents:** structured documents formally describing research studies, including enrolment and exit criteria, data queries, a human-readable description of the study, and other study-related metadata;
- **Pseudonymized health data for research:** citizen health data queried from the phone, pseudonymized/anonymized, and sent to a research centre;
- **Digitally signed consent:** a formal agreement between a citizen and a research centre about the participation of a citizen to a research study, or his/her withdrawal from it;
- **Enrolment and exit notifications:** messages indicating the successful enrolment of a citizen into a study, or his/her leaving of the study.

For the representation of health data, as well as queries and criteria, the Protocol adopts the FHIR standard [FHIR], as does the entire InteropEHRate project. This design choice allows the retrieval of health data from citizens' S-EHRs directly, without requiring further data conversion mechanisms. Beyond FHIR itself, the

Protocol requires the data contained in S-EHRs to conform to InteropEHRate's highest functional level of interoperability, called "Research Sharing", as specified by the deliverable [D3.1], in order to ensure that cross-border data collection leads to meaningful results.

2.4. Processes

The execution of a research study, from its initial proposal by a researcher until its closure and archival, is a long and complex process that can last years, even for retrospective studies where medical data are readily available. Typically, the entire process involves the following macro-steps:

1. Pre-acceptance (GO / NO-GO)
2. *Formulation of requests to execute a given research study (as a formal research description)*
3. Approvals from the Ethical Committee as well as w.r.t. feasibility
4. Setting up of research environment
5. *Setting up the cohort, including citizen consent*
6. *Retrieval of data*
7. Preparation and linkage of datasets
8. Data analysis for the research experiment
9. Control of access to results
10. Archival of experiment and results
11. Closure

Addressing all of the macro-steps above is out of the scope of the InteropEHRate project and of the Protocol itself. The Protocol's focus, instead, is the way in which medical data are retrieved directly from citizens' smartphones, with all the necessary handling of consent, privacy, and security aspects of the operation. For this reason, the Protocol only covers the macro-steps relevant to these operations (in italics above), namely:

- **Formulation of request:** only to the extent that the research study is defined in the form of a formal, machine-processable RDD document. The Protocol does not cover *how* the RDD is created, but it does rely on the RDD in the operations it defines. The precise format of the RDD is defined in a separate deliverable on the InteropEHRate Interoperability Profiles [D2.9].
- **Setting up the cohort:** this covers the verification of enrolment criteria, as well as gathering citizen consent. Citizens are provided with the possibility of subscribing and being enrolled into specific research studies, as well as withdrawing from them.
- **Retrieval of data:** the citizens' data are transferred from their smartphones to their respective RRCs.

Accordingly, the Protocol consists of the following macro-steps or phases:

1. **OPT-IN:** the Citizen opts in to be invited in research studies in general.
2. **ENROLLMENT:** the consenting Citizen is enrolled into a specific study.
3. **DATA RETRIEVAL:** relevant health data are retrieved from the Citizen's phone.
4. **WITHDRAWAL:** the Citizen decides to withdraw from providing further data to a given study.
5. **OPT-OUT:** the Citizen decides to opt out from a given study or from all current and future studies.

3. ARCHITECTURE AND INTERFACES

The figure below shows the main software systems, their exposed remote APIs, and their corresponding human users (actors) whose actions and communication are covered by the Protocol.

The protocol defines and exploits, as shown in Figure 2, the following APIs:

- **Research Interface (RDSI):** remote API offered by the Research Centre Information System, allowing any S-EHR App to communicate the consent for a specific research study, receive enrolment-related information, and sharing citizen health data .
- **Research Definition Document Download Interface (RDDI):** remote API offered by the Central Node, allowing the S-EHR App to download Research Definition Documents.

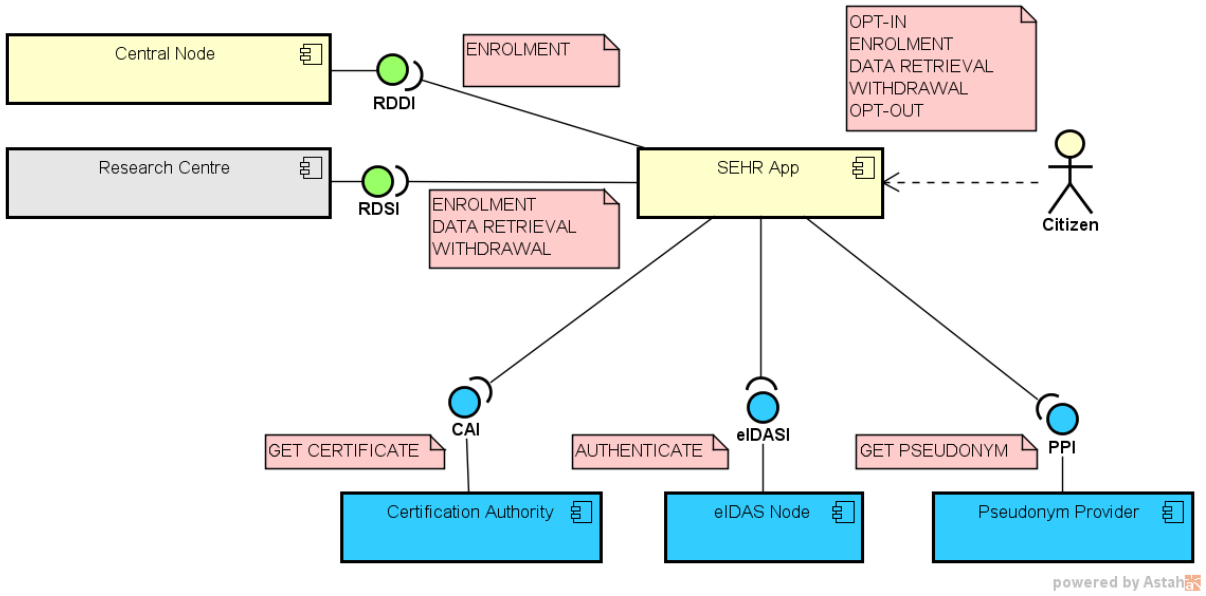

Figure 2 - Systems, actors, and communication channels of the Protocol

The meaning of colours in the figure is the following:

- blue for standard legacy interfaces and systems (used but not defined by the Protocol);
- yellow for new systems defined in this deliverable;
- green for new interfaces defined in this deliverable.

3.1. Human-Computer Interfaces and Use Cases

This section describes the user interfaces that are required by the Protocol, from a high-level functional perspective of use cases. The Protocol covers the interactions of the Citizen with the Research Network. The Protocol does not specify how this human interaction happens, in particular, it does not require the usage of specific local APIs for executing them but specifies only how the input and output of these human interactions are related to input and output of specified remote APIs. Other user interactions with the mentioned systems are possible, but they are not covered by the specification of the Protocol because they do not constraint the usage of the Protocol APIs.

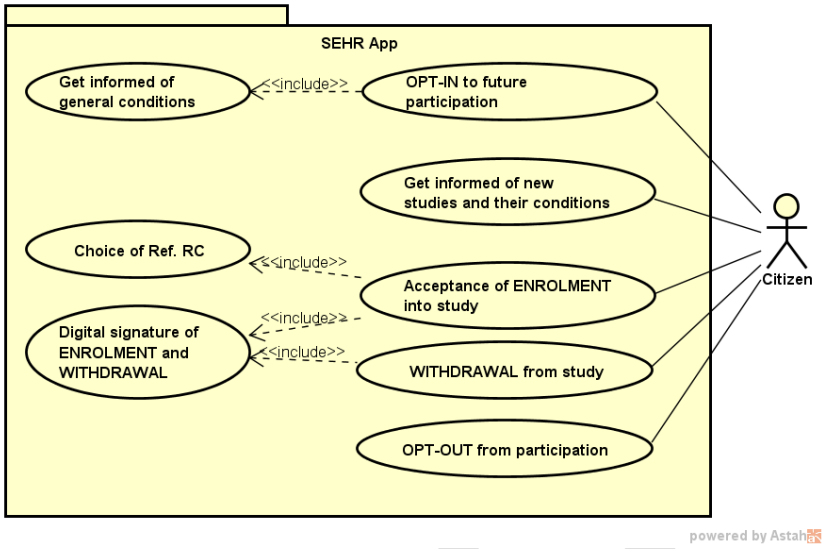

Figure 3 - Use case diagram for the interaction of the Citizen with the S-EHR App

- **OPT-IN to future participation:** the Citizen sets his/her status on the smartphone as "interested" in participating in future studies. Before doing so, the Citizen is informed of what this entails (namely, the silent verification of enrolment criteria on his/her phone by accessing his/her health data, without sharing any citizen data with third parties). This allows the phone regularly to retrieve information about studies.
- **Get informed of new studies and their conditions:** the Citizen is informed about every study for which his/her health data meet the eligibility criteria, including the purpose and details of the study, the data collected, etc.
- **Acceptance of ENROLMENT into study:** the Citizen formally accepts to participate in a given study.
- **Choice of Reference Research Centre:** as part of accepting the enrolment into a study, the Citizen chooses a reference research centre (RRC) from a list of possibilities corresponding to his/her geographical region of stay.
- **Digital signature of enrolment and of withdrawal:** for both enrolling into a study and withdrawing from it, a formal contract needs to be signed between the Citizen and his/her Reference Research Centre. These contracts are digitally signed by the Citizen, requiring his/her explicit participation.
- **WITHDRAWAL from study:** the Citizen formally signals the decision to stop sending data for a given study.
- **OPT-OUT from participation:** the Citizen sets his/her status on the smartphone as "not interested" anymore in participating in future studies.

3.2. Remote APIs Defined by the Protocol

This section provides a succinct description of the endpoints of the major interfaces defined by the Protocol: RDDI and RDSI, including the high-level security-related interfaces. Further technical details (such as the RESTful syntax or the API errors returned) will be provided in the next version of the deliverable.

3.2.1. RDDI - Central Node

The Central Node provides the services exposed through the *Research Dataset Definition Interface* (RDDI). RDDI is a RESTful interface.

RDDI	Caller	Input Parameters	Return value	Description
GET open-studies	S-EHR App	void	SignedRDD List	An open endpoint that allows any caller, but primarily a S-EHR App, to retrieve the digitally signed list of currently open studies, to which enrollment is possible. Returns a list of RDDs, each describing a study.

Table 1 - Methods of the RDDI Interface

In the table below more details are provided for the *Research Dataset Definition Interface* (RDDI) APIs.

Property	Value
FHIR Resource	Bundle
HTTP Method	GET
Header Params	• Content-Type: `application/fhir+json`
URL	`http://<BASE_ADDR>/open-studies`
Client Search Params	no input parameters
Return Value	• Instance of Bundle of type `collection` containing instances of ResearchStudyBundle of type `document` represented with `Content-Type` corresponding to requested Accept parameter (JSON). • Version: DSTU3 • Profile: http://hl7.org/fhir/uv/ips/StructureDefinition/composition-uv-ips

HTTP Return Codes	**200 Successful**: request was successfully processed. **400 Bad Request**: search could not be processed or failed basic FHIR validation rules. **401 Not Authorized**: authorization is required for the interaction that was attempted. **403 Forbidden**: client is not allowed to access requested resources due to security policy. **404 Not Found**: resource type not supported, or not a valid FHIR end-point. **406 Not Acceptable**: client requested a not supported content-type format. **500 Internal Server Error**: server encountered an unexpected internal error, the request could not be processed.

Example:

```
GET http://<BASE_ADDR>/open-studies
```

3.2.2. RDSI - Research Centre Information System

The Research Centre Information System provides the services exposed through the *Research Data Sharing Interface* (RDSI). RDSI is a RESTful interface. The detailed API descriptions for the RDSI will be included in the second version of this deliverable, D4.9.

RDSI Endpoint	Caller	Input Parameters	Return value	Description
sendEnrollmentConsent	S-EHR App	studyID, citizenIdentification, citizenPseudo, signedConsent, sehrAppId	SignedContract	Send the Citizen's electronically signed consent of enrolling into a specific study. The consent also includes personal identification information on the citizen, the newly generated study-specific pseudonym or pseudo-identity, as well as the S-EHR App ID. The receiving RC checks the signature validity of the signedConsent, signs and returns the contract signed by both parties
sendExitNotification	S-EHR App	studyID, citizenPseudo, reason, citizenSignature	void	Send a notification that the Citizen is exiting a study due to the exit criteria being met. If the RRC fails to satisfy the call, a corresponding RESTful API Error is returned.
sendWithdrawal	S-EHR App	studyID, citizenPseudo, citizenSignature	void	Send a notification that the Citizen is withdrawing from an ongoing research study. If the RRC fails to satisfy the call, a corresponding RESTful API Error is returned.
sendHealthData	S-EHR App	studyID, citizenPseudo, healthData	void	Allows a S-EHR App to send citizen health data to the RRC. The receiving RC verifies and decrypts the encrypted and signed payload *healthData* and retrieves the FHIR bundle contained within. If the RRC fails to satisfy the call, a corresponding RESTful API Error is returned.
retrievePseudoIdentity	S-EHR App	studyID	PseudoIdentity	Allows a S-EHR App to receive a pseudo identity which has been generated at the RRC.

Table 2 - Methods of the RDSI Interface

3.3. Standard APIs Used by the Protocol

This section enumerates existing remote services and their APIs that are used by the Protocol but are not defined by it. These services are used for pseudonymisation and security purposes.

3.3.1. CAI - Certificate Authority

The Certificate Authority provides the services exposed through the *Certificate Authority Interface* (CAI) [EJBCA 2021]. CAI is a web service interface. More information regarding how the interface is involved in the protocol will be provided in the context of D3.10.

CAI Endpoint	Caller	Input Parameters	Return value	Description
getCertificate	S-EHR App, RRC	void	X509Certificate	An open endpoint that allows any caller to retrieve a valid X509 Certificate associated with a public key. If the caller fails to satisfy the call, a corresponding CertificateException Error is returned.

Table 3 - Methods of the CAI Interface

3.3.2. eIDASI - eIDAS Node

The eIDAS Node provides the services exposed through the *eIDAS Interface* (eIDASI). eIDASI offered public interfaces are provided in [eIDAS 2017]. More information will be provided in the context of D3.10.

eIDASI Endpoint	Caller	Input Parameters	Return value	Description
generateEIDASAuthnRequest	S-EHR App	String destination, String serviceProvider, int qaal, PersonalAttributeList personalAttributeList, String assertionConsumerServiceURL	byte[]	A SAML assertion token as well as the necessary identification attributes retrieved by an eIDAS Node in byte[] from.

Table 4 - Methods of the eIDASI Interface

3.3.3. PPI - Pseudonym Provider

The Pseudonym Provider provides the services exposed through the *Pseudonym Provider Interface* (PPI) [1609.2-2016]. PPI is a RESTful interface. More information will be provided in the context of deliverable [D3.10].

PPI Endpoint	Caller	Input Parameters	Return value	Description
getPseudonym	S-EHR App	String[] token	Pseudonym	The S-EHR App anonymously requests a pseudonym by providing

				an anonymous SAML assertion token acquired by eIDAS authentication.

Table 5 - Methods of the PPI Interface

3.4. Technical Binding of API Datatypes

The technical binding of the API datatypes is based on the data model as described in [D2.8], where several FHIR Implementation Guides are described. The Implementation Guide for Research Data Sharing defines the data model for the management of research related studies.

Several constraints apply to the technical binding that go beyond the definition of a data model:

- An overarching and unique study identifier related to a specific research study must be provided and used with the data exchange related to that study.
- Healthcare data related to a study must not contain any identifying data such as a citizen's name, instead the provided healthcare data should be anonymized or pseudonymized.
- Identifiers used to uniquely identify a citizen, either on a device, or in a data sharing environment, must be replaced with pseudo-identifiers (or pseudonyms) related to a citizen in the context of a specific study. This way, the identity of the citizen is protected.
- The provision of research data is the result of the application of the enrolment criteria, specifying a citizen cohort with certain features, and the data selection criteria, defining exactly which information about a citizen are used to compile the research data. Thus, only the data defined by the data selection criteria of citizens matching the cohort specification are provided. The selected research data must be anonymized and all identifiers must be replaced with pseudo-identifiers (or pseudonyms) as described above in order to protect the citizen's identity.
- The enrolment criteria and the data selection criteria are provided as part of a research study.
- A trusted organization, such as the research center, may be in possession of the citizen's unique identifiers and the citizen's pseudo-identifiers and is able to re-identify a citizen by managing their correspondence.

The following table shows the API parameter binding.

No	API parameter	technical binding (see D2.8)		description
		implementation guide	HL7 FHIR resources / profiles (Cardinality)	
1	SignedRDDList	Implementation Guide for Research Data Sharing	- Research Study (1..N) - Cohort (1..N) - Data Set Definition (1..1) - Reference	A digital signed list of currently open studies. It consists of a FHIR bundle containing a combination of the resources shown in this table and profiled according to the Implementation Guide for Research Data Sharing [D2.8].

			Research Centres (1..N)	
2	studyID	Implementation Guide for Research Data Sharing	Research Study . identifier (1..1)	A ResearchStudy object which contains the attribute identifier.
3	citizenIdentification	Implementation guide for Cross Border Data Exchange	Patient	Identifying attributes of a citizen should not be transmitted together with the corresponding pseudonyms used for a specific study. Therefore, the Patient profile is used.
4	citizenPseudo	Implementation Guide for Research Data Sharing	- Research Subject . pseudoID (1..1) - Research Subject . patient . identifier	The API parameter citizenPseudo contains either a pseudonym or a pseudo-identity of the citizen, and is transmitted based on the ResearchSubject profile. If the connection between a pseudo and a patient must be transmitted, the different identifiers can be transmitted within the same resource.
5	signedConsent	Implementation Guide for Research Data Sharing	Research Subject . Citizen Consent (1..1)	The API parameter signedConsent is based on the Consent profile referenced by the Research Subject profile.
6	sehrAppId	Implementation Guide for Research Data Sharing	- Device . identifier (1..1) - Device . patient . id (1..1)	The device identifier attribute is used in order to specify the potentially identifying S-EHR App Id. A non-anonymous patient identifier is also provided in order to allow the assignment to an identified citizen.
7	Reason	Implementation Guide for Research Data Sharing	profile/extension on Research Subject . status	
8	HealthData	Implementation guide for Cross	all	A digitally signed and encrypted bundle of FHIR

		Border Data Exchange		resources according to the specified implementation guide. All identifying information has been removed or replaced in the context of the research study. Instead, the research subject's pseudo id and anonymized demographic information is contained.

Table 6 - Technical bindings of the API parameters

4. ANONYMIZATION AND PSEUDONYMIZATION

This section describes data pseudonymisation and anonymization of health data shared by citizens for research purposes. Both mechanisms are used on the citizen's phone, so that the data sent to researchers does not include any unique identifier that could lead to the identification of an entity. In the context of InteropEHRate, depending on the mechanism by which the pseudo-identifier of a citizen is generated, one of two variants of pseudonymisation will be used: *pseudo-identity-based* or *pseudonym-based*, both of which are presented in detail in section 4.1 below.

The difference between anonymization and pseudonymisation is, while in neither case should the identity of a participating citizen be revealed, pseudonymisation still allows the identification of the citizen. This possibility is foreseen only within the context of the Reference Research Center and in exceptional cases, such as upon the discovery of a severe illness of a citizen. The RRC can map the pseudonym to a citizen's unique identifiers, or can request this mapping to be done by the Pseudonym Provider.

The procedure works as follows: first of all, the query is executed on the citizen's phone, and the dataset defined in the RDD is extracted. Then it is anonymized or pseudonymized, depending on the scenario, before being sent to the RRC.

The researcher is responsible to choose whether to use data pseudonymisation or data anonymization. This is also set inside the Research Definition Document (RDD), which includes policies for both cases. In case of pseudonymisation all personal information, which is found in the requested data, will be replaced with a pseudonym, whereas in case of anonymization all personal information will be removed so that it can be impossible for someone to lead to the identification of a citizen. Afterwards, the pseudonymized or anonymized data will be sent to the Reference Research Center and by extension to the researcher who conducts the research.

The execution of the operation related to pseudonymisation and anonymization will be further analysed in the sections below.

4.1. Study-Specific Pseudonymisation

The Principal Investigator (PI) of the study publishes the Research Definition Document (RDD). The RDD contains the policies and the dataset needed and mentions whether pseudonymisation or anonymization should be implemented on this dataset.

In case of pseudonymisation, there are two ways to replace all personal information of the citizens, depending on the pseudonymisation policy adopted for the research study. Once the citizen is authenticated at the eIDAS node, and if he or she gives consent to participating in the study, either a (more conventional) *pseudo-identity* or a (stronger but more rarely used) *pseudonym* [Camenisch 2017] will be created for the citizen. The pseudo-id will be generated by the Principal Investigator at each Research Centre participating in the study, whereas the pseudonym will be generated by the Pseudonym Provider (PP). Depending on study-specific policies defined within the RDD and initially set by the PI of the study, the mechanism of either using pseudo-identities or pseudonyms is chosen.

The steps, as described above, are the following: first, the data query is executed on the mobile device, extracting the dataset defined in the RDD. In the case the data needs to be pseudonymized, a pseudonym is created and stored locally on the phone, based on the study's predefined policies, and replaces the citizen's identifiers as described above. Finally, the pseudonymized dataset is sent to the Reference Research Center.

4.1.1. Variant #1: Data Pseudonymisation through Study-Specific Pseudo-Identities

The pseudo-identity is essentially a string generated after a pattern that consists of numbers and/or letters. It replaces all attributes, and in particular all direct and indirect identifiers, which are not included in the requested data and can lead to the identification of a citizen, for example the name and age. The pseudo-id consists of three parts:

[PREFIX] [INCREMENTED_NUMBER] [SUFFIX]

The prefix is an alphanumeric sequence that depends on the study, and is stated in the given RDD. The incremented number is increased every time the PI creates a pseudo-id for a citizen for a specific study. The suffix is a random sequence, placed right after the aforementioned values.

A different pseudo-id is produced for each study and for each citizen. The pseudo-id is not only sent to the RRC along with the requested data, but it is also stored locally on the citizen's mobile device (S-EHR application). This method has its pros and cons: the advantage of using this method is that the pseudo-ids are human-readable, which corresponds to current well-established practices of hospitals and researchers. The disadvantage is that the pseudo-ids have limited randomness and they are more vulnerable to unauthorized de-identification, as opposed to the pseudonyms.

4.1.2. Variant #2: Data Pseudonymisation through Short-term Pseudonyms

Pseudonyms are certificates that are only valid if they are signed by a root CA and only for a short time [Eckhoff2011]. As mentioned before, the citizen is authenticated at the eIDAS node and retrieves a SAML response that includes an anonymous token. This token is used for the citizen's authentication to the Pseudonym Provider, in order to receive a high-entropy pseudonym. Entropy provides the measure of the uncertainty to identify the citizen that is participating in a study among a set of citizens.

After the pseudonyms have been received, the citizen can send his/her signed consent (signed with his/her private key) to the RRC and, finally, can start sharing (anonymously signed) data. The latter also contains a blind signature so that the RRC can always verify that received data originates from a user that has already provided a signed consent. The PI of the RRC will be able to get access to the mapping of pseudonyms with the IDs of the citizens that acquired them in order to be also able to perform quick pseudonym resolution. All other actors of the RRC (as well as other research centres) will not be able to link data back to the users/citizens, unless as an emergency, in which case they will be allowed to request the necessary pseudonym resolution to a *trusted third party*, defined in the next version of this deliverable (see section 8). The actors that can request pseudonym resolution (beyond the PI of an RRC that directly has access to the mapping of pseudonyms to the IDs of citizens) will have to prove both their identity and the reason why this request needs to take place (e.g. medical emergency).

4.2. Anonymization Operations

The researcher may choose anonymization instead of or alongside pseudonymisation for their research. This is stated within the RDD. In the case of pseudonymisation, the researcher could re-identify a citizen based on his/her pseudonym in specific predefined cases. In the case of anonymization, this is not feasible,

since all unique identifiers of a citizen are removed locally on the phone before the requested data are sent to the RRC.

Once the query is executed, the data requested by the RDD are extracted and anonymized within the S-EHR app before being sent to the researcher. Not only are removed all direct identifiers such as name or surname, but also all irrelevant information not explicitly requested by the researcher and that the citizen has consented to share.

The data anonymization operation can be split into two major categories: anonymization of structured and unstructured data. For structured data, the process is simpler as---apart from the case where the data structure embeds unstructured information, such as paragraphs of free text---the identification of the data values to be anonymized amounts to enumerating the corresponding structured data attributes (in our case, the FHIR resource attributes). When it comes to unstructured data, the process is more challenging, as there is no standard and robust automated way to identify the pieces of data to be removed or replaced.

Anonymization of unstructured data is even more challenging when it needs to be executed on a mobile device, due to technological limitations in the application of resource-intensive information extraction methods. For example, in the case of an electrocardiogram (ECG), information regarding the citizen may not only be included in the metadata of the image, but inside the image pixel data itself. Hence, it might be very difficult or even impossible to conceive generic methods that can be run on a mobile device in order to delete such information. A possible solution is to create two instances of the data (e.g. of the image or unstructured text document) at the data source (e.g. the hospital): the first one may include information about the citizen, while the second one will be an anonymized version which will not include any information at all and it will be used for research purposes. In this way, unstructured data can be supported in the S-EHR app.

The precise anonymization policies adopted by the Research Data Sharing Protocol will be defined in the second version of this deliverable. Our current understanding is that the identification and flagging of data values inside structured data will need to be done as an upstream operation, before the health record is loaded onto the mobile device. The S-EHR App will then only need to perform the substitution of flagged data values. As for unstructured data, either the same approach can be used, or---especially in the case of imaging or other binary data---a fully anonymized copy of the data should be uploaded into the S-EHR App alongside the original one.

5. PROTOCOL SECURITY

In order to be able to have the secure exchange of messages a) between the S-EHR App and the Central Node in the context of RDDs and b) between the S-EHR App and the reference Research Center in the context of shared data, we need the user requirements listed in Table 7. Based on these requirements, the main security aspects we need to consider are the encryption of the communication channel between the S-EHR App and the reference Research Center in order to achieve confidentiality of the shared medical data, authenticity and integrity of the medical information and RDDs, and mutual authentication between the citizen and the reference Research Center but with the requirement of privacy-preserving authentication from the RRC side.

Establishing a Public Key Infrastructure (PKI) is one of the most important tasks in security concerning communication over the internet. To be able to do that an existing third party trusted Certificate Authority is required to be in place. The role of the Certificate Authority (CA) is to a) issue certificates, b) confirm the identity of a certificate owner and c) to provide proof that the certificate is valid. For multiple mechanisms and security concepts to work, the above requirements must be fulfilled. EJBCA[2] covers all our needs - from certificate management, registration and enrolment to certificate validation. EJBCA is platform-independent and is easily scalable to match the needs of InteropEHRate PKI requirements. In addition, in asymmetric schemes, the pseudonym issuance process is similar to certificate issuance in a PKI. In the literature it is typically proposed that CAs manage and issue long-term identity certificates while pseudonyms are issued by separate Pseudonym Providers (PP) [PETIT2015].

To guarantee data confidentiality, an authenticated key agreement protocol will be used to securely exchange a symmetric session and AES256 for the actual encryption. The Station-to-Station (STS) [STS1992] protocol is a cryptographic key agreement scheme based on classic Diffie–Hellman, and provides mutual key and entity authentication. Unlike the classic Diffie–Hellman, which is not secure against a man-in-the-middle attack, this protocol assumes that the parties have signature keys, which are used to sign messages, thereby providing security against man-in-the-middle attacks. In the context of the research scenario, the STS scheme will be used securely to establish a symmetric encryption key. The authentication form the reference Research Center side is basically encapsulated in the consensus to enrolment as defined in Section 3, where the reference Research Center needs to authenticate the citizen in a privacy-preserving manner through the appropriate use of pseudonyms.

In order to apply all the security needs in the protocol, a necessary bootstrap phase exists in order for all the participants to issue their certificates from a Certificate Authority, and pseudonym certificates from a Pseudonym Provider. This phase is out of the scope of the protocol, however it is mandatory for the successful completion of the security steps. Section 3.2 provides all the interfaces of the Research Data Sharing Protocol, including the CA APIs as well as the necessary security parameters, such as *encryptedBundle*. Last but not least, apart from the application-level security, *Transport Layer Security v1.2* must be enabled at the transport layer in both communication channels, the RDSI and RDDI to protect from data breaches and Distributed Denial of Service (DDoS) attacks. In a nutshell, TLS is a protocol which provides privacy between communicating applications and their users, or between communicating services. When a server and client communicate, well-configured TLS ensures that no third party can eavesdrop or tamper with any message.

[2] https://www.ejbca.org

#id	User requirement title	Security Requirements
#69	Non-repudiable data provenance tracking	Non repudiation
#70	Integrity of medical information	Integrity
#86	Digital signature by Reference Research Centre of Citizen's consent	Authenticity
#87	Citizen's digital signature of consent to share health data for a given study	Authenticity
#88	Citizen's digital revocation of consent to share health data for a given study	N/A
#130	Pseudoidentity restricted to single research protocol	Privacy
#150	Identification and authorisation of organisations and researchers accessing to IRS	Authentication & Authorisation

Table 7 - Security Requirements

5.1. Security Prerequisites

The correct execution of the Protocol supposes that the following prerequisites are respected regarding credentials:

- the Central Node, the S-EHR App, and all Reference Research Centres have retrieved their certificates from a central Certification Authority, as described in the context of other protocols in [D3.6];
- the S-EHR Apps have retrieved the pseudonym certificate from the Pseudonym Provider or the pseudo-identity from the RRC. All the necessary external APIs will be provided in the context of [D3.6] and [D6.8];
- upon installation, the S-EHR App has downloaded the public key of the Central Node along with its connection address (URI).

5.2. Security of the Research Data Sharing Channel

Below we present in detail the operations performed by Research Centres as well as the S-EHR App in order to establish the security of the RDS communication channel. The table below maps the security operations to the Protocol steps in which they are executed, as described in section 6.

Purpose: Confidentiality of the communication channel between the S-EHR App and the reference Research Center, Integrity and Authenticity of the shared medical data, Mutual Authentication between the Citizen and the reference Research Center.

Actors and components: Citizen, S-EHR App, reference Research Centre, Pseudonym Provider, Principle Investigator of the RRC, Certificate Authority.

Preconditions: All actors have installed the necessary credentials and certificates signed and verified by a CA. For the pseudonym-based variant of pseudonymisation, the citizen must also be already authenticated to the trusted Pseudonym Provider using a valid eIDAS SAML assertion and retrieve the necessary assertion (e.g. anonymous certificate). TLSv1.2 or greater needs to be established.

Steps:

Step	Security Operation	Protocol Step (see Section 6)
1	The S-EHR App uses a Key Derivation Function (KDF) to derive a one-time key K that will be used for encrypted communication between the S-EHR App and the reference Research Centre. Then encrypts asymmetrically the generated key K with the public key of the reference Research Center.	ENROLMENT Step 4
2	The S-EHR App re-sings the assertion that verifies that the Citizen is authenticated and acquired signed from the eIDAS Node, using the pseudo-identity/pseudonym that acts as a short-term signing key.	ENROLMENT Step 5
3	The S-EHR App encrypts the double-signed assertion asymmetrically with the public key of the reference Research Center.	ENROLMENT Step 5
4	The S-EHR App sends to the reference Research Center the concatenated encrypted double-signed assertion and the encrypted generated symmetric key.	ENROLMENT Step 7
5	The reference Research Center receives the message, verifies the S-EHR App's signature, decrypts the encrypted key with its private key then verifies the double-signed assertion.	ENROLMENT Step 7
6	The reference Research Center adds its own digital signature to the signed consent document using its private key, and encrypts it with the S-EHR App's public key, before sending the counter-signed consent document back to the Citizen.	ENROLMENT Step 8
7	Upon reception of the counter-signed consent document, the Citizen decrypts it using his/her private key, and checks that it is signed by the reference Research Center using the Center's public key.	ENROLMENT Step 8
8	The S-EHR App encrypts symmetrically and sings the anonymised data to be shared with the pseudo-identity or pseudonym.	DATA RETRIEVAL Step 5

9	The reference Research Centre receives the encrypted and anonymised health data and validates the signature. Then decrypts data using the established key.	DATA RETRIEVAL Step 6
10	The Citizen withdraws his/her participation in a study, digitally signing a withdrawal message (contract) before sending it to the reference Research Center.	WITHDRAWAL Step 1

Table 8 - Mapping of RDS security operations to the Protocol steps in section 6

These security operations correspond to three major security requirements:

- **mutual authentication** between the S-EHR App and the Reference Research Centre before starting sharing the data;
- **confidentiality** of the shared data through encryption, so that unauthorized individuals cannot access or use them;
- **non-repudiation** of the data sharing contract or its withdrawal, through digital signatures.

5.3. Security of the Research Data Definition Channel

Below we present in detail the operations performed by the Central Node as well as the S-EHR App in order to establish the security of the RDD communication channel. The table below maps the security operations to the Protocol steps in which they are executed, as described in section 6.

Purpose: Integrity and Authenticity of RDDs, Non-repudiation of the Central Node signatory.

Actors and components: Citizen, S-EHR App, Central Node, Principal Investigator of the Study, Certificate Authority.

Preconditions: All actors should have installed the necessary credentials and certificates signed and verified by a CA. The Central Node should also acquire its own certificate signed by the trusted Certificate Authority. TLSv1.2 or greater should be used. The PI has already authenticated to the Central Node and published a digitally signed RDD.

Steps:

Step	Security Operation	Protocol Step (see Section 6)
1	The Central Node digitally signs the RDDs, to ensure that it is not accidentally altered or changed with her/his private key.	Before ENROLMENT Step 1
2	The S-EHR App downloads the published RDDs from the Central Node along with their digital signatures and certificate.	ENROLMENT Step 1
3	The S-EHR App checks the validity of each certificate and signature.	ENROLMENT Step 2

	If the validation is successful, the S-EHR App can be sure that the study was not altered (integrity), while the Central Node cannot deny having published the RDDs (authenticity, non-repudiation).	

Table 9 - Mapping of RDD security operations to the Protocol steps in section 6

The rationale behind these security aspects of RDDI is to download the RDDs and to be sure that nothing has been altered during the download. Integrity and non-repudiation are necessary aspects of health data, documents, and reports. For this reason, InteropEHRate should comply with the Electronic Signatures Directive 1999/93/EC and EU Regulation 910/2014 of 23 July 2014 on electronic identification (eIDAS). Digital signature ensures authenticity and integrity of the RDDs, while at the same time prevents the Central Node signatory from being able to repudiate (deny) his involvement. Such properties make the adoption of digital signatures an integral part of RDDI.

6. PROCESS DEFINITIONS

This section describes the communication process among the human actors and systems involved, as defined by the Protocol the sequence diagrams corresponding to each phase of the protocol. The Protocol is divided into phases as defined in section 2. For each phase, both a high-level activity diagram and a lower-level and more formal sequence diagram are provided. The diagrams focus on the interactions between components and thus contain little detail on operations internal to single components. Green colour marks the swimlanes and actions of human agents (Citizen), while light yellow represents actions by automated systems.

6.1. OPT-IN phase

Purpose: In the OPT-IN phase, the Citizen signals the general intention of participating in research studies in the future. (S)he gives his/her consent to the S-EHR App on the phone regularly to poll the Research Network for new studies soliciting enrolment.

Actors and components: Citizen, S-EHR App.

Preconditions: The S-EHR App has been installed on the Citizen's mobile device. The solicitation of the Citizen could possibly happen in the last stages of the installation process.

Steps:

1. The S-EHR App solicits the Citizen for a general agreement to future participation in research studies.
2. The Citizen chooses either *yes* (opt-in) or *no* (opt-out). The S-EHR App may also allow the Citizen to postpone the decision.
3. The Citizen's answer is recorded in the S-EHR App. In case the Citizen decided to opt out, the S-EHR App does not send any other research-related solicitations to the Citizen in the future.

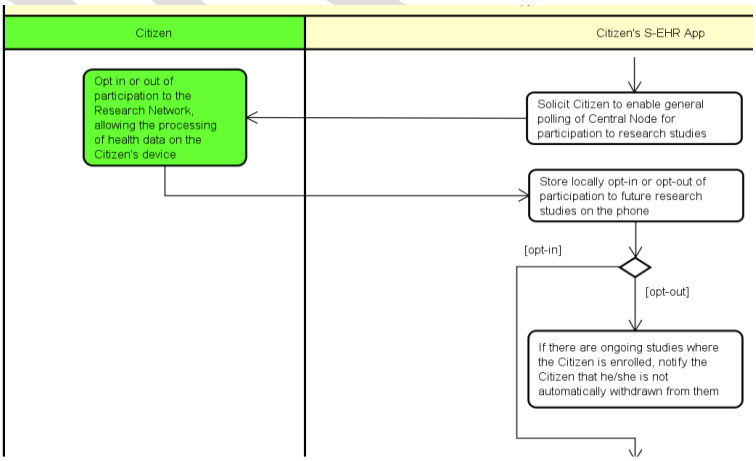

Figure 4 - High-level data flow of the OPT-IN phase

6.2. ENROLLMENT phase

Purpose: In the enrolment phase, for each newly published study, the S-EHR App evaluates whether the Citizen's health data matches the enrolment criteria, and if so, asks for the Citizen's consent to be enrolled in the study. Upon enrolment, a Reference Research Centre is selected by the Citizen for the study.

Actors and components: Citizen, S-EHR App, Reference RC, Central Node.

Preconditions: The Citizen has previously opted in for participating in studies. A study has been published on the Central Node. The Citizen has a S-EHR installed on his/her smartphone that contains health data in a formal, structured, and standardised representation, such as the one defined by the InteropEHRate Interoperability Profile [D2.8].)

Steps:

1. If the Citizen has chosen to opt in to studies in the OPT-IN phase, the S-EHR App regularly (e.g. daily) polls the Central Node to retrieve the RDDs of currently open studies. The set of such RDDs is downloaded by the S-EHR App from the Central Node as a digitally signed document.
2. The S-EHR App checks the digital signature on the set of RDDs retrieved.
3. For each new study RDD retrieved, the S-EHR App silently extracts and verifies the enrolment Criteria with respect to the patient data of the Citizen. In case the patient data does not meet the enrolment Criteria, the RDD is silently deleted and no further action is taken with respect to it.
4. If the patient data meets the enrolment Criteria, the Citizen is solicited by the S-EHR App for his/her (digitally signed) consent to participate in the study, displaying to him/her the goals, purposes, and conditions of research and the data collected. In case the Citizen declines participation, the RDD is deleted and no further action is taken with respect to it.
5. In case the Citizen agrees and digitally signs his/her consent, (s)he is prompted to choose first a Reference Region and then a corresponding Reference Research Centre (RRC), from the list of regions and research centres contained in the RDD.
6. An anonymous study-specific identifier is generated to be used for data pseudonymisation in the data retrieval phase. Depending on the study definition within the RDD, either a pseudo-identity or a pseudonym is generated (see section 4):
    a. in case the study operates with pseudo-identities:
        i. the S-EHR App retrieves a pseudo-identity from the RRC chosen by the Citizen;
        ii. the consent to enrolment, digital signature, and S-EHR App identifier are then sent to the RRC chosen by the Citizen. The S-EHR App identifier is needed so that the RRC can contact the Citizen if necessary, such as if the Citizen's health data point to an important and so far undiagnosed pathology.
    b. in case the study operates with short-term pseudonyms:
        i. the S-EHR App retrieves a study-specific pseudonym, created by the Pseudonym Provider that represents the Citizen;
        ii. the consent to enrolment, digital signature, S-EHR App identifier, and pseudonym are then sent to the RRC chosen by the Citizen. The S-EHR App identifier is needed so that the RRC can contact the Citizen if necessary, such as if the Citizen's health data point to an important and so far undiagnosed pathology.

7. The RRC receives this enrolment notification and stores it in its list of citizens enrolled into the study.
8. The RRC counter-signs the consent using digital signature, and sends back the counter-signed document to the S-EHR App, which saves it locally.

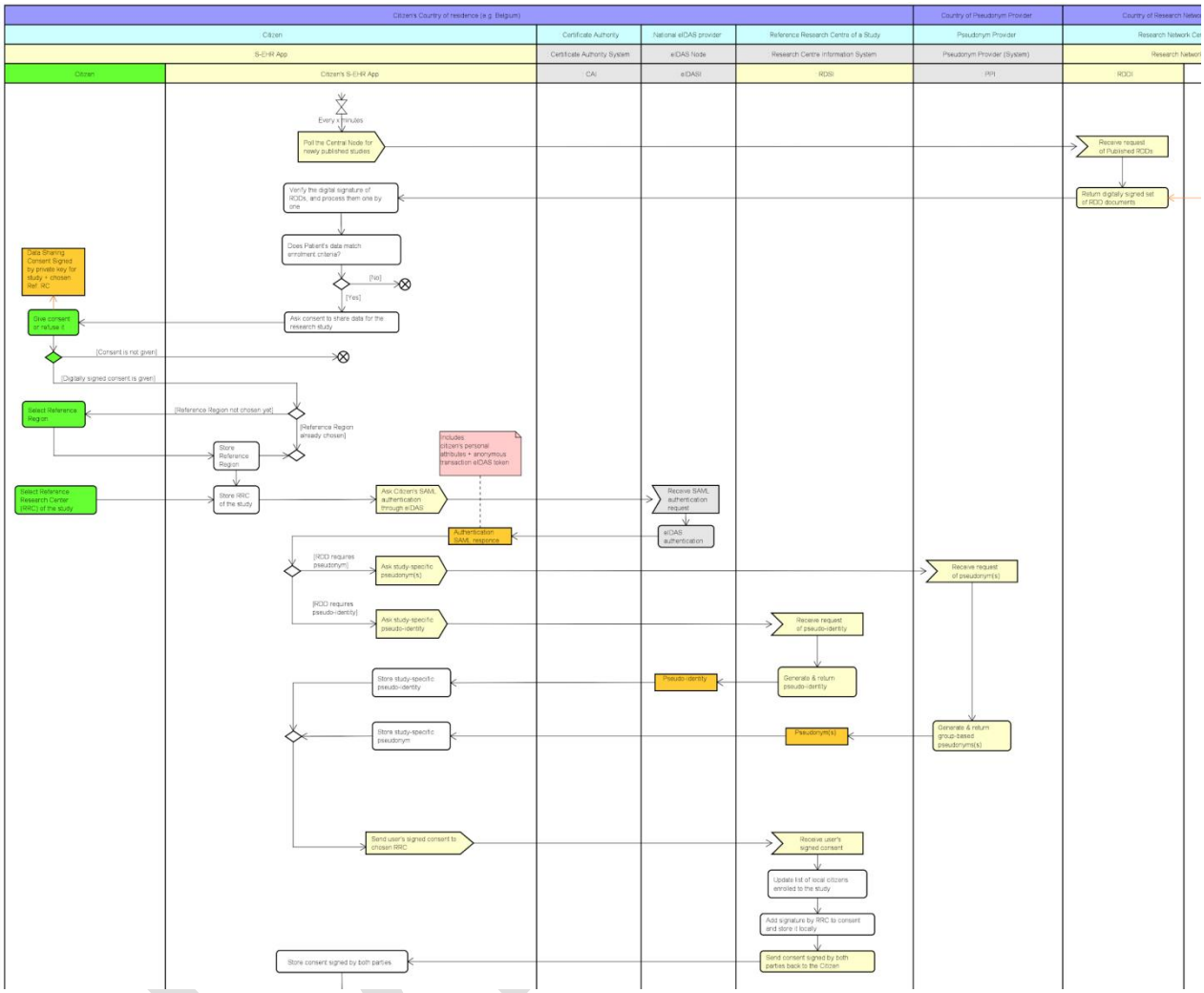

Figure 5 - High-level data flow of the ENROLLMENT phase

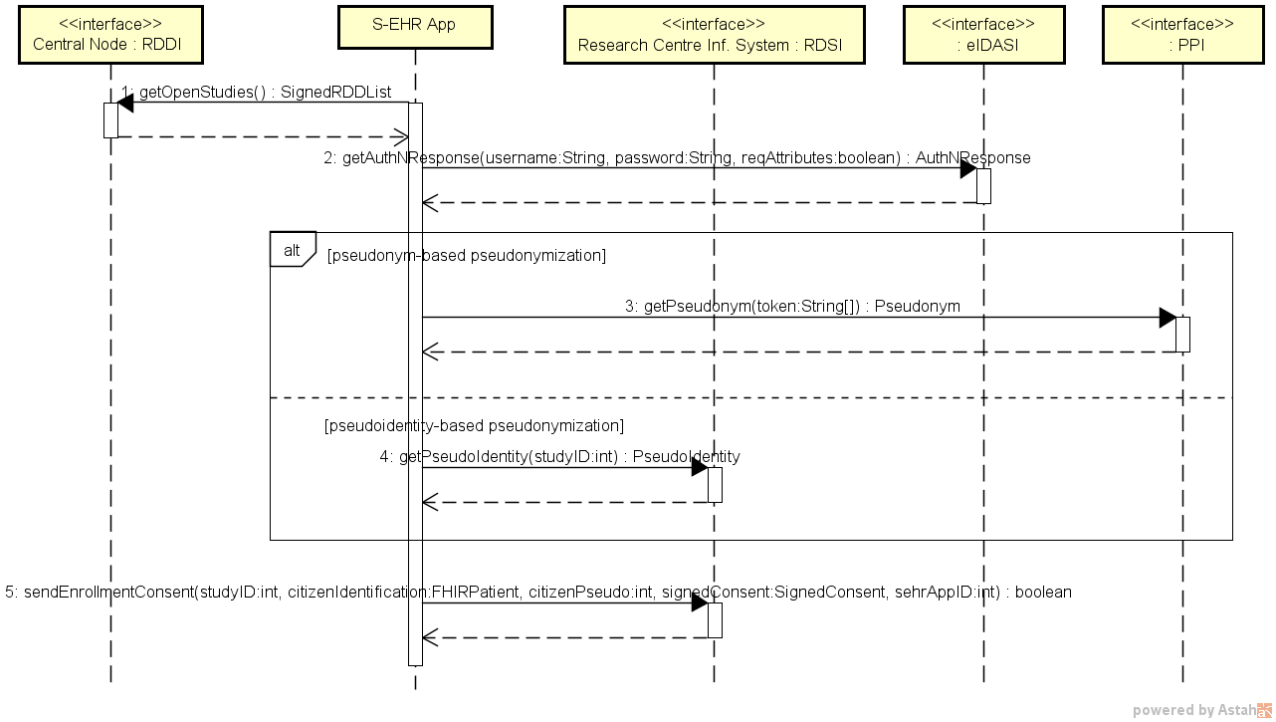

Figure 6 - Sequence diagram for the ENROLLMENT phase

6.3. DATA RETRIEVAL phase

Purpose: In the DATA RETRIEVAL phase, data relevant to the study is gathered from the Citizen's smartphone, either a single time (for retrospective studies) or repetitively (for prospective studies). After in-phone anonymization, the data are sent to the Citizen's RRC.

Actors and components: S-EHR App, Reference RC.

Preconditions: The Citizen has been enrolled in the study.

Steps:

1. If the study is retrospective AND all necessary data are available, the following steps 2-6 will be executed only once. Otherwise, they will be executed periodically, as defined in the RDD. Within each data retrieval period, the S-EHR App regularly (e.g. daily) checks for new, updated data to be available, and executes steps 2-X as soon as this is the case.
2. The S-EHR App verifies again if the Citizen still meets the enrolment criteria, and does not meet the exit criteria. If the result is negative, an Exit Notification is sent to the RRC containing the reason (enrolment criteria negative or exit criteria positive). Upon the reception of this message, the RRC updates the list of citizens enrolled into the study.
3. If the data still meet the criteria, the S-EHR App retrieves the data to be sent to the RRC, building and executing the query based on the dataset definition included in the RDD.
4. The S-EHR App applies anonymization and pseudonymisation to the data retrieved, based on the requirements included in the RDD.

5. The S-EHR App sends the anonymized/pseudonymized data to the RRC. Depending on policy/settings, the S-EHR App may notify the Citizen that data has been sent.
6. The RRC receives the data.
7. At the end of the data retrieval period, upon the decision of the PI of the RRC, the RRC forwards the data collected from all citizens under its control to the CRC where the PI of the Study resides.

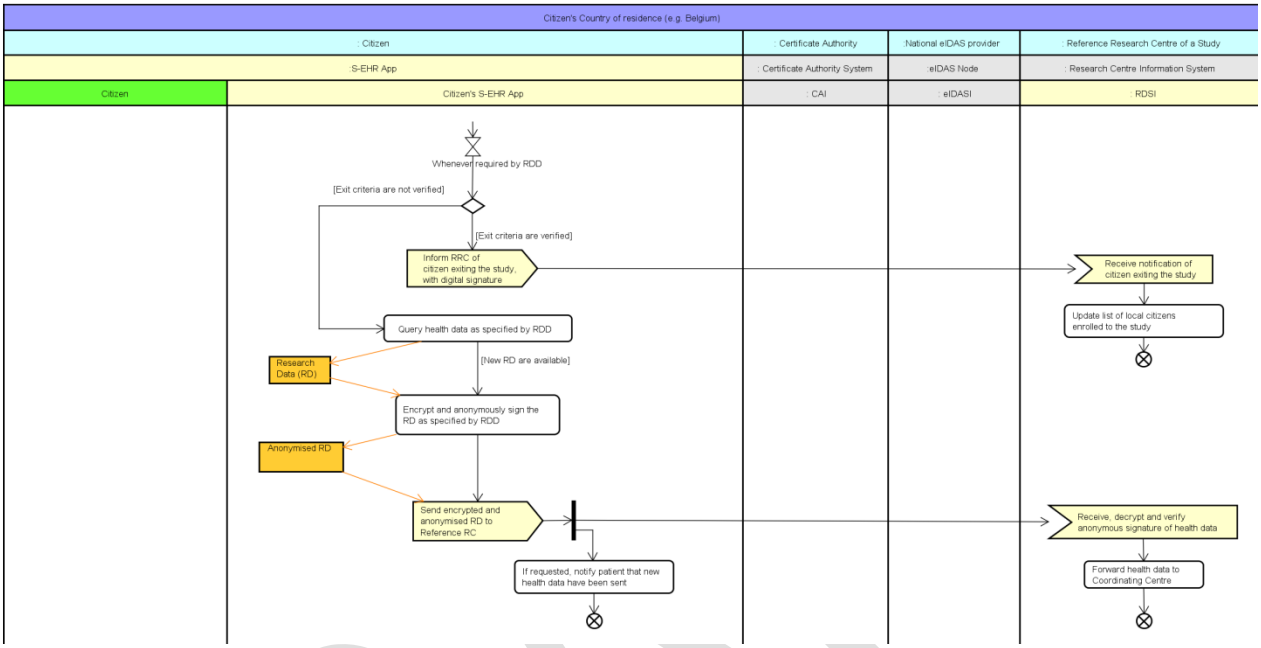

Figure 7 - High-level data flow of the DATA RETRIEVAL phase

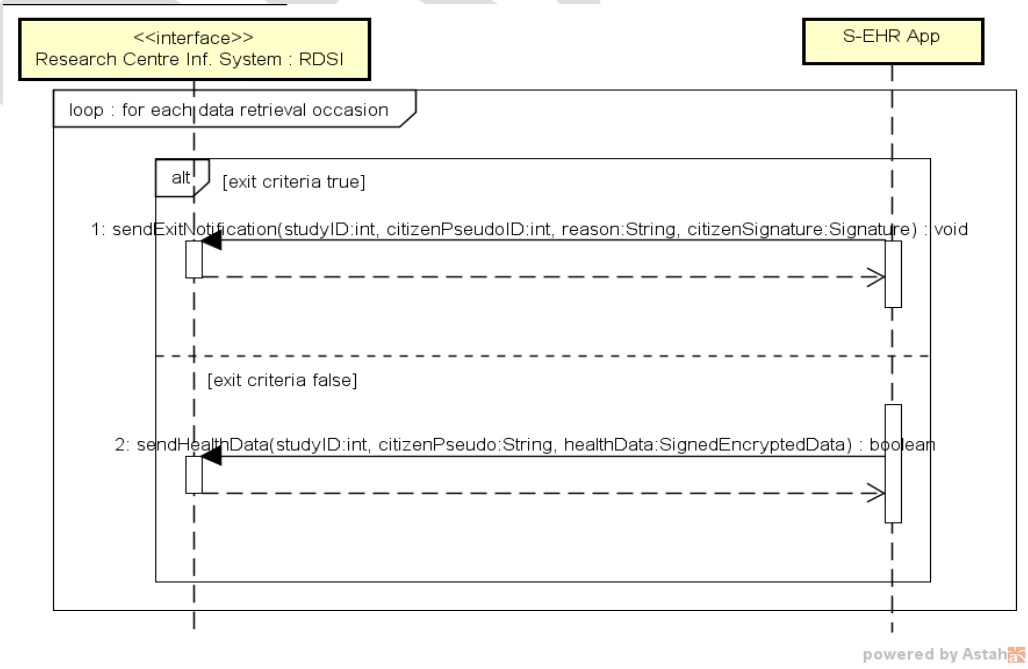

Figure 8 - Sequence diagram for the DATA RETRIEVAL phase

6.4. WITHDRAWAL phase

Purpose: In the WITHDRAWAL phase, the Citizen decides to end his/her ongoing participation to a given study. Upon this request, all further data retrieval operations are suspended, previously collected data are deleted, and the Citizen is deleted from the list of enrolled patients at the RRC.

Actors and components: Citizen, S-EHR App, Reference RC.

Preconditions: The Citizen has been enrolled in the study. At the time of the withdrawal, the study is either still in the enrolment phase or it is already running. It is not possible to withdraw from a study once the data retrieval period has ended.

Steps:

1. The Citizen decides to withdraw from a specific study. In order to do so, (s)he retrieves from the S-EHR App the list of studies to which (s)he is enrolled, and selects "withdraw". (S)he digitally signs the withdrawal.
2. The S-EHR App sends the signed withdrawal message to the RRC.
3. The RRC receives the notification. It acknowledges it, deletes all data retrieved from the Citizen so far, and updates its local list of citizens enrolled into the study.

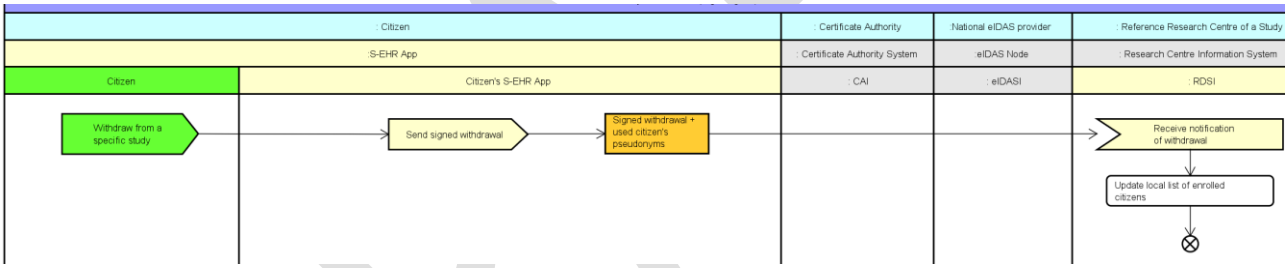

Figure 9 - High-level data flow of the WITHDRAWAL phase

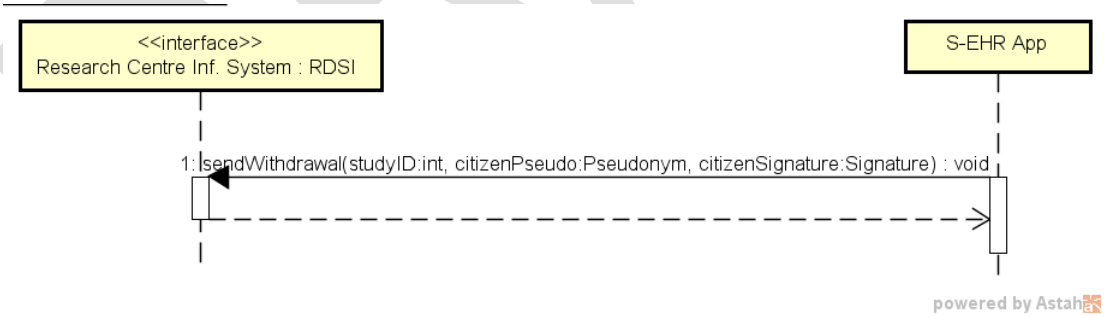

Figure 10 - Sequence diagram for the WITHDRAWAL phase

6.5. OPT-OUT phase

Purpose: In the OPT-OUT phase, the Citizen decides to opt out from any future study.

Actors and components: Citizen, S-EHR App.

Preconditions: The Citizen has previously opted in to research studies.

Steps:

1. The Citizen decides to opt out from all future studies. (S)he lets this know to the S-EHR App.
2. The S-EHR App may consider this decision only for future studies, or also for ongoing studies, in which case a separate withdrawal is necessary for each study. In case there are ongoing studies where the Citizen is enrolled, the S-EHR App prompts the Citizen whether (s)he really wants to withdraw from these ongoing studies as well.
3. The decision to opt out from future studies is stored locally in the S-EHR App. From this moment on, the S-EHR App will not poll the Central Node for future studies anymore.
4. In case there are ongoing studies where the Citizen is enrolled, and the Citizen has explicitly signalled to want to withdraw from these as well, the S-EHR App executes a withdrawal process separately for each study.

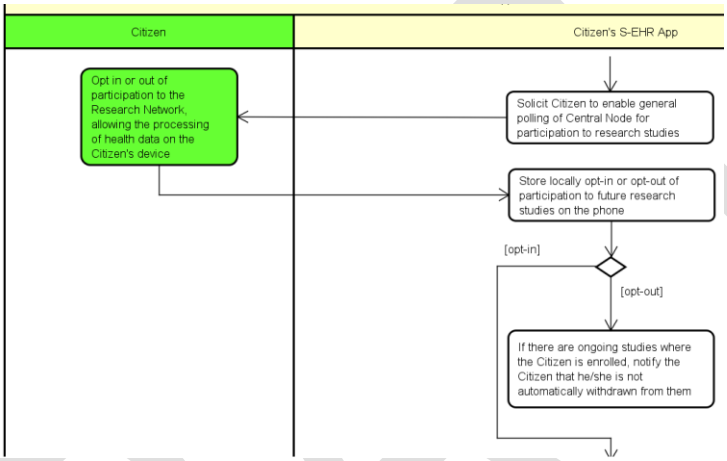

Figure 11 - High-level data flow of the OPT-OUT phase

7. RELATED WORK

The Protocol described in this deliverable is novel and unique in its approach of retrieving health data directly from citizens' personal devices. The conventional approach so far has been, for retrospective studies, to transform and reuse data from existing sources under centralised control (of a hospital, a region, or an entire country), such as clinical patient data or death records, based on general prior consent given by patients. This is the process assumed and implemented, for instance, in the project *Healthcare Data Safe Havens*, as presented in [HDSH]. In prospective studies, typically a much smaller number (e.g. hundreds) of patients are involved, with consent and data collection happening physically at and carried out by the research centre (or multiple research centres in the case of multicentric studies). The Research Data Sharing Protocol covers both the retrospective and the prospective use cases, but entirely decentralises the data collection process.

In terms of cross-border interoperability, the Protocol adopts the state-of-the-art approach of relying on international data representation standards. Existing projects such as [EMIF], [EHDEN], or Healthcare Data Safe Havens [HDSH] rely on the OMOP CDM standard [OMOP], which is a structured data representation specifically developed for research data interoperability. Our Protocol, on the other hand, relies on the FHIR standard for representing data for research. This choice is not justified by FHIR being inherently better for this purpose—we consider both FHIR and OMOP adequate for the majority of use cases—but by the seamless interoperability it provides with FHIR-based electronic health records, such as the smartphone-based *smart health data* [D2.8]. This way, data can be directly retrieved from an already cross-border interoperable representation, without requiring any complex data transformation that would be tedious to implement on smartphones. However, the rest of the Protocol is designed so that it does not rely on any specific underlying data format: it is up to the implementation to ensure that the query format used in the RDD matches the health record data format on the smartphone, or else to implement data conversion mechanisms.

8. CONCLUSIONS AND NEXT STEPS

This deliverable specifies a first version of the Research Data Sharing Protocol. This first version should be considered as consolidated with respect to the goals and the scope of the Protocol (in terms of operations covered), its actors, and the high-level system architecture it presupposes. The upcoming second and last version will provide additional details with respect to the data models, the communication processes, security, and anonymization aspects. These additional details and improvements will be based on experience gathered through a first demonstrator implementation of systems and libraries that realize the Protocol. In particular, we see the following aspects of the Protocol as not yet final:

- APIs of the RDSI and RDDI interfaces: while an effort was made to align the API definitions in section 3 with the requirements as much as possible, the future evolution of requirements or implementation experience may lead to their modification. Minor technical details, still to be specified, and concrete binding of the RESTful APIs, here reported only at abstract level, will be documented in the next version of this deliverable.

- The RDD data model: the detailed data structure of the RDD are defined as part of separate deliverables, namely [D2.8] (released in M24) and [D2.9] (to be released in M36) that describe the interoperability profiles of the InteropEHRate project. The next version of this deliverable will be aligned with these future results, in terms of the technical bindings and API definitions in section 3.

- Pseudo-identity generation: the generation of pseudo-identities (described in section 4) according to patterns predefined by medical researchers has been foreseen as a requirement and described in its generality; however, the precise representation of such patterns in the RDD and how the subsequent pseudo-identity generation happens are yet to be detailed.

- Pseudo-identity mapping management: the mappings between S-EHR App identifiers and pseudo-identities, necessary to contact patients in exceptional cases of emergency, is ideally maintained by a trusted third party (other than the research centre itself). The involvement of this party is yet to be defined.

- Anonymization: while pseudonymisation and anonymization have been considered in their generality, it has not yet been specified what kind of in-device anonymization techniques, if any, will be applied, and which kinds of data (structured, unstructured, images, image metadata) will need it.

- Security: while the principal security mechanisms have already been described in sections 5 and 6, some details have been left for future development, such as the precise certificates to be provided by the Certificate Authority. Also, the activity and sequence diagrams are not yet showing all communication details.

- Questionnaire for prospective studies: for prospective studies, information will also be gathered through questionnaires presented to citizens through their mobile devices. This form of data collection is a new requirement and has not yet been specified as part of the Protocol.

These still underspecified elements of the Protocol will be addressed in the future v2 of this deliverable.

Similar to the other communication protocols, the RDS protocol is intended to be supported by different systems, potentially provided by different vendors. Different implementations will be interoperable if compliant to this specification.

For the readers interested to experiment with the RDS protocol without implementing it themselves, a reference implementation is provided by deliverable **[D4.17]** including libraries for the mobile devices, the research centres, and the Central Node. The design of this specific implementation is documented by the deliverable [D4.10].

REFERENCES

- **[D2.2]** InteropEHRate Consortium, *Deliverable D2.2—Requirements Specification V2*, 2020. www.interopehrate.eu/resources/#dels
- **[D2.5]** InteropEHRate Consortium, *Deliverable D2.5—InteropEHRate Architecture - V2, 2020.* www.interopehrate.eu/resources/#dels
- **[D2.7]** InteropEHRate Consortium, *Deliverable D2.7—FHIR profile for EHR interoperability v1*, 2019. www.interopehrate.eu/resources/#dels
- **[D2.8]** InteropEHRate Consortium, *Deliverable D2.8—FHIR profile for EHR interoperability v2*, 2020. www.interopehrate.eu/resources/#dels
- **[D2.9]** InteropEHRate Consortium, *Deliverable D2.9—FHIR profile for EHR interoperability v3*, 2021. www.interopehrate.eu/resources/#dels
- **[D3.1]** InteropEHRate Consortium, *Deliverable D3.1—Specification of S-EHR mobile privacy and security conformance levels - V1, 2020. www.interopehrate.eu/resources/#dels*
- **[D3.4]** InteropEHRate Consortium, *Deliverable D3.4—Specification of remote and D2D IDM mechanisms for HRs Interoperability - V2, 2021. www.interopehrate.eu/resources/#dels*
- **[D3.5]** InteropEHRate Consortium, *Deliverable D3.5—Specification of data encryption mechanisms for mobile and web applications - V1, 2020.* www.interopehrate.eu/resources/#dels
- **[D3.6]** InteropEHRate Consortium, *Deliverable D3.6—Specification of data encryption mechanisms for mobile and web applications - V2, 2021. www.interopehrate.eu/resources/#dels*
- **[D3.10]** InteropEHRate Consortium, *Deliverable D3.10—Design of libraries for HR security and privacy services - V2, 2021.* www.interopehrate.eu/resources/#dels
- **[D6.8]** InteropEHRate Consortium, *Deliverable D6.8—Design of a mobile service for data anonymization and aggregation, 2021. www.interopehrate.eu/resources/#dels*
- **[D4.10]** InteropEHRate Consortium, *Deliverable D4.10—Design of library for health data sharing for research v1*, 2021. www.interopehrate.eu/resources/#dels
- **[D4.17]** InteropEHRate Consortium, *Deliverable D4.17- Libraries for research health data sharing - V1, 2021.* www.interopehrate.eu/resources/#dels
- **[FHIR]** HL7 FHIR Specifications. https://www.hl7.org/fhir/
- **[Camenisch 2017]** J. Camenisch and A. Lehmann, "Privacy-Preserving User-Auditable Pseudonym Systems," 2017 IEEE European Symposium on Security and Privacy (EuroS&P), Paris, 2017, pp. 269-284, doi: 10.1109/EuroSP.2017.36.

- **[eIDAS 2017]** European Commission — DIGIT Unit D3, eIDAS-Node Installation, Configuration and Integration Manual, Version 1.3, 2017
- **[EJBCA 2021]** PrimeKey, EJBCA WS Support, 2021 https://download.primekey.se/docs/EJBCA-Enterprise/latest/ws/index.html
- **[EMIF]** The *European Medical Information Framework* project, http://www.emif.eu
- **[EHDEN]** The *European Health Data and Evidence Network* project, http://www.ehden.eu
- **[HDSH]** G. Bella et al., *Cross-Border Medical Research using Multi-Layered and Distributed Knowledge*. Prestigious Applications of Intelligent Systems, ECAI 2020.
- **[OMOP]** OHDSI, *The OMOP Common Data Model.* https://www.ohdsi.org/data-standardization/the-common-data-model/
- **[PETIT2015]** Petit, F. Schaub, M. Feiri and F. Kargl, "Pseudonym Schemes in Vehicular Networks: A Survey," in IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pp. 228-255, First quarter 2015, doi: 10.1109/COMST.2014.2345420.
- **[STS1992]** W. Diffie, P. van Oorschot and M. Wiener, "Authentication and Authenticated Key Exchange'', Designs, Codes and Cryptography, 2, 1992, pp.107-125.
- **[1609.2-2016]** "IEEE Standard for Wireless Access in Vehicular Environments--Security Services for Applications and Management Messages," in IEEE Std 1609.2-2016 (Revision of IEEE Std 1609.2-2013), vol., no., pp.1-240, 1 March 2016, doi: 10.1109 / IEEESTD 2016 7426684.
- **[Eckhoff2011]** D. Eckhoff, R. German, C. Sommer, F. Dressler and T. Gansen, "SlotSwap: strong and affordable location privacy in intelligent transportation systems," in IEEE Communications Magazine, vol. 49, no. 11, pp. 126-133, November 2011, doi: 10.1109 / MCOM.2011.6069719.