PAPAYA

PlAtform for PrivAcY preserving data Analytics

D5.1 – EHEALTH USE CASE VALIDATION

Work Package	WP 5, Platform Validation
Lead Author	Eleonora Ciceri (MCI, Marco Mosconi (MCI)
Contributing Author(s)	Francesco Serra (MCI), Alex Calovi (MCI), Matteo Eccher (MCI), Simone Fischer-Hübner (KAU), Farzaneh Karegar (KAU), Angel Palomares Perez (ATOS), Nuria Ituarte Aranda (ATOS), Ron Shmelkin (IBM), Boris Rozenberg (IBM)
Reviewers	Nuria Ituarte Aranda (ATOS), Bastien Vialla (ORA)
Due date	31.07.2021
Date	26.07.2021
Version	1.0
Dissemination Level	PU (Public)

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, through the PAPAYA project, under Grant Agreement No. 786767. The content and results of this deliverable reflect the view of the consortium only. The Research Executive Agency is not responsible for any use that may be made of the information it contains.

Revision History

Revision	Date	Editor	Notes
0.1	24.02.2021	Eleonora Ciceri (MCI)	TOC definition
0.2	01.03.2021	Eleonora Ciceri (MCI), Marco Mosconi (MCI)	Initial contribution
0.3	20.06.2021	Eleonora Ciceri (MCI), Marco Mosconi (MCI), Francesco Serra (MCI), Matteo Eccher (MCI), Alex Calovi (MCI), Angel Palomares Perez (ATOS), Nuria Ituarte Aranda (ATOS), Ron Shmelkin (IBM), Boris Rozenberg (IBM)	Finalization of contribution regarding the architectural description and the requirements validation
0.4	30.06.2021	Simone Fischer-Hübner (KAU), Farzaneh Karegar (KAU)	Validation via end users introduced
0.5	08.07.2021	Eleonora Ciceri (MCI), Bastien Vialla (ORA), Nuria Ituarte Aranda (ATOS)	Addressed comments from the first review round
0.6	26.07.2021	Eleonora Ciceri (MCI), Bastien Vialla (ORA), Nuria Ituarte Aranda (ATOS)	Addressed comments from the second review
1.0	26.07.2021	Melek Önen (EURC), Eleonora Ciceri (MCI)	Quality check and final version

Table of Contents

Executive Summary 5
Glossary of Terms 6
1 Introduction 7
1.1 Summary of contributions 7
2 UC1: Privacy-Preserving Arrhythmia Detection 8
2.1 Use case description in a nutshell 8
2.2 Use cases specification validation 10
2.2.1 Revision of GDPR roles 10

2.2.2 Coverage of use cases 11
2.2.3 Coverage of privacy requirements 11
2.2.4 Integration with PAPAYA platform....................12
2.2.5 Applications implementation: interfaces 14
2.3 Requirements validation....................21
2.4 Validation by stakeholders 22
3 UC2: Privacy-preserving stress management 22
3.1 Use case description in a nutshell....................23
3.1.1 Physiological parameters suggesting stress 23
3.1.2 Building a NN model that detects stress....................24
3.1.3 Building a dataset to train the NN model 25
3.1.4 The need for large datasets: how can we collect enough data to build a better performing model? 26
3.2 Use cases specification alidation 26
3.2.1 Revision of GDPR roles 26
3.2.2 Coverage of use cases 27
3.2.3 Coverage of privacy requirements 29
3.2.4 Integration with PAPAYA platform....................30
3.2.5 Applications implementation: interfaces 33
3.3 Requirements validation....................44
3.4 Validation by stakeholders 45
3.4.1 Evaluation Objectives 45
3.4.2 Evaluation phases 46
3.4.3 Results....................48
3.4.4 Conclusions 58
4 Conclusions 60

List of Tables

Table 1 Coverage of use cases related to “Privacy-preserving arrhythmia detection” 11
Table 2 Coverage of privacy requirements for “Privacy-preserving arrhythmia detection” 11
Table 3 Integrated PAPAYA component for “Privacy-preserving arrhythmia detection” 13
Table 4 Requirements coverage for Use Case 1 21

Table 5 Coverage of use cases related to “Privacy-preserving stress management” 27
Table 6 Description of UC-STR-1a 27
Table 7 Description of UC-STR-1b 28
Table 8 Coverage of privacy requirements for “Privacy-preserving stress management” 29
Table 9 Integrated PAPAYA component for “Privacy-preserving arrhythmia detection” 30
Table 10 Requirements coverage for Use Case 2....44
Table 11 Features in functionality list potentially could be covered by the metaphors. 53

List of Figures

Figure 1 CardioPharma service, prior to its integration with PAPAYA 8
Figure 2 CardioPharma service, integrated with PAPAYA 9
Figure 3 Integration of CardioPharma and the PAPAYA platform....14
Figure 4 CardioPharma pharmacist application: initial dashboard 15
Figure 5 CardioPharma pharmacist application: patient profile 16
Figure 6 CardioPharma pharmacist application: access to DS Tool 1 16
Figure 7 CardioPharma pharmacist app: content of DS Tool 1 17
Figure 8 CardioPharma pharmacist application: starting a new analysis 18
Figure 9 CardioPharma pharmacist application: exam in progress....19
Figure 10 CardioPharma cardiologist application: overview of an exam....20
Figure 11 CardioPharma cardiologist application: arrhythmia classification 20
Figure 12 CardioPharma cardiologist application: report creation....21
Figure 13 A heartbeat showing its R peak [10] 24
Figure 14 The stress-detection pipeline....25
Figure 15 Integration of CardioPharma, Stress app and the PAPAYA platform....32
Figure 16 Collaborative training: CardioPharma installations for each participating company 33
Figure 17 CardioPharma: creation of tagging campaign 34
Figure 18 CardioPharma: visualization of tagging campaigns....35
Figure 19 CardioPharma: labeling beats with arrhythmia classes....36
Figure 20 CardioPharma: labeling ECG tracks as stressed or not stressed 36
Figure 21 CardioPharma: collaborative training dashboard 37
Figure 22 CardioPharma: possible statuses for collaborative training task....38
Figure 23 PAPAYA uStress application: login and pairing to the monitoring device 39
Figure 24 PAPAYA uStress dashboard 40
Figure 25 PAPAYA uStress application: notification of stress statuses 41
Figure 26 PAPAYA uStress application: access to DS tools....42
Figure 27 Welcome page and list of preferences from the Privacy Engine....43
Figure 28 Choice of privacy preferences by workers using the Privacy Engine....44
Figure 29 High-level data analysis scenario in UC2 of PAPAYA (shown to our interviewees-see 46
Figure 30 Our method to reach our objective 47
Figure 31 The study design of the interviews conducted in phase 3....48
Figure 32 Metaphor to convey about differential privacy for centralized models....51
Figure 33 Metaphor to convey about differential privacy for centralized models in the context of federated learning....52

Executive Summary

This deliverable reports the outcome of the work carried out in Task T5.1 (*“Validation through eHealth UC”*). Specifically, the document reports the validation process of the eHealth use cases, namely, *Privacy-preserving arrhythmia detection* (UC1) and *Privacy-preserving stress management* (UC2), in terms of:

- adherence to the initial specifications of the use cases, as defined in Deliverable D2.1;
- adherence to the requirements defined in Deliverable D2.2;
- validation through the identified stakeholders.

This document, when complemented with the outcomes of T5.2 and T5.3, serves as a tool for validating the outcome of the PAPAYA projects, in terms of adherence to the initial design and applicability to real-world scenarios.

Glossary of Terms

2PC	Two-Party Computation
CNN	Convolutional Neural Network
DP	Differential Privacy
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DS	Data Subject
ECG	ElectroCardioGram
GDPR	General Data Protection Regulation
HRV	Heart Rate Variability
IT	Information Technology
MCI	MediaClinics Italia
NN	Neural Network
PET	Privacy Enhancing Technology
PP	Privacy-preserving
UC	Use case

1 Introduction

This document reports the results of the work carried out in Task T5.1 ("*Validation through eHealth Use cases"*). The aim of this work is to perform validation of the PAPAYA platform via the eHealth use cases proposed in Deliverable D2.1. This validation is done by pursuing the following objectives:

1. **Use case coverage**. Deliverable D2.1 proposed a set protocols and privacy requirements that each use case should follow. Validation should ensure that the proposed use cases and privacy requirements have been covered during the implementation of the PAPAYA platform and its integration with the eHealth demonstrators;
2. **Integration with PAPAYA.** This deliverable shall prove that the integration approach that has been adopted to integrate PAPAYA technologies and the demonstrators has been carried out smoothly, as a demonstration of the fact that the PAPAYA platform is flexible and could be integrated by any IT developer in any field (healthcare included) in an easy way;
3. **Requirements validation.** This deliverable shall provide a comprehensive view of how requirements proposed in Deliverable D2.2 had been covered;
4. **Validation by stakeholders.** This deliverable shall prove that the proposed solution (which integrates healthcare solutions and PAPAYA solutions) is considered valuable by the stakeholders of the healthcare scenario. Such stakeholders have been identified and described in previous deliverables (see, e.g., the Annex to Deliverable D6.4), and shall give a point of view on the proposed solution both on usability and applicability to the market.

The work presented in this deliverable, if complemented to the one presented in Deliverable D5.2 (where the use cases related to the mobile and phone usage scenario are validated) and the one presented in Deliverable D5.3 (where recommendations and refinements for the PAPAYA framework are provided), completes the validation of the work performed in the PAPAYA project. Indeed, the whole validation process (whose outcome is presented in these three deliverable) succeeds in: i) demonstrating the usability and applicability of the PAPAYA technologies in real-world scenarios (i.e., eHealth and mobile scenario); ii) demonstrating the validity of the PAPAYA framework from a business and technical perspective.

1.1 Summary of contributions

This document is structured as follows:

- Section 2 presents the validation of UC1 (i.e., Privacy-preserving arrhythmia detection);
- Section 3 presents the validation of UC2 (i.e., Privacy-preserving stress management);
- Section 4 concludes the document summarizing the major findings.

2 UC1: Privacy-Preserving Arrhythmia Detection

In this chapter, we present the validation activities carried out for the first use case in the eHealth scenario, namely, the *Privacy-preserving arrhythmia detection* use case.

2.1 Use case description in a nutshell

The *privacy-preserving arrhythmia detection* use case targets patients who need to perform cardiac parameters analysis for some clinical reason, e.g., in case patients suffer from a chronic condition, with the goal of verifying the presence/absence of arrhythmias.

MediaClinics Italia (MCI) already offered the ECG analysis to patients well before the beginning of the PAPAYA project, through a service called CardioPharma. Figure 1 presents the architectural layout of the offered service.

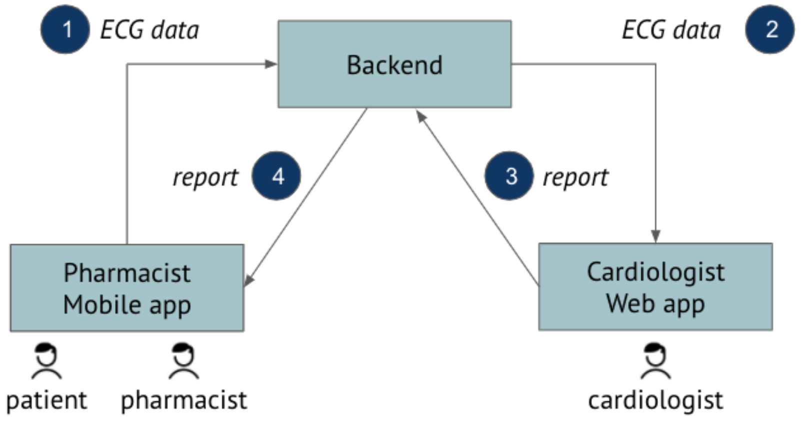


Figure 1 CardioPharma service, prior to its integration with PAPAYA

As Figure 1 shows, to get access to the service, a patient could go to a pharmacy affiliated with MCI and request an analysis. Such an analysis would require the pharmacist to deliver the patient with an ECG sensor (that would read ECG data for 24 or even 48 hours), collect the ECG data when ready (via the CardioPharma mobile application, see step 1) and send it to the cardiologist (see step 2). As a final step, the cardiologist would analyze manually the long stream of ECG data (via the CardioPharma Web application) to understand if there are arrhythmias, write a report on the findings and make it available for the patient once finalized (see steps 3 and 4).

This solution came with its own limitations, specifically at the analysis side. Indeed, analyzing a long stream of ECG data is a huge work for a cardiologist alone, and thus two options become viable: i) either analyze manually a subset of ECG data (avoiding to go through the whole 24/48 hours), or ii) make a machine go through a pre-processing phase during which it spots arrhythmia automatically, which are then signalled to the cardiologist to be analyzed and reported. Both these approaches, unfortunately, come with their own problems:

- a limited manual analysis of ECG data could make the cardiologist overlook some important signals of arrhythmia;
- an automatic analysis of ECG data would require competences in AI (that a SME could not have), and moreover a full, in-house analysis of long ECG tracks could require computational power that may be difficult to have in-house.

Thus, the definition of the Privacy-preserving arrhythmia detection use case presented in D2.1 proposed a scenario where:

- the 24/48-long ECG track is analyzed fully on cloud, where large computational resources are available, and the complexities of ECG analysis have been tackled by sector experts;
- the PAPAYA platform protects the data during the whole analysis process, avoiding the exposure of sensitive data to external actors.

Figure 2 presents the new architecture of the proposed solution:

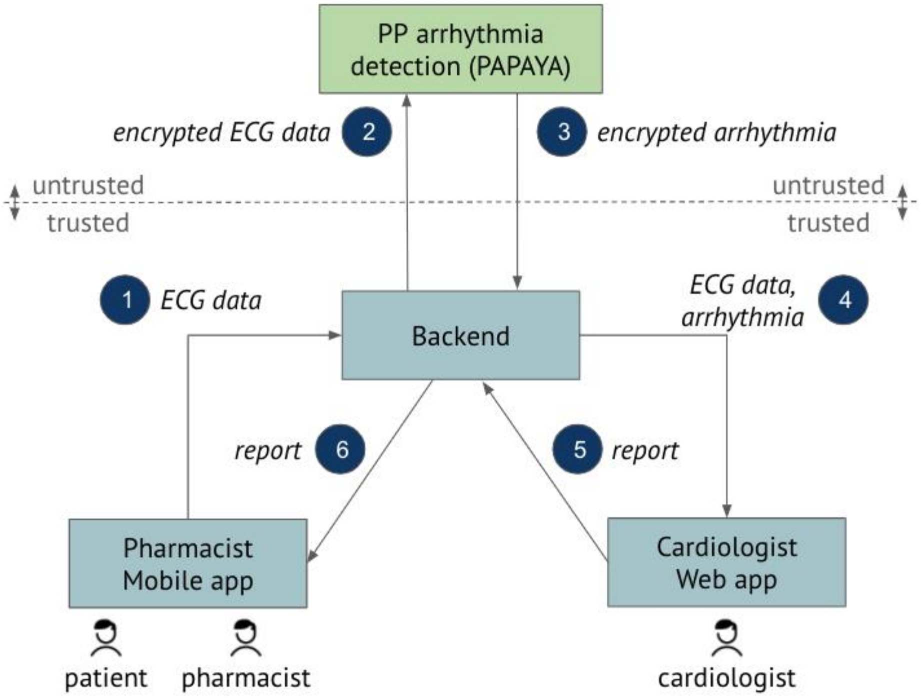


Figure 2 CardioPharma service, integrated with PAPAYA

In this newly proposed solution, the service remains the same for the involved parties, meaning that the pharmacist and the cardiologist continue to operate in the same way they operated before the integration with PAPAYA. The major difference is that the cardiologist facilitated in the production of the report, as he is provided also with hints on where arrhythmias could reside in the provided ECG track, and thus he does not have to go through the whole set of data.

2.2 Use cases specification validation

In this section, we validate the implementation against the use case definition specified in Deliverable D2.1.

2.2.1 Revision of GDPR roles

In this section, we briefly review the GDPR roles declared in Deliverable D2.1, as they were redefined after a deeper evaluation of the involved stakeholders and business model (see the Annex of Deliverable D6.4 for further details). In the following, we make clear how the GDPR roles change depending on the way the service is sold, as with different configurations the purposes of treatment are selected by different stakeholders.

2.2.1.1 Distribution of the service via hospitals

When the CardioPharma service is sold to hospitals, it is the hospital that selects CardioPharma as a service to treat its patients' data out of the market. Thus, the purposes of treatment are chosen by the hospital itself. Consequently (as reported in the Annex of D6.4):

- the patient is the **data subject**;
- the hospital is the **data controller**; the cardiologist and the nurse (replacing the pharmacist in this scenario) are its employees;
- MCI is the **data processor**;
- the external cloud provider is a **data processor**.

2.2.1.2 Distribution of the service via nursing homes

When the CardioPharma service is sold to nursing homes, the dynamics of choices and role distributions is quite similar to the one put in place with hospitals:

- the patient is the **data subject**;
- the nursing home is the **data controller**; the nurse (replacing the pharmacist in this scenario) is its employee;
- the cardiologist is a **data processor**;
- MCI is a **data processor**;
- the external cloud provider is a **data processor**.

2.2.1.3 Distribution of the service via pharmacies

When the CardioPharma service is sold to pharmacies, the dynamics of choices and role distribution differs from the ones we discussed regarding hospitals and nursing homes. Specifically:

- the patient is the **data subject**;

- the pharmacy and MCI are **joined data controllers**, as the pharmacy decides the purposes of processing for the biographical data of its clients, and MCI decides the purposes of processing for health-related (ECG) data collected from the patients;
- the cardiologist is a **data processor**;
- the external cloud provider is a **data processor**.

2.2.2 Coverage of use cases

In the following, we present the coverage table for the use cases presented in Deliverable D2.1:

Table 1 Coverage of use cases related to "Privacy-preserving arrhythmia detection"

Use Case	Status
UC-ECG-1 Detect arrhythmias in a patient's ECG	Implemented

The use case has been successfully implemented with the integration between CardioPharma and the PAPAYA platform. See Sections 2.2.4 and 2.2.5 for details about the integrated architecture and the app functionalities provided to the user.

2.2.3 Coverage of privacy requirements

In the following we present the coverage table for the privacy requirements presented in Deliverable D2.1:

Table 2 Coverage of privacy requirements for "Privacy-preserving arrhythmia detection"

Requirement	Status
Patients' data shall be pseudonymized when sent to the cardiologist	The cardiologist Web application does not visualize patients' biographical data (see Figure 10)
Patients' data shall be protected via PETs before sending them to the external cloud for classification	ECG data is never processed in cleartext when sent to the external cloud, as it is protected and analyzed via 2PC
Re-identification of patients on data outsourced to the PAPAYA platform shall not be possible	Re-identification of patients is not allowed outside premises, as identifiers are not shared with other actors and ECG data is protected and analyzed via 2PC

Consent shall be handled by the system, as a lawful basis for processing	Consent is collected at pharmacy side, when the patient accepts to have his data analyzed by CardioPharma
Performing any other analytics for other purposes rather than the one specified in the consent (i.e., analysis of ECG data to detect arrhythmias) shall be infeasible	The consent collected at pharmacy side is customized so that the purposes of processing are related only to the analysis of ECG data to detect arrhythmias
Processing shall be denied when a valid consent from the patient is not provided	Collection of consent at pharmacy side is the entrypoint to the service: as the pharmacist is a representative of the data controller, whenever this first step of interaction with the patient (i.e., the collection of consent) is not performed, the access to CardioPharma is denied to the patient and no interaction (and data collection) is initiated

2.2.4 Integration with PAPAYA platform

In this section, we describe the integration activities performed in task T5.1. Firstly, we list the PAPAYA components that were used (and thus, integrated) for UC1. Then, we present an architectural view of the integrated solution.

2.2.4.1 Integrated PAPAYA components

As reported in Section 2.1, the CardioPharma solution is structured in the following way:

- **Backend application**: this is a microservice communicating via REST interface, that detains the information regarding the performed analyses, registered patients etc.
- **Frontend applications**: these applications are the ones used by the final users (i.e., the pharmacist and the cardiologist):
    - **Mobile application:** the application used by the pharmacist (or the nurse in case the service is distributed to hospitals and nursing homes) to register the patient to the service and collect his data (that is, biographical data and ECG data);
    - **Web application:** the application used by the cardiologist to visualize the ECG and arrhythmia data and produce a report.

The PAPAYA platform, as briefly stated during the use case definition, allows us to perform two operations: a) on one side, it allows us to outsource the classification process of long ECG tracks in untrusted environments, by protecting data before outsourcing them, performing the classification on protected data, and unprotect the result only once it reaches the trusted environment; b) on the other side, it allows us to inform the data subject about the operations performed on his data.

Table 3 lists the PAPAYA tool we integrated for this use case, specifying also the CardioPharma components that were subjected to the integration. The next section will provide a more detailed view of the integration, depicting the architectural design of the integrated solution.

Table 3 Integrated PAPAYA component for "Privacy-preserving arrhythmia detection"

PAPAYA tool		**CardioPharma**	
DS tool 1	This tool explains the patient the basic functioning of the technologies behind PAPAYA	Mobile application	The pharmacists application allows patients to visualize PAPAYA's data subject tool 1 by opening an external browser. The integration has been held using an external browser in order not to give access to the patient with the complete application.
2PC	This component enables the outsourcing of the arrhythmia classification in an untrusted environment	CardioPharma backend *(through a specialized integration service)*	The Cardiopharma backend coordinates with PAPAYA's 2PC component in order to require the arrhythmia classification, and store it once computed. The classification is finally shown to the cardiologist that can elaborate a diagnosis.

2.2.4.2 Integrated architecture

Figure 3 shows the architectural representation of the integrated CardioPharma-PAPAYA solution. The figure shows how components (both from CardioPharma and PAPAYA) are connected.

On the one hand, the integration of 2PC has been performed by implementing an integration service (in the figure referenced as "CardioPharma-2PC integration service") that translates the entities coming from the domain of CardioPharma into the format expected by PAPAYA. This component communicates via REST interface, as all the CardioPharma and PAPAYA services. Its functionalities include:

1. the registration of the analyses performed via the PAPAYA 2PC component;
2. the storage and the update over time of their statuses (being it "*in progress"* when the result is not reached, "*completed"* when the classification has been performed and "*rejected"* when the classification process ended with an error).

On the other hand, the integration of DS Tool 1 has been performed on the CardioPharma mobile application, which is the touchpoint for the pharmacist (and thus, for the patient).

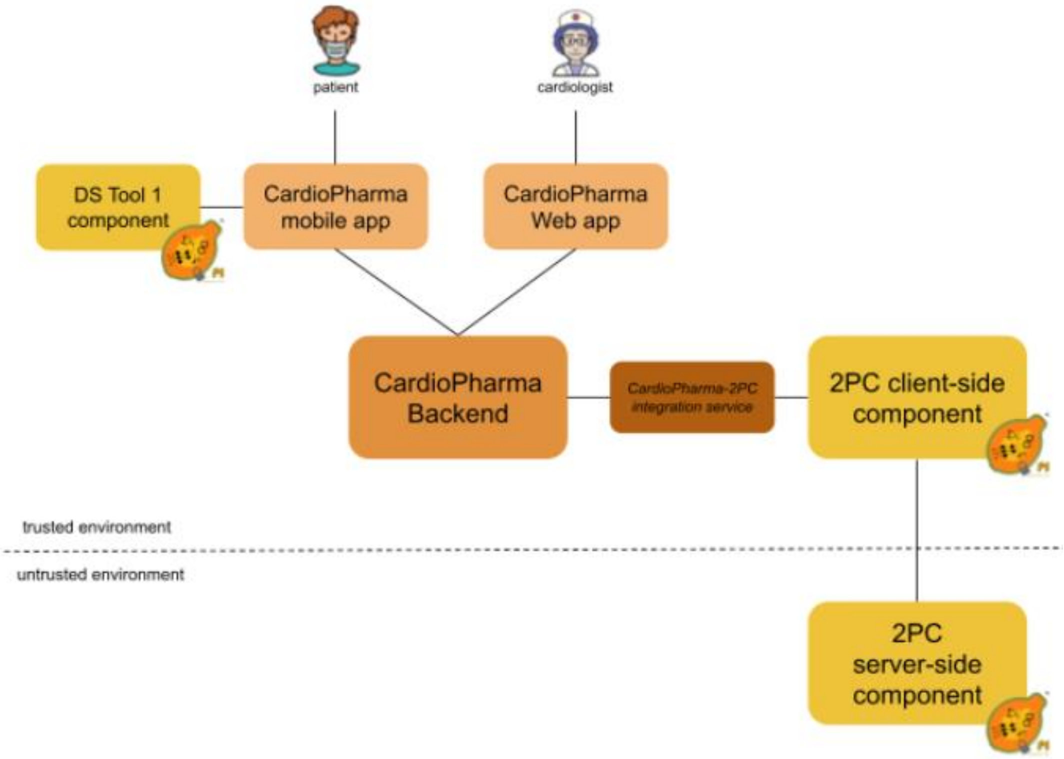


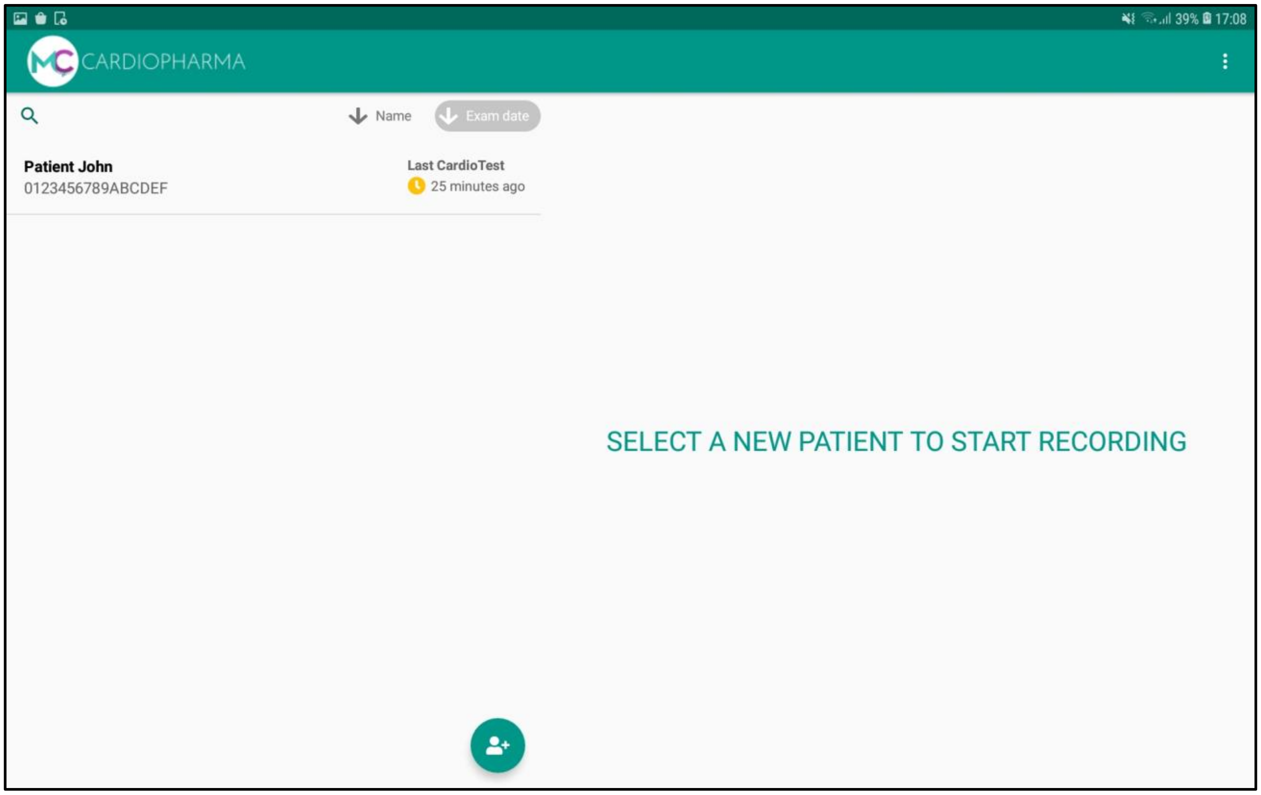
Figure 3 Integration of CardioPharma and the PAPAYA platform

2.2.5 Applications implementation: interfaces

In this section, we present a tour of the functionalities offered by the CardioPharma service integrated with the PAPAYA platform.

2.2.5.1 CardioPharma mobile application

In this section we show the interfaces offered by CardioPharma to allow a pharmacist to register new exams for patients and send them for analysis.



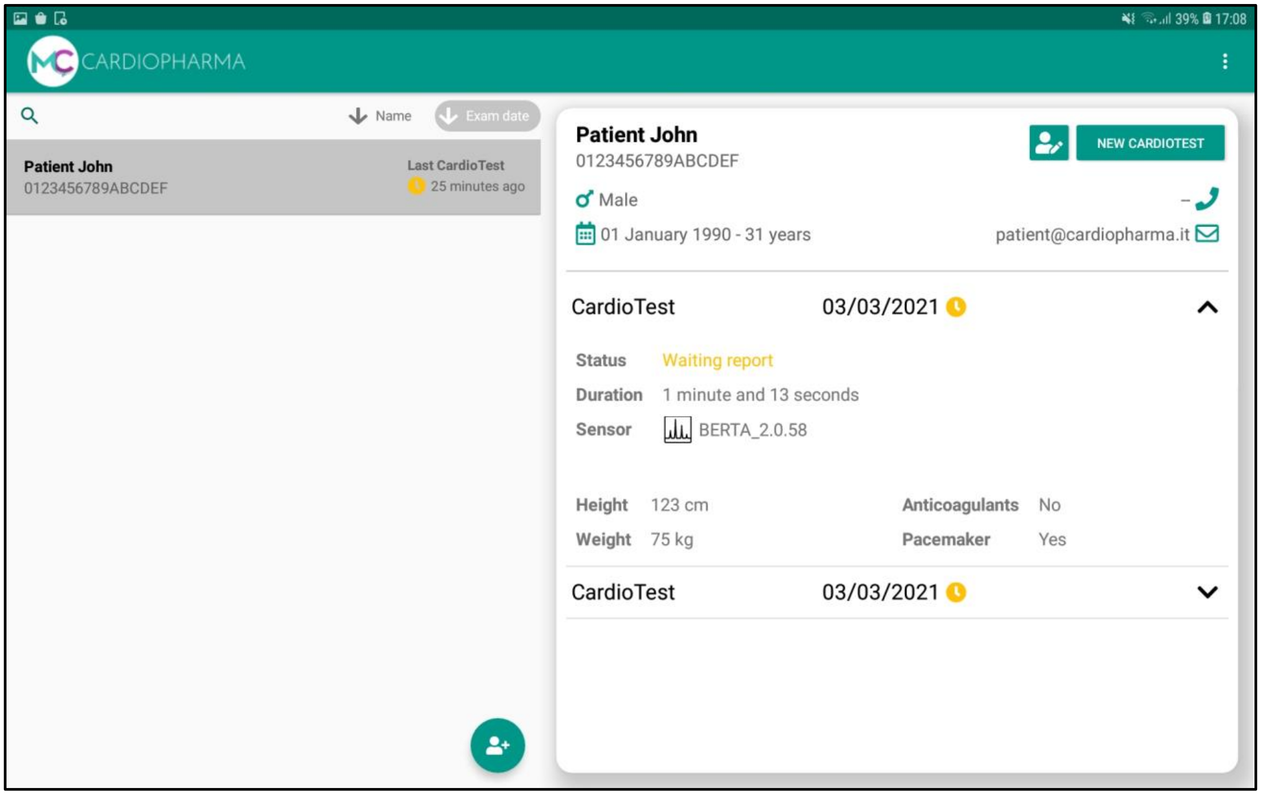
Patient John
0123456789ABCDEF

Last CardioTest
25 minutes ago

SELECT A NEW PATIENT TO START RECORDING

Figure 4 CardioPharma pharmacist application: initial dashboard

Figure 4 shows the initial dashboard of the mobile application (in this case, installed on a tablet). From here, the pharmacist can decide either to register a new patient (by pressing the button on the lower part of the screen) or to select an existing patient to start a new ECG recording (i.e., a new analysis).



Patient John
0123456789ABCDEF
Last CardioTest
25 minutes ago

Patient John
0123456789ABCDEF
Male
–
01 January 1990 - 31 years
patient@cardiopharma.it

CardioTest 03/03/2021

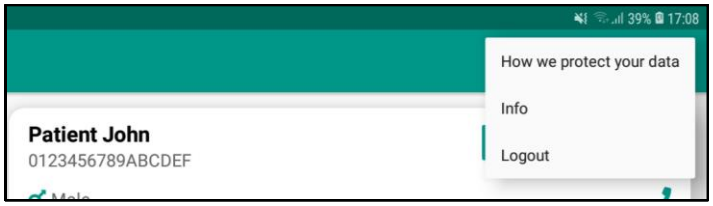
Status	Waiting report
Duration	1 minute and 13 seconds
Sensor	BERTA_2.0.58

Height	123 cm	Anticoagulants	No
Weight	75 kg	Pacemaker	Yes

CardioTest 03/03/2021

Figure 5 CardioPharma pharmacist application: patient profile

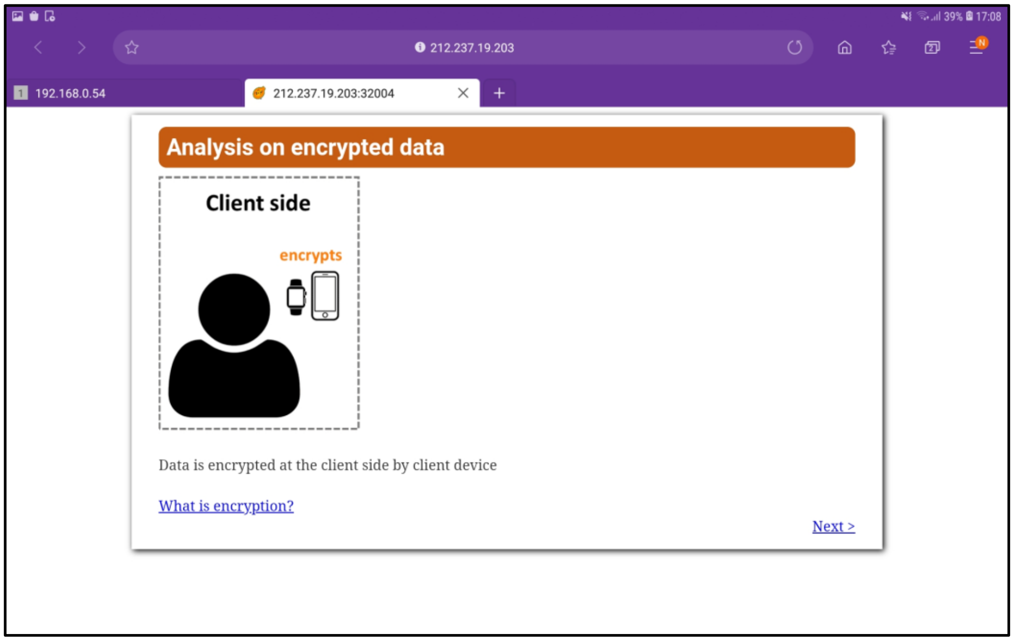
Once a patient is selected, we can see his biographical data and the analyses (indicated as "CardioTest") he did in the past. We can also start another analysis, or we can show him the content of DS tool 1.



Patient John
0123456789ABCDEF

Figure 6 CardioPharma pharmacist application: access to DS Tool 1

Figure 6 shows how to access the content of DS Tool 1 from the patient card.



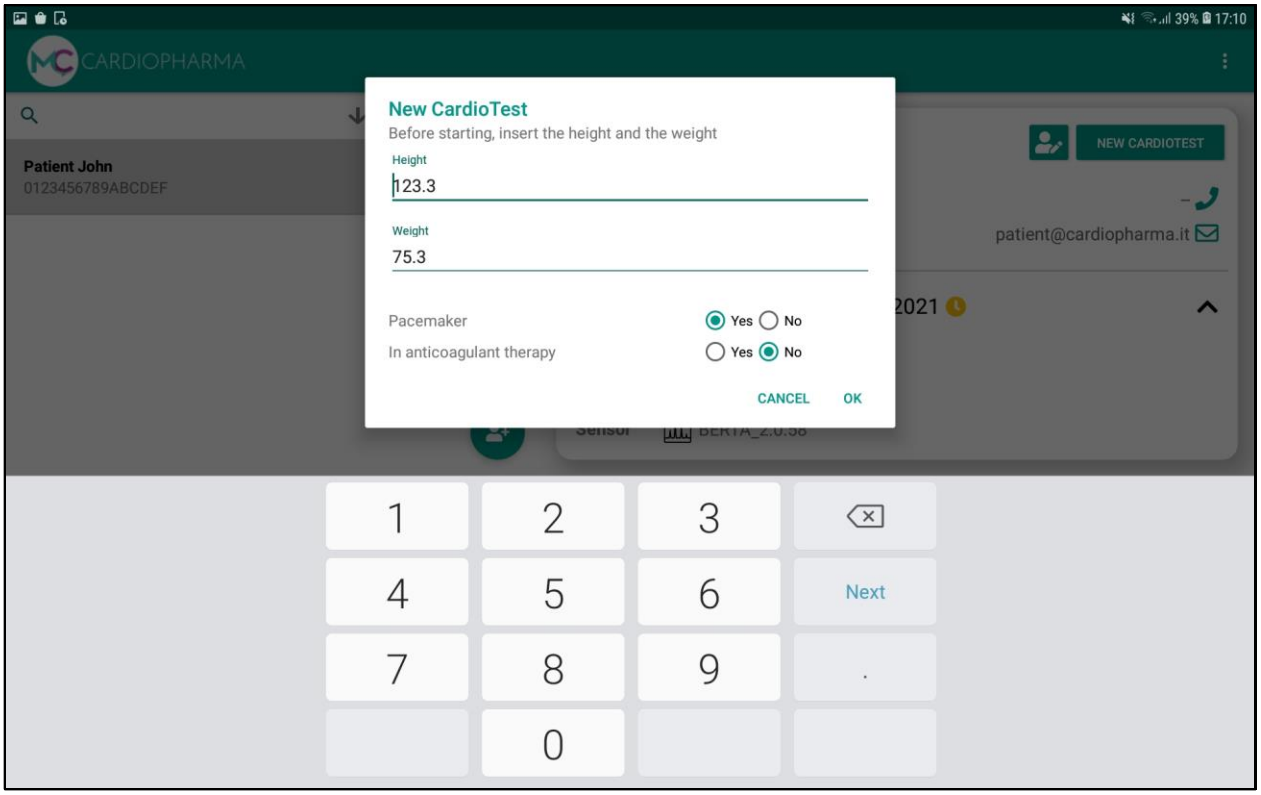
Analysis on encrypted data

Client side

Data is encrypted at the client side by client device

Figure 7 CardioPharma pharmacist app: content of DS Tool 1

Figure 7 instead shows the interface for DS tool 1.



Patient John
0123456789ABCDEF

patient@cardiopharma.it

Sensor BERTA_2.0.58

New CardioTest

Before starting, insert the height and the weight

Height: 123.3

Weight: 75.3

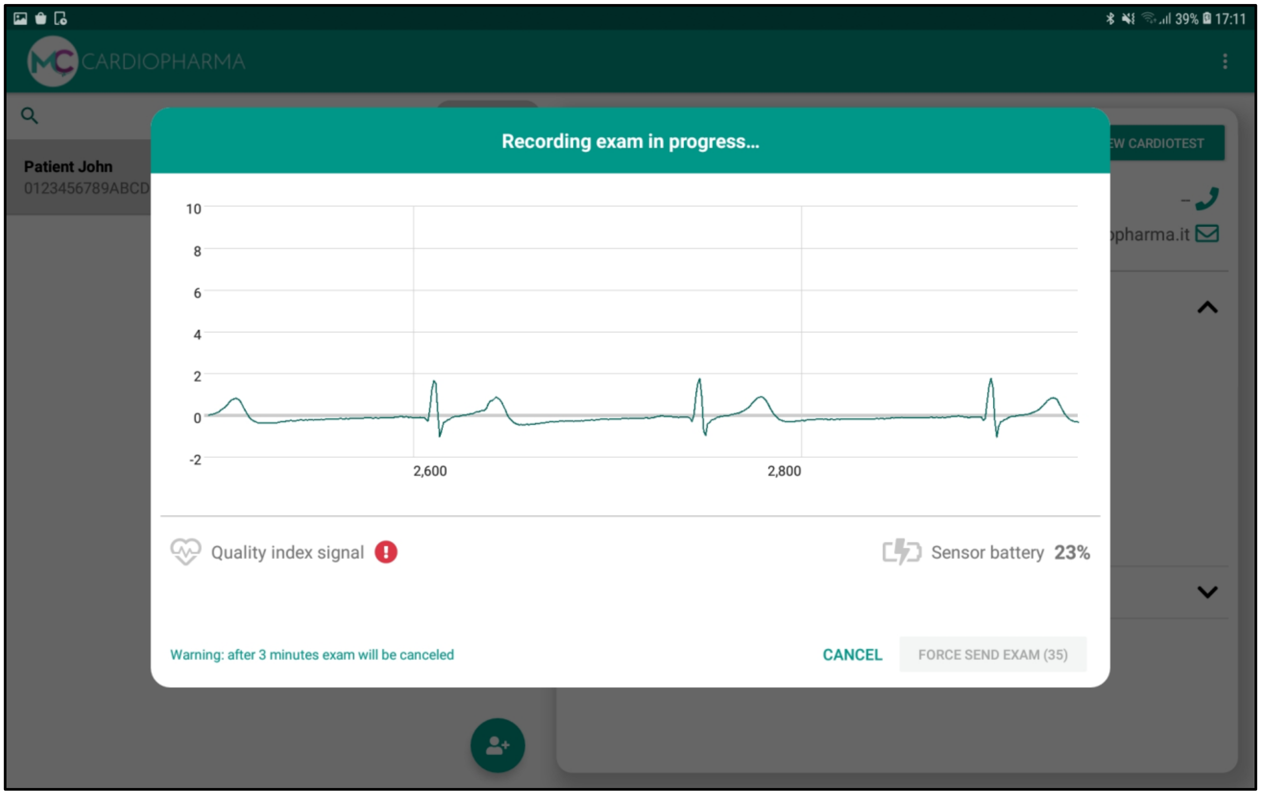
Pacemaker ● Yes ○ No

In anticoagulant therapy ○ Yes ● No

CANCEL OK

Figure 8 CardioPharma pharmacist application: starting a new analysis

Figure 8 shows the interface that the pharmacist uses to start a new ECG analysis for the selected patient. At first, she has to specify some anamnesis fields (e.g., height, weight, presence of pacemaker, usage of anticoagulants). Then, she has to pair the application with the ECG monitoring device that will be given to the patient. After that, the exam can start.


Recording exam in progress...

Patient John
0123456789ABCD

Quality index signal

Sensor battery 23%

Warning: after 3 minutes exam will be canceled

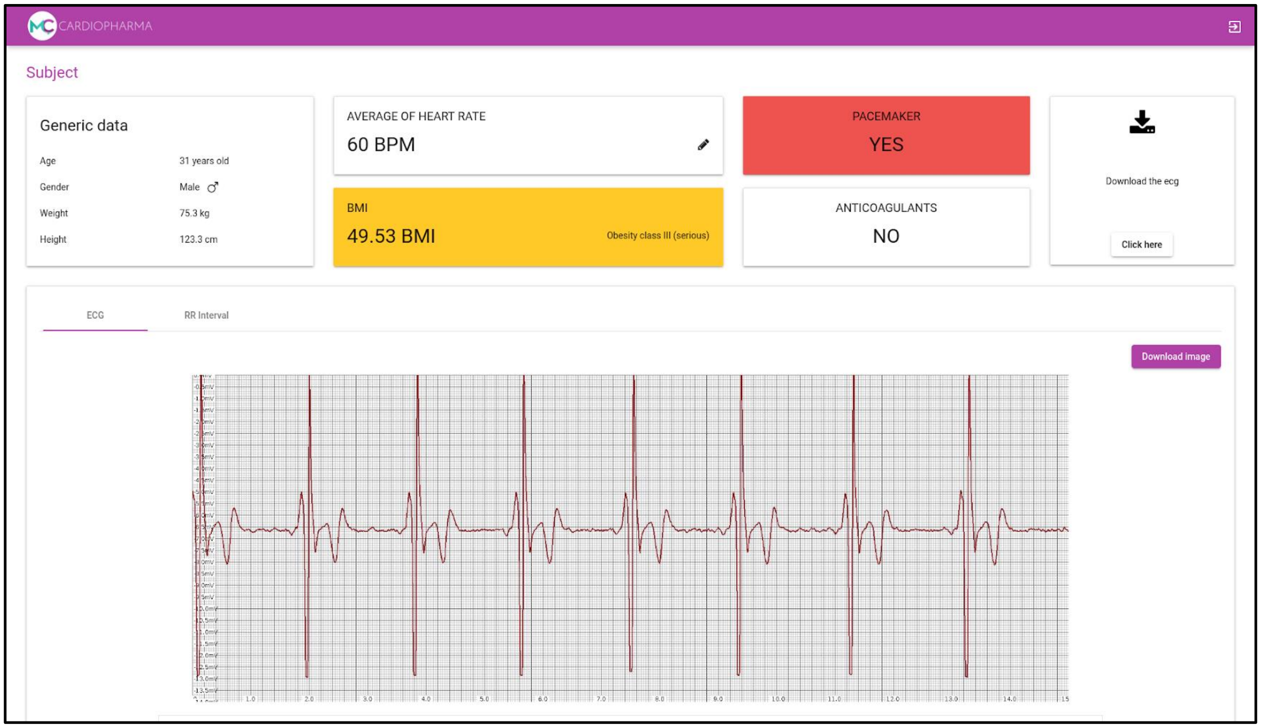
Figure 9 CardioPharma pharmacist application: exam in progress

Figure 9 shows an exam in progress. It can be seen that the ECG signal is being recorded and the sensor battery level is available on the screen. Also, there is a measurement of the signal quality, that assesses the quality of the measured ECG, which could be used by the pharmacist to understand if it is the case to move the sensor a bit on the patient's skin to get a better result.

Once the registration of ECG is over, it is automatically sent to the CardioPharma backend for arrhythmia analysis, and from there it will be sent to the cardiologist Web application for reporting (see next section). It is also possible for the pharmacist to force sending the exam for analysis before the monitoring period is over, in case shorter exams are needed.

2.2.5.2 CardioPharma Web application

In this section, we show the interfaces offered by the CardioPharma cardiologist Web application, that he can use to produce a report of the arrhythmias for the patient.



CARDIOPHARMA

Subject

Generic data

Age	31 years old
Gender	Male ♂
Weight	75.3 kg
Height	123.3 cm

AVERAGE OF HEART RATE
60 BPM

PACEMAKER
YES

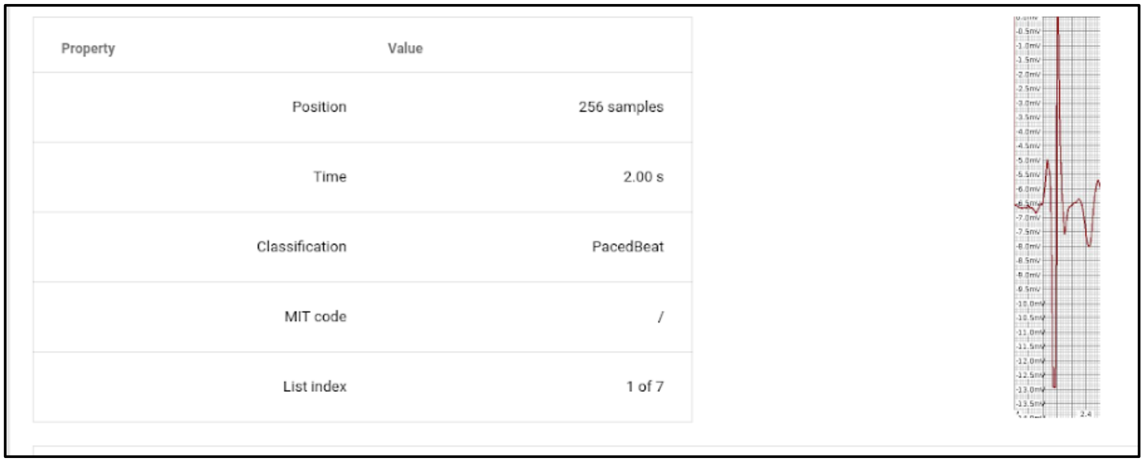
BMI
49.53 BMI
Obesity class III (serious)

ANTICOAGULANTS
NO

Download the ecg
Click here

Figure 10 CardioPharma cardiologist application: overview of an exam

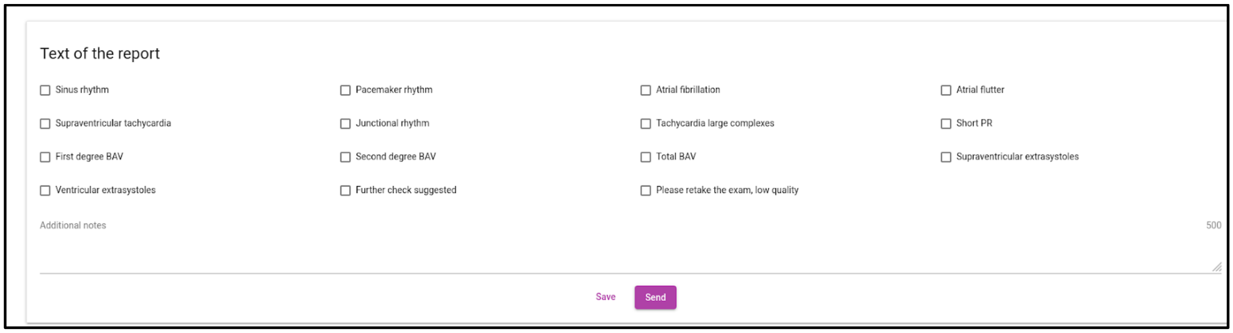
Figure 10 shows the overview of a registered exam, where the cardiologist can see part of the patient's biographical data (but not the identifiers such as name and surname), the ECG and related data, the anamnesis.



Property	Value
Position	256 samples
Time	2.00 s
Classification	PacedBeat
MIT code	/
List index	1 of 7

Figure 11 CardioPharma cardiologist application: arrhythmia classification

Figure 11 shows the classification given by the classification (performed via PAPAYA 2PC component) for each beat. The cardiologist can filter beats and visualize in detail the ones that the system recognizes as arrhythmia, and use this data to build his report.



Text of the report

☐ Sinus rhythm ☐ Pacemaker rhythm ☐ Atrial fibrillation ☐ Atrial flutter

☐ Supraventricular tachycardia ☐ Junctional rhythm ☐ Tachycardia large complexes ☐ Short PR

☐ First degree BAV ☐ Second degree BAV ☐ Total BAV ☐ Supraventricular extrasystoles

☐ Ventricular extrasystoles ☐ Further check suggested ☐ Please retake the exam, low quality

Additional notes 500

Figure 12 CardioPharma cardiologist application: report creation

Finally, Figure 12 shows the interface that the cardiologist can use to create the report. Once he selects the recognized and validated arrhythmia from the ones specified in the interface, the report is created and sent back to the pharmacist application. From there, the pharmacist can download and print it for the patient.

2.3 Requirements validation

In this section, we validate the implementation against the requirements specified in Deliverable D2.2 (Section "*Requirements derived from Interviews with medical professionals (UC1)*")".

Table 4 Requirements coverage for Use Case 1

ID	Title	Acceptance Criteria	Validation
UC1.EUR.HCI.1	Communicating protection of outsourced data	Stakeholders SHOULD be informed by introductory tutorials and/or consent forms.	Covered; the integration of the Data Subject toolbox with this use case ensures that the technologies behind PAPAYA are well explained, and that it is clear that outsourcing is done only upon protection
C.EUR.HCI.2	Assurance guarantees	Assurance Certification by a recognised authority, and /or reports on validated research study SHOULD be made available and be communicated to stakeholders.	Partially covered; CardioPharma is produced by MediaClinics Italia, which is certified ISO 13485. If the proposed application (CardioPharma with integration with PAPAYA components) hits the market, it would be done under certification, where the software would be certified ISO/IEC 62304)

			the PAPAYA components would be treated as Software of Unknown Provenance (as per ISO/IEC 62304) the PAPAYA components would need to undergo thorough testing before getting to certification
C.EUR.HCI.3	Communication privacy and utility benefits and trade-offs	Detailed information about the PIA process and evaluator SHOULD be made available by the user interface or by other means. The PIA SHOULD be conducted by a qualified expert.	Partially covered; currently MediaClinics Italia is undergoing a process in which the DPIA for new products (such as the cardiology suite) is redacted with the help of the company DPO. This documentation should be made available for consultation
UC1.EUR.HCI.2	Informing doctors	Introductory tutorials or other sources of information for informing doctors SHOULD exist.	Covered; in this scenario, cardiologists are collaborators of MediaClinics, and thus they would be engaged in the usage of CardioPharma with an explanation of how it works.
UC1.EUR.HCI.3	Informing patients on technical privacy measures	Usable consent and policy information and/or information leaflets SHOULD be in place to inform patients accordingly.	Covered; The DS Tool for UC1 informs patients in different layers of detail about the data analysis that is conducted on encrypted data.

2.4 Validation by stakeholders

In this section, we validate the implementation by presenting it to the stakeholders.

The data subject tool for explaining data analytics on encrypted data for UC1 was already tested with end users during the iterative development process. The evaluation results were reported in D3.4 [12]. Moreover, interviews on UC1 with eHealth stakeholders were already performed during the requirement elicitation phase, preliminary results related to the elicited requirements were summarised in D2.2. The interviews were later thoroughly evaluated by conducting a thematic analysis, which revealed detailed insights about the eHealth stakeholders' perceptions of PAPAYA in the context of UC1. For detailed results of this stakeholder evaluation of UC1, we refer to our publications [13, 14].

3 UC2: Privacy-preserving stress management

In this chapter, we present the validation activities carried out for the second use case in the eHealth scenario, namely, the *Privacy-preserving stress management* use case.

3.1 Use case description in a nutshell

The *privacy-preserving stress management* use case targets workers who suffer from stress and need to keep their stress levels in check.

People suffering from stress have the problem of not being always able to recognize that their stress level is rising, and when they realize it the stress level may be so high that it negatively affects their health status. It would then be very helpful to have an automatic tool that is able to recognize stress in patients *at its onset*, and suggest mitigation actions (e.g., mindfulness exercises) to lower their stress levels before they become too high to be handled.

Thus, this use case is targeted at building an automatic tool for stress management that recognizes automatically (i.e., via an AI trained model) stress situations in workers and gives suggestions about when it is time to apply countermeasures to lower the stress.

At the time the use case was proposed in Deliverable D2.1, research about how to properly detect stress in workers (even manually) was still ongoing, as this functionality is not a diffused one and there is still a lot of research going on to find better and better tools to properly recognize stress and anxiety from physiological parameters. This is why at the beginning of the PAPAYA project this use case was presented as a research-oriented use case: while its application had definitely value from a business perspective, its development had to start from a less mature point of view in comparison with UC1. Now that we performed a deep-dive into the related literature and we identified a sound picture of how one can assess stress/no-stress situations by observing physiological parameters, we can review what was declared in Deliverable D2.1 by adding more details to it and suggesting corrections to what was previously declared. The following subsections accompany the reader through this path.

3.1.1 Physiological parameters suggesting stress

The detection of stress (being it manual or automatic) can be done by monitoring several physiological parameters, according to the reviewed literature. We list the ones the literature cites the most in the following:

- the **heart rate variability** (or, **HRV**) [1][2][3][4][5][6][7];
- the **galvanic skin response** [1][2][6][7][8][9];
- the **electrocardiogram** (or, **ECG)** [1][2][3][4][5][9];
- the **electroencephalogram** (or, **EEG)** [1][2];
- the **blood volume pulse** (or, **BVP**) [1][2][6][7].

There are many other parameters that can be used to track stress (e.g., blood pressure, electromyogram, skin temperature, respiration, body language, facial expression, pupil dilation, blink rate, voice, etc.). These parameters can be even combined to gain a better perspective of the stress level of a person. Nevertheless, these ones are less cited in the literature, and are either derivable from the ones listed before, or difficult to measure (unless we are willing to make the worker wear several sensors and cables along his working day, which would not be convenient, and would make a potential business service lose applicability).

Thus, we decided to exclude these parameters and focus on the most used ones. Specifically, when reviewing the literature, we found out that the **HRV parameter** is one of the most frequently used (even by cardiologists) to understand if a person is subjected to stress.

The capability of HRV to capture stress-related situations is directly related to its definition. Indeed, the Heart Rate Variability (HRV) [11] is a measure that expresses the physiological phenomenon of variation in the time interval between heartbeats. It is commonly measured by the variation in the RR intervals, i.e., the time period between two consecutive R peaks in the ECG signal.

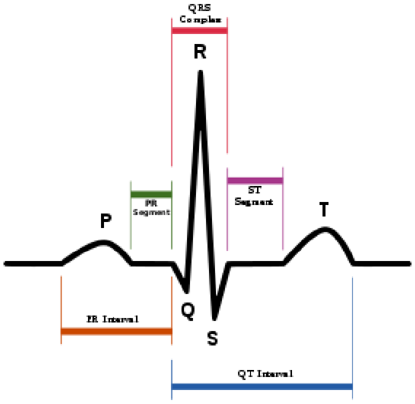


Figure 13 A heartbeat showing its R peak [10]

Variation in the beat-to-beat interval can be affected by the status in which a person finds himself: indeed, variability increases during relaxation and decreases during stress. This is due to the fact that HRV is typically higher when the heart beats slowly and decreases when the heart beats more quickly (e.g., during stressful situations or during physical activity). Hence, when a person is at rest (e.g., when working in an office) a reduction in HRV could suggest a stress situation for the person.

3.1.2 Building a NN model that detects stress

Several computational techniques have been applied in the literature to recognize stress situations from physiological parameters, e.g., Naive Bayes classification [1,9], decision trees [1,9], artificial neural networks [1,2,3,9], support vector machines [1,2,6], hidden Markov models [1], linear discriminant analysis [2,3] etc. We focus on the ones based on artificial neural networks, so as to be able to exploit the techniques made available by PAPAYA.

The work in [3] proposes a method for automatically recognizing the stress in workers by training a **convolutional neural network** on a dataset with the following form:

- **Feature vector:** HRV features
- **Label:** not stressed, stressed due to interruptions, stressed due to time pressure

The following Figure shows a summary of the pipeline (proposed by the paper) that has to be applied in order to extract a stress/no-stress label from an ECG:

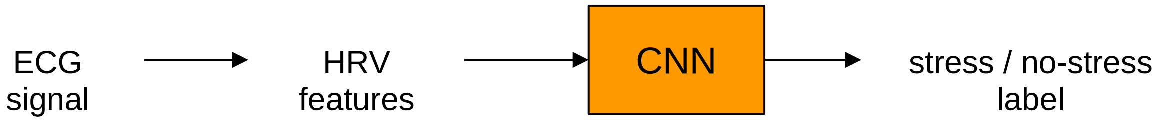


Figure 14 The stress-detection pipeline

This pipeline can be easily applied in the real-world when users are workers in an office: by using a sensorized T-shirt able to measure continuously ECG tracks, the ECG can be translated in real-time into HRV features, that are fed into the CNN to get a classification.

3.1.3 Building a dataset to train the NN model

The aforementioned pipeline can be easily reproduced whenever we are in possession of a dataset that relates HRV features with stress/no-stress situations. Unfortunately, the work in [3] did not release the dataset used to train the CNN, and hence one can re-implement the work done in [3]:

1. either by using a publicly available dataset (such as the one in [9]; in such a case the labels “stress due to interruptions” and “stress due to time pressure” are converted into a single “stress” label);
2. or by building a labeled dataset out of data collected on workers.

Speaking of this second option, our preliminary understanding (also reported in Deliverable D2.1) was that we could build a dataset with the help of workers. In such a case, each worker that was willing to participate in the data collection would: a) wear a sensorized T-shirt (able to detect their ECG signal and transform it into HRV features); b) use a mobile application to signal every time he would feel stressed.

Nevertheless, after careful review of the literature, we can claim that the environment in which these measurements would be taken and the actual impossibility (for a person) of feeling the stress rising at its onset would prevent a worker to create a training dataset that is of enough quality to train a neural network.

Hence, we decided to change the strategy used to collect the dataset, in a way that is more coherent with respect to what we found in the literature, and that is actually done by cardiologists specialized in stress. This change of strategy will have an impact on the use cases declared in Deliverable D2.1, see Section 3.2 of this deliverable for further details. The new strategy used to collect a labeled dataset for performing stress/no-stress classification is as follows:

- we use a sensorized T-shirt to collect **ECG tracks** on workers and transform them into their HRV features;
- a **cardiologist** (instead of the worker) is asked to tag the several collected ECG tracks in stress / no-stress classes, by looking at the related HRV graphs and features and understanding if they are associated with a stress situation or not.

3.1.4 The need for large datasets: how can we collect enough data to build a better performing model?

The process of creating a dataset for this type of classification is burdensome, as: a) on one side, it requires to gather a log of ECG tracks from workers, which may not be willing to participate to data collection or even not willing to wear the sensorized T-shirt for a long time; b) on the other side, it requires to find a cardiologist that is able to understand HRV graphs in relation to stress, and collect tons of labels (one per each collected ECG track, say, of 5 minutes).

Suppose that a company decides to put in place an automatic tool to detect stress in its workers. Building the aforementioned model is not complicated, but on the contrary the collection of the training set can be problematic, specifically in case of SMEs that may have a small set of employees.

Thus, in this use case we propose to enable companies to participate in **collaborative training tasks**, where:

- each company shares its own model, trained locally with its workers' data;
- a joint model is built by merging the knowledge contained in the locally trained models;
- the joint model is shared among all the companies that participated in the collaborative training, so that they can use it to recognize stress in their workers.

In this way, companies make the most out of the small set of data they possess, and are enabled to put in place services that promote the welfare of their workers.

3.2 Use cases specification alidation

In this section, we validate the implementation against the use case definition specified in Deliverable D2.1.

3.2.1 Revision of GDPR roles

As for the case of UC1 "*Privacy-preserving arrhythmia detection*", the GDPR roles declared in Deliverable D2.1 were revised based on the revised exploitation plan of the service. Specifically:

- the worker is the **data subject**
- the company is a **joint data controller** and buys the service from MCI
- MCI is a **joint data controller**
- the external cloud provider is a **data processor**

3.2.2 Coverage of use cases

In the following, we present the coverage table for the use cases presented in Deliverable D2.1:

Table 5 Coverage of use cases related to "Privacy-preserving stress management"

Use Case	Status
UC-STR-1 Collect stress-related dataset	Reviewed, implemented
UC-STR-2 Train collective model	Implemented
UC-STR-3 Classify worker's data to identify a stress condition	Implemented

As reported in the table, and discussed in Section 3.1.3, **UC-STR-1** was revised and actually divided into two different use cases:

- **UC-STR-1a** Collect ECG track
- **UC-STR-1b** Tag ECG track with stress label

Notice that to collect the dataset and make a cardiologist tag it, we will use the CardioPharma service described in Section 4 for UC1.

In the following, we provide the description of such use cases. These will replace the definition of UC-STR-1 provided in Deliverable D2.1.

Table 6 Description of UC-STR-1a

ID and name	***UC-STR-1a*** Collect ECG track
Primary actor	MC shirt
Secondary actors	Worker
Description	Workers in a workplace contribute to the creation of a stress-related dataset, by collecting their ECG tracks via a sensorized T-shirt. The T-shirt collects an ECG track every X minutes (X chosen as parameter).
Preconditions	***PRE-1*** Worker's company is registered to the service ***PRE-2*** Worker is registered to the service ***PRE-3*** *A consent to treat worker's data is signed by the worker* ***PRE-4*** Worker wears the MCI sensorized T-shirt
Postconditions	***POST-1*** A set of ECG tracks is collected
Normal flow	**1a.0 Dataset collection** 1. The sensor in the T-shirt collects X minutes of ECG track

	2. The sender sends the registered ECG track to CardioPharma 3. CardioPharma stores the track in the dataset
Alternative flow	-
Exceptions	-
Assumptions	-

Table 7 Description of UC-STR-1b

ID and name	***UC-STR-1b*** Tag ECG track with stress label
Primary actor	Cardiologist
Secondary actors	-
Description	The cardiologist visualizes an untagged ECG track and the related HRV graphs, and, based on their content, tags the track as either “stress” (i.e., stress-related) or “non-stress” (i.e., not stress-related).
Preconditions	***PRE-1*** Cardiologist is signed to the platform ***PRE-2*** The identity of the cardiologist is verified
Postconditions	***POST-1*** The ECG track is tagged
Normal flow	**1b.0 ECG Labelling in stress situations** 1. The cardiologist asks to visualize the ECG track 2. The system visualizes the track together with its HRV graphs 3. The cardiologist recognizes that the track is related to a stress situation and thus tags the track with the “stress” label
Alternative flow	**1b.1 ECG Labelling in non-stress situations** 1. The cardiologist asks to visualize the ECG track 2. The system visualizes the track together with its HRV graphs 3. The cardiologist recognizes that the track is NOT related to a stress situation and thus tags the track with the “non-stress” label
Exceptions	-
Assumptions	-

3.2.3 Coverage of privacy requirements

In the following, we present the coverage table for the privacy requirements presented in Deliverable D2.1:

Table 8 Coverage of privacy requirements for "Privacy-preserving stress management"

Requirement	Status
Neural network model input shall be protected via PETs before outsourcing them to the external cloud	Neural network models (weights) are actually protected before outsourcing using the privacy-preserving collaborative training functionality of the PAPAYA platform
Re-identification of workers from models outsourced to the PAPAYA platform shall not be possible	This is ensured by the privacy-preserving collaborative training functionality
Consent shall be handled by the system, as a lawful basis for processing	The requirement is covered as processing without specific consent is not allowed by the system
Performing any other analytics for other purposes rather than the one specified in the consent shall be infeasible	The requirement is covered as processing without specific consent is not allowed by the system
Processing shall be denied when a valid consent from the worker is not provided	The requirement is covered as processing without specific consent is not allowed by the system
Privacy preferences (e.g., decision on when to monitor workers and what data to monitor) shall be handled by the system	The requirement is covered as processing without specific consent is not allowed by the system
Data collection shall be performed only when compliant to the privacy preferences specified by the workers	The requirement is covered as processing without specific consent is not allowed by the system
Retrieving data from single workers shall not be feasible	This is ensured by the privacy-preserving collaborative training functionality
Computation of models on data coming from a single worker shall not be feasible (to avoid re-identification)	This is ensured by the privacy-preserving collaborative training functionality
Analytics shall not be performed before multisource data are aggregated	This is ensured by the privacy-preserving collaborative training functionality
Right to erasure shall be supported by the system	The privacy preferences that the user can specify cover also the right to erasure

3.2.4 Integration with PAPAYA platform

In this section, we describe the integration activities performed in task T5.1. Firstly, we list the PAPAYA components that were used (and thus, integrated) for UC2. Then, we present an architectural view of the integrated solution.

3.2.4.1 Integrated PAPAYA components

As reported in Section 3.1, our implementation plan is to:

- use CardioPharma (i.e., the MCI production-ready service used also for UC1) to **collect the ECG dataset** and **make the cardiologist label it** (as stress/non-stress related);
- use CardioPharma to **build the model** to detect stress situations;
- use a mobile application to allow the worker to **monitor his stress levels** and **receive notifications** when these levels are rising.

The CardioPharma components are the same as the ones discussed in Section 2.2.4.1. The main difference is that here the Web application:

- is used by the cardiologist to label ECG data as stress/no-stress;
- is used by the system administrator to open new dataset tagging campaigns, by selecting some dataset (e.g., a set of unlabeled ECG tracks) and selecting the labels a labeler (e.g., the cardiologist) could use to label data.

The PAPAYA platform, as briefly stated during the use case definition, allows us to perform two operations: a) on the one side, it allows us to outsource the model construction by creating consortiums of companies that share their local models to train a better performing model (via collaborative training); b) on the other side, it allows us to inform the data subject about the operations performed on his data, and give his permissions to modify his privacy preferences.

Table 9 lists the PAPAYA tool we integrated for this use case, specifying also the MCI components (i.e., either CardioPharma components or the mobile application) that were subjected to the integration. The next section will provide a more detailed view of the integration, depicting the architectural design of the integrated solution.

Table 9 Integrated PAPAYA component for "Privacy-preserving arrhythmia detection"

PAPAYA tool		**CardioPharma**	
DS tool 1	This tool explains the worker the basic functioning of the technologies behind PAPAYA	Mobile application	The mobile application allows workers to visualize PAPAYA's data subject tool 1 by opening an external browser. The integration has

			been held using an external browser in order not to give access to the worker with the complete application.
DS tool 2	This tool provides the worker with information about data disclosure	Mobile application	The mobile application allows workers to visualize PAPAYA's data subject tool 1 by opening an external browser. The integration has been held using an external browser in order not to give access to the worker with the complete application.
Privacy Engine	This tool provides the worker with the possibility of choosing his privacy preferences	Mobile application	The mobile application interacts with the Privacy Engine allowing the worker to set up his privacy preferences
Collaborative training	This component enables the construction of a single, better performing model using local models from several companies	CardioPharma backend *(through a specialized integration service)*	The Cardiopharma backend coordinates with PAPAYA's 2PC component in order to require the arrhythmia classification, and store it once computed. The classification is finally shown to the cardiologist that can elaborate a diagnosis.

3.2.4.2 3.2.4.2 Integrated architecture

Figure 15 shows the architectural representation of the fully integrated CardioPharma-PAPAYA solution for UC2. The figure shows how components (both from CardioPharma and PAPAYA) are connected. The pictorial representation shows:

- on the left side, what is needed by a company to share its locally trained model and participate to a collaborative training session;
- on the right side, what is needed by a company to use the collaboratively trained model to recognize stress levels in its workers.

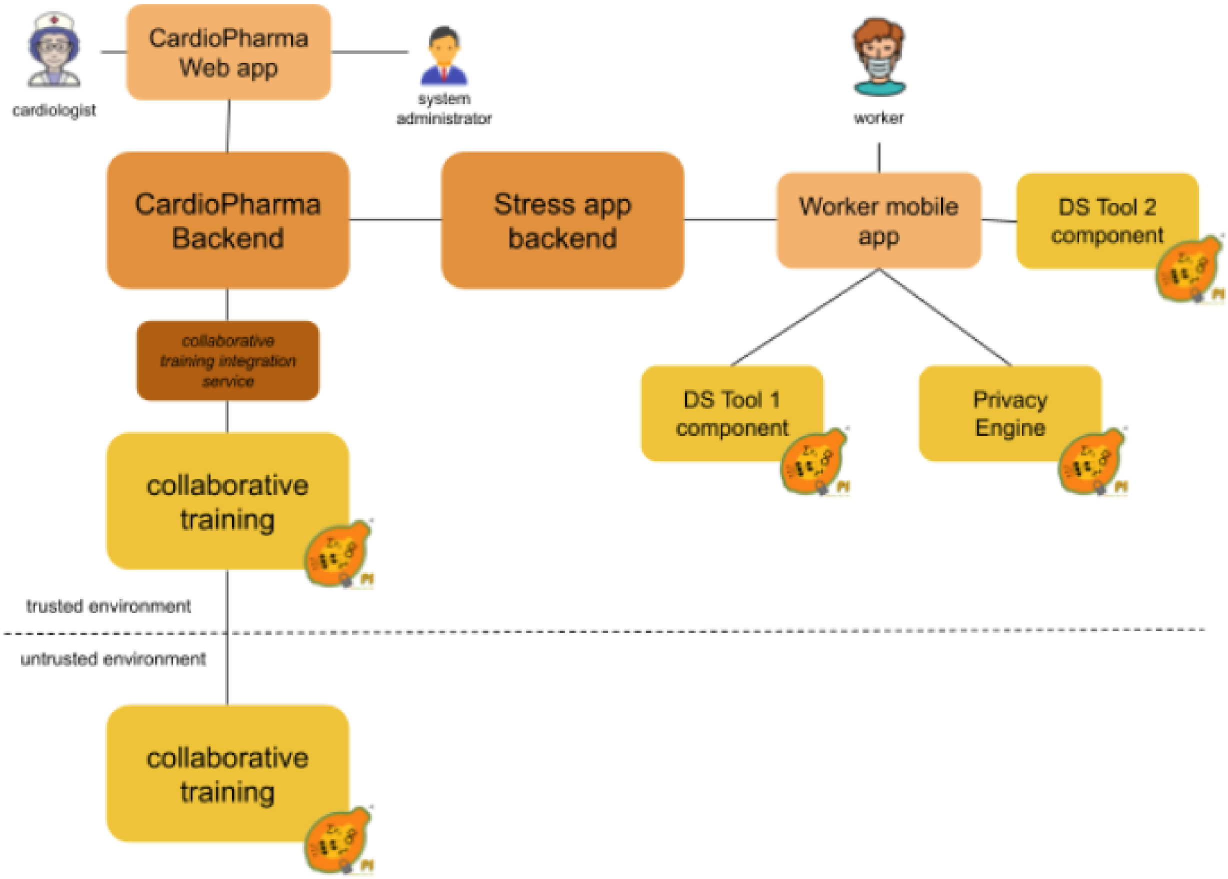


Figure 15 Integration of CardioPharma, Stress app and the PAPAYA platform

In the following, we explain in more details the content of the Figure.

Each company that is interested in using the collaborative training component (so as to access a better trained model for stress detection) installs in its premises a CardioPharma instance. Such an installation can thus declare itself as a participant to a collaborative training session, create a locally trained model and send it to the collaborative training server side for integration with other models created by other companies. As for the first use case, the integration of the collaborative training component has been performed by implementing an integration service (in the figure referenced as “collaborative training integration service”) that translates the entities coming from the domain of CardioPharma into the format expected by PAPAYA. This component communicates via REST interface, as all the CardioPharma and PAPAYA services. Its functionalities include:

1. the registration of the collaborative training processes over time;
2. the monitoring of their statuses (being, e.g., “*in progress”* when the result is not reached, or “*completed”* when the model has been trained).

This mechanism is obviously replicated for many participants, as shown in Figure 16: there are many CardioPharma installations, one for each participating company, and each one of these installations retains its own tagged dataset. When they decide to participate in a common training session, all the CardioPharma installations coordinate between each other to declare themselves as participants to the training, train locally all their models, share the locally trained model and obtain the collaboratively trained model.

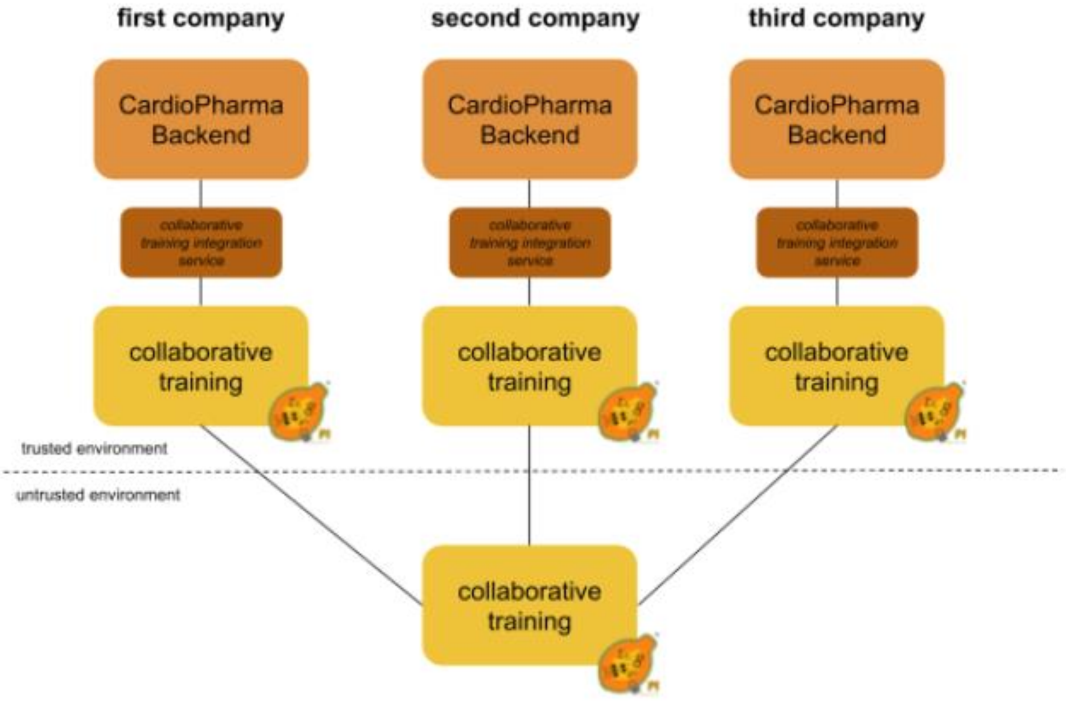


Figure 16 Collaborative training: CardioPharma installations for each participating company

Once the server-side collaborative training component is done with its work (i.e., it has created the collaboratively trained model), this model is redistributed to all the companies that participated in the training session, and it is ready to be used by each single company to recognize stress levels in its workers. To do so, we created a separate backend (in Figure 15, called “Stress app backend”), that retains the collaboratively trained model, and is queried by the Worker mobile app to classify ECG tracks in stress/no-stress classes.

Finally, the worker’s mobile application integrates DS Tool 1, DS Tool 2, and the Privacy Engine.

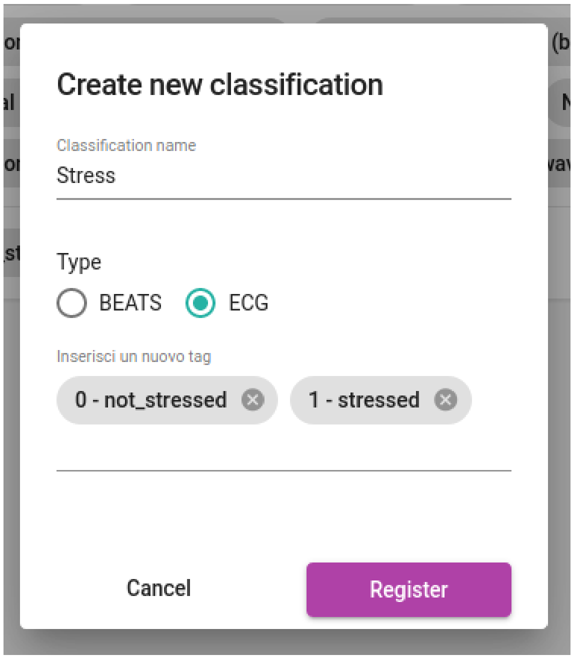
3.2.5 Applications implementation: interfaces

In this section, we show the implemented functionalities through screenshots of the implemented applications.

3.2.5.1 CardioPharma: dataset tagging

In this section, we show the interfaces offered by CardioPharma to: a) allow system administrators to start tagging campaigns; b) allow cardiologists to tag ECG tracks with stress/non-stress labels.

A tagging campaign allows a user (e.g., a cardiologist) to label entries (e.g., either ECG tracks or single ECG beats) with class labels (e.g., “stressed” or “not_stressed”). The creation of a tagging campaign allows a company to have its datasets, created in-house, and useful for building classification models.



Create new classification

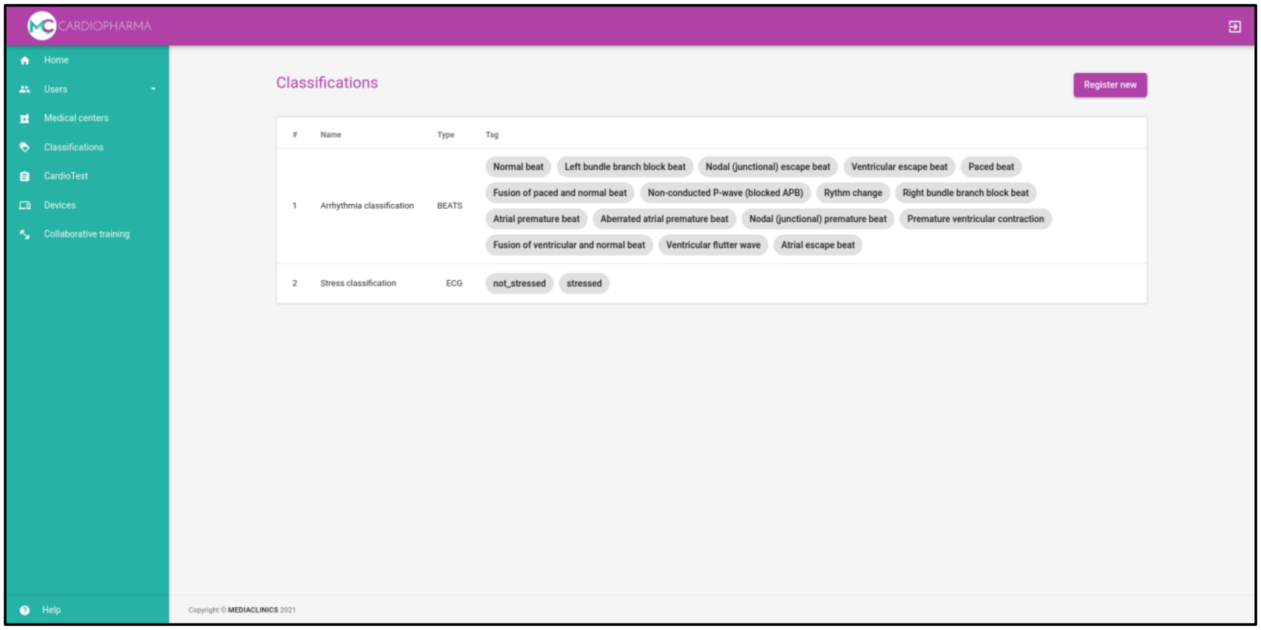
Classification name: Stress

Type: ○ BEATS ● ECG

Inserisci un nuovo tag: 0 - not_stressed, 1 - stressed

Figure 17 CardioPharma: creation of tagging campaign

Figure 17 shows the form the system administrator uses to create a new tagging campaign. Each tagging campaign has a name, a type of associated data indicating what will be the object of tagging (i.e., either a single ECG beat or a whole ECG track) and the possible class labels with which the cardiologist can tag data. The campaign shown in the Figure is exactly the one that can be used to create a training dataset for UC2: whole ECG tracks (and related HRV figures) are tagged by cardiologists as either stress-related or non-stress-related.



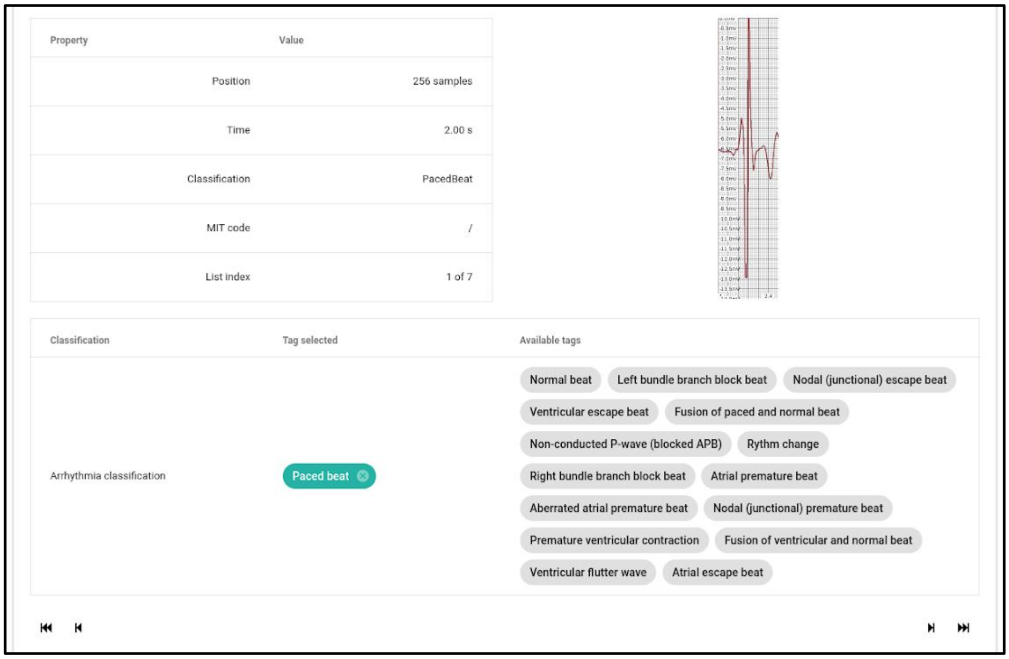
Classifications

#	Name	Type	Tag
1	Arrhythmia classification	BEATS	Normal beat Left bundle branch block beat Nodal (junctional) escape beat Ventricular escape beat Paced beat Fusion of paced and normal beat Non-conducted P-wave (blocked APB) Rythm change Right bundle branch block beat Atrial premature beat Aberrated atrial premature beat Nodal (junctional) premature beat Premature ventricular contraction Fusion of ventricular and normal beat Ventricular flutter wave Atrial escape beat
2	Stress classification	ECG	not_stressed stressed

Figure 18 CardioPharma: visualization of tagging campaigns

Figure 18 shows instead the interface dedicated to the system administrator that visualizes the tagging campaign he created in the past. As an example, the Figure shows:

- a tagging campaign for arrhythmia classification (where each beat can be classified with 16 different arrhythmia classes, to create a dataset which can be used to train a model for UC1);
- a tagging campaign for stress classification (where each ECG track can be classified with either “stressed” or “not stressed” labels, coherently with what UC2 requires).

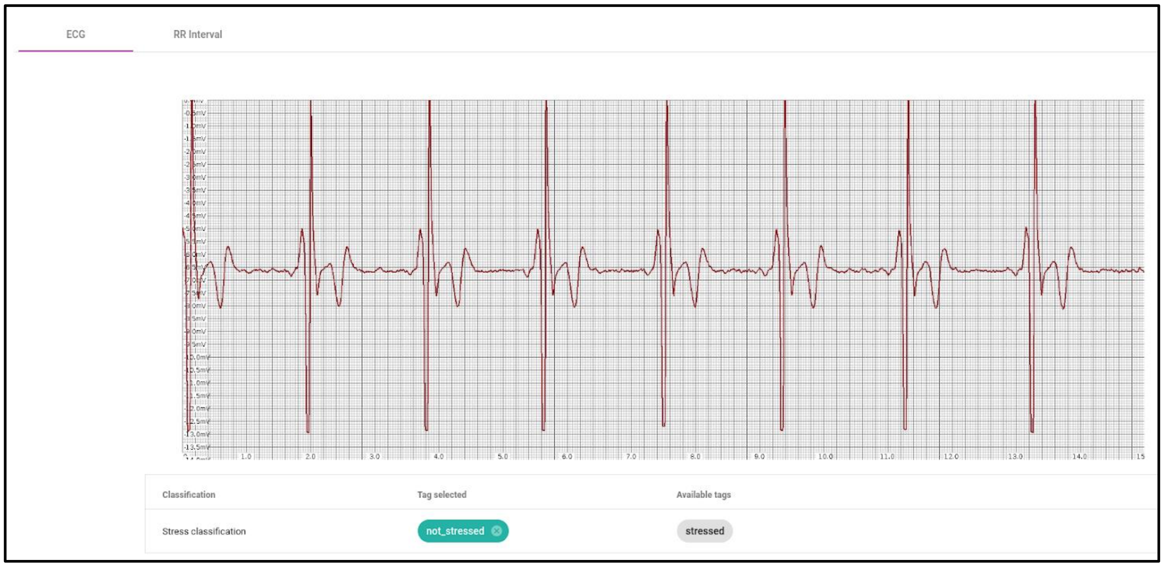


Property	Value
Position	256 samples
Time	2.00 s
Classification	PacedBeat
MIT code	/
List index	1 of 7

Classification	Tag selected	Available tags
Arrhythmia classification	Paced beat ⊗	Normal beat Left bundle branch block beat Nodal (junctional) escape beat Ventricular escape beat Fusion of paced and normal beat Non-conducted P-wave (blocked APB) Rythm change Right bundle branch block beat Atrial premature beat Aberrated atrial premature beat Nodal (junctional) premature beat Premature ventricular contraction Fusion of ventricular and normal beat Ventricular flutter wave Atrial escape beat

Figure 19 CardioPharma: labeling beats with arrhythmia classes

Figure 19 shows the interface used by the cardiologist to tag an entry (in this case, a beat) with one of the allowed labels (in this case, arrhythmia classes).



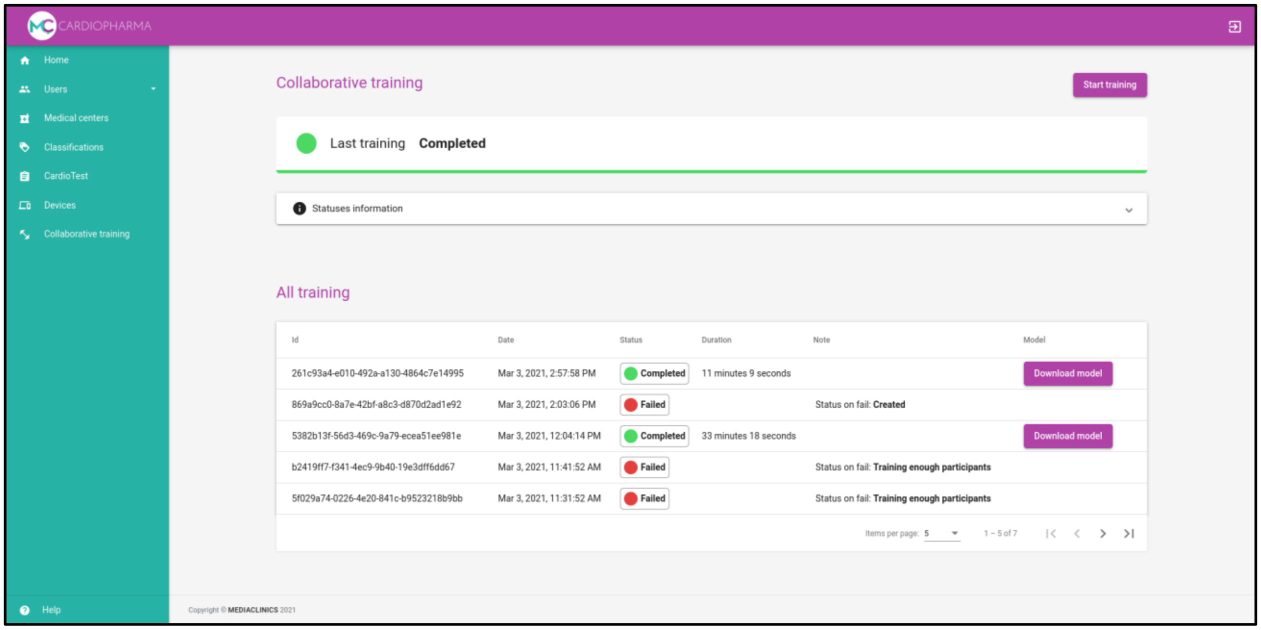
Classification	Tag selected	Available tags
Stress classification	not_stressed ⊗	stressed

Figure 20 CardioPharma: labeling ECG tracks as stressed or not stressed

Similar interfaces are proposed for each labeling campaign; Figure 20 shows how UC2 dataset can be created via a similar interface, where whole ECGs (and related HRV features) can be tagged as either “stressed” or “non-stressed”.

3.2.5.2 CardioPharma: collaborative training

In this section, we show the interfaces offered by the collaborative training dashboard, used to take part into a collaborative training session.



Collaborative training

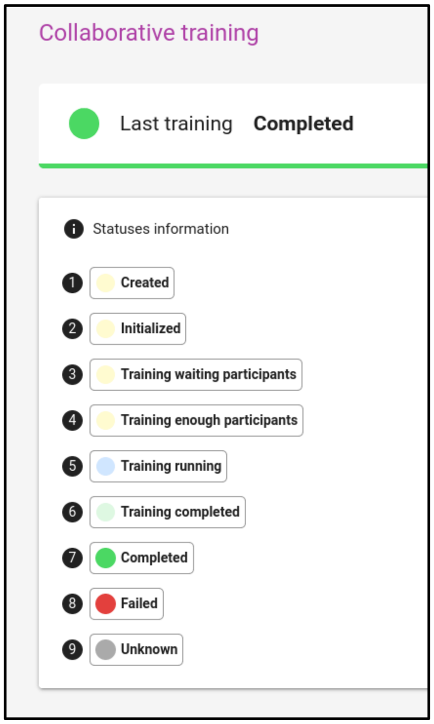
● Last training **Completed**

All training

Id	Date	Status	Duration	Note	Model
261c93a4-e010-492a-a130-4864c7e14995	Mar 3, 2021, 2:57:58 PM	🟢 Completed	11 minutes 9 seconds		Download model
869a9cc0-8a7e-42bf-a8c3-d870d2ad1e92	Mar 3, 2021, 2:03:06 PM	🔴 Failed		Status on fail: **Created**	
5382b13f-56d3-469c-9a79-ecea51ee981e	Mar 3, 2021, 12:04:14 PM	🟢 Completed	33 minutes 18 seconds		Download model
b2419ff7-f341-4ec9-9b40-19e3dff6dd67	Mar 3, 2021, 11:41:52 AM	🔴 Failed		Status on fail: **Training enough participants**	
5f029a74-0226-4e20-841c-b9523218b9bb	Mar 3, 2021, 11:31:52 AM	🔴 Failed		Status on fail: **Training enough participants**	

Figure 21 CardioPharma: collaborative training dashboard

Figure 21 shows the collaborative training dashboard that can be used by the CardioPharma administrator to start new trainings or check the status of the ones started in the past. For the ones that failed, the system reports the reason why they failed (e.g., the training needed more participants to get to the end). For the ones that are completed, instead, it is possible to download the model. This is a useful operation in our case, as the collaboratively trained model will be uploaded to the uStress app backend, and used by uStress to classify workers’ stress in real time.



Collaborative training

● Last training **Completed**

Statuses information

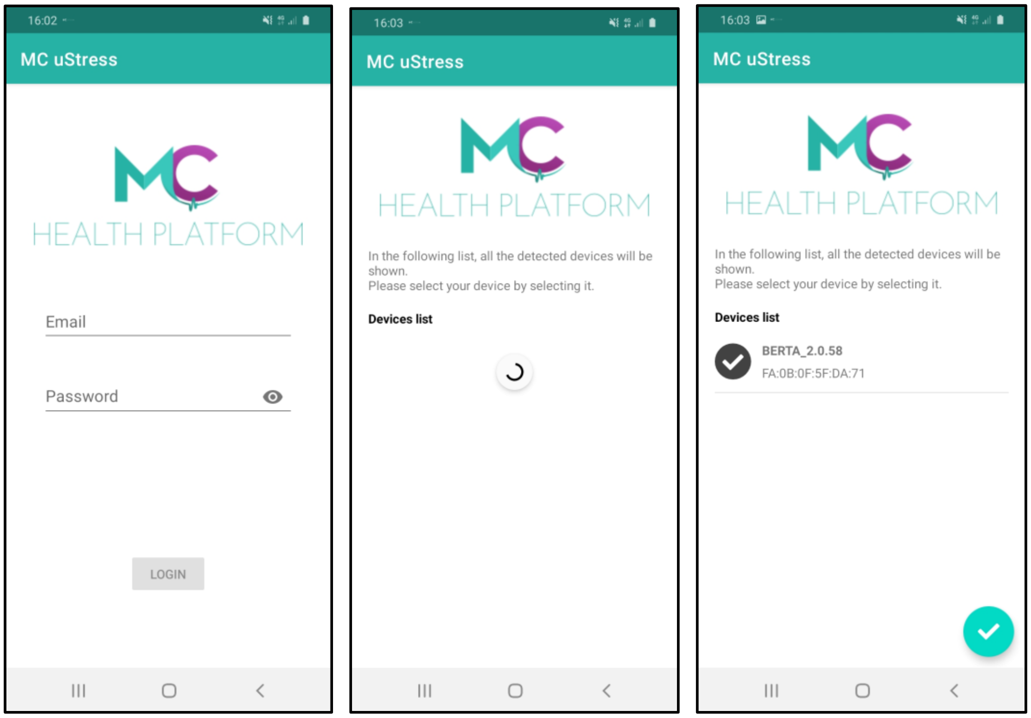
1. ● Created
2. ● Initialized
3. ● Training waiting participants
4. ● Training enough participants
5. ● Training running
6. ● Training completed
7. ● Completed
8. ● Failed
9. ● Unknown

Figure 22 CardioPharma: possible statuses for collaborative training task

Finally, Figure 22 shows all the possible statuses for a collaborative training task.

3.2.5.3 uStress mobile application

In this section, we show the interfaces offered by the worker’s mobile application, called uStress.



In the following list, all the detected devices will be shown.
Please select your device by selecting it.

Devices list

In the following list, all the detected devices will be shown.
Please select your device by selecting it.

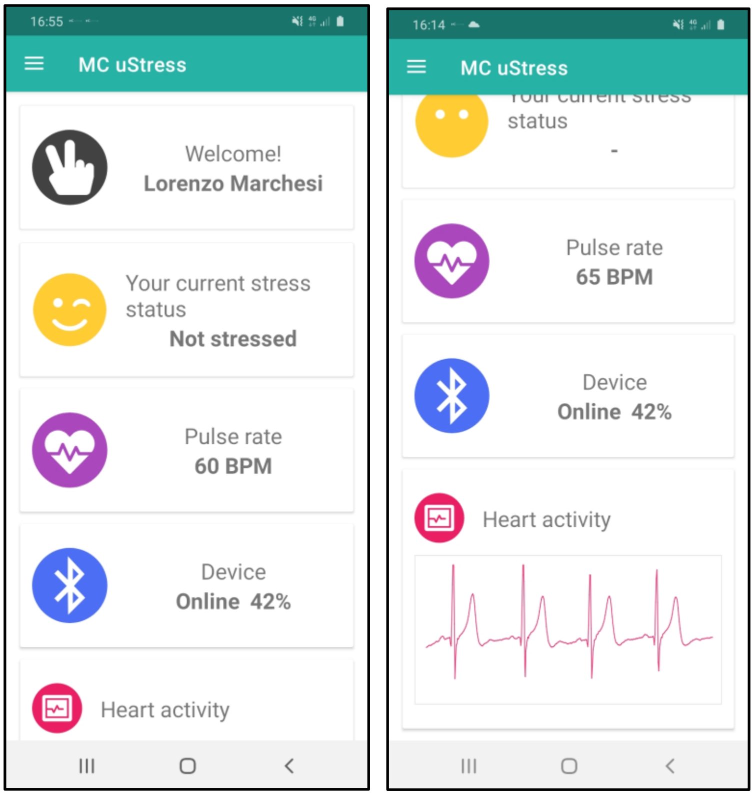
Devices list

BERTA_2.0.58
FA:0B:0F:5F:DA:71

Figure 23 PAPAYA uStress application: login and pairing to the monitoring device

Figure 23 shows the initial screens of the application, where the worker can: a) identify himself via login; b) visualize the list of ECG monitoring devices that are available for pairing; c) pair the ECG monitoring device that was given to him with the application.

From this point on the worker can wear the sensorized T-shirt, that will start acquiring ECG data and send them to the Stress app backend for classification (5 minutes of ECG track every minute). Classification is done using the collaboratively trained model.



16:55

MC uStress

Welcome!
Lorenzo Marchesi

Your current stress status
Not stressed

Pulse rate
60 BPM

Device
Online 42%

Heart activity

16:14

MC uStress

Your current stress status
-

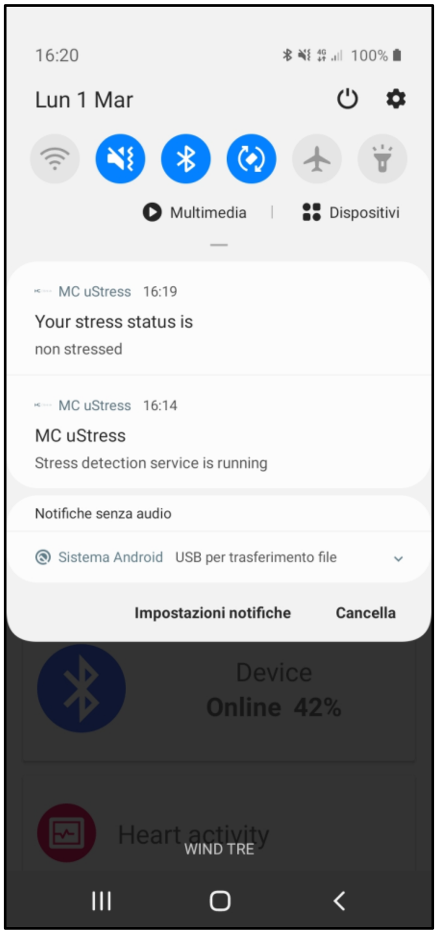
Pulse rate
65 BPM

Device
Online 42%

Heart activity

Figure 24 PAPAYA uStress dashboard

Figure 24 shows the PAPAYA uStress dashboard the worker can visualize at any moment. This dashboard reports a real-time visualization of the worker's heart activity, his pulse rate, the device status (connected or not, battery level) and the **worker's stress status**.



MC uStress 16:19

Your stress status is

non stressed

MC uStress 16:14

MC uStress

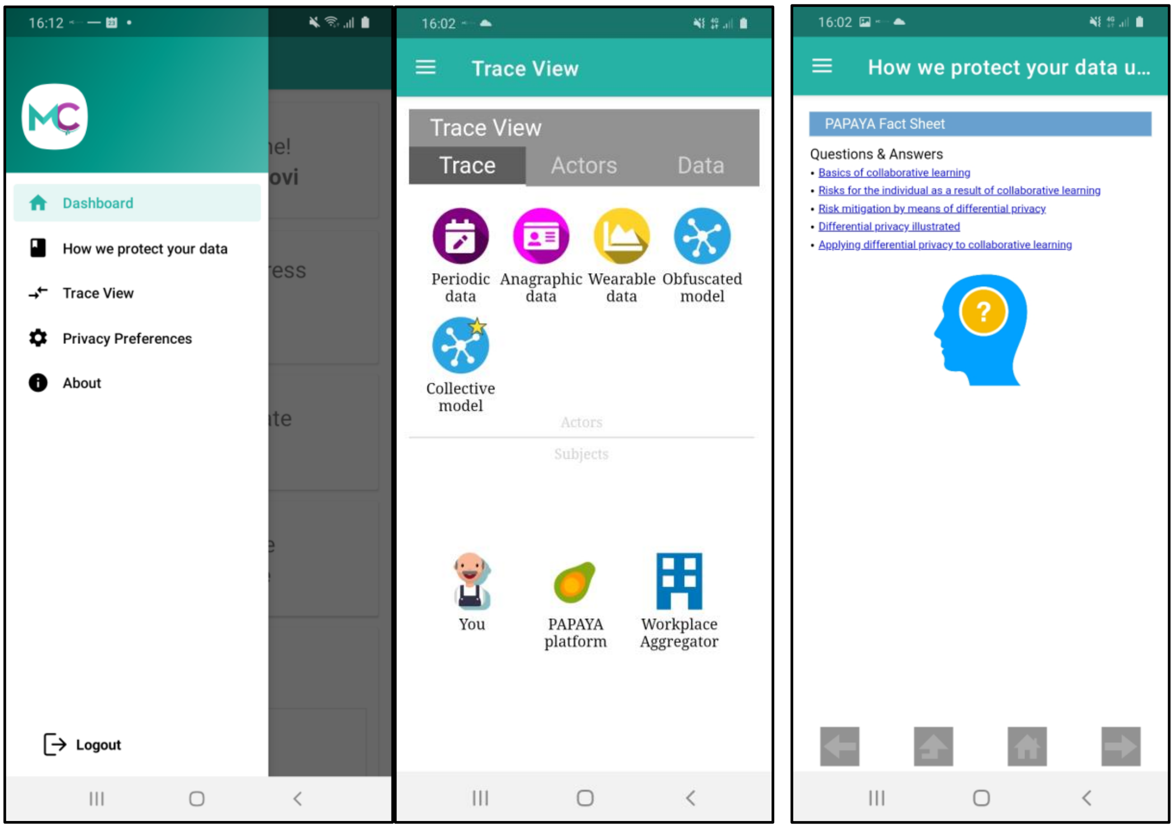
Stress detection service is running

Notifiche senza audio

Sistema Android USB per trasferimento file

Figure 25 PAPAYA uStress application: notification of stress statuses

Notice that, as shown in Figure 25, notifications on stress levels are sent at system level, so that even when the application is in background the worker is aware of his stress level.

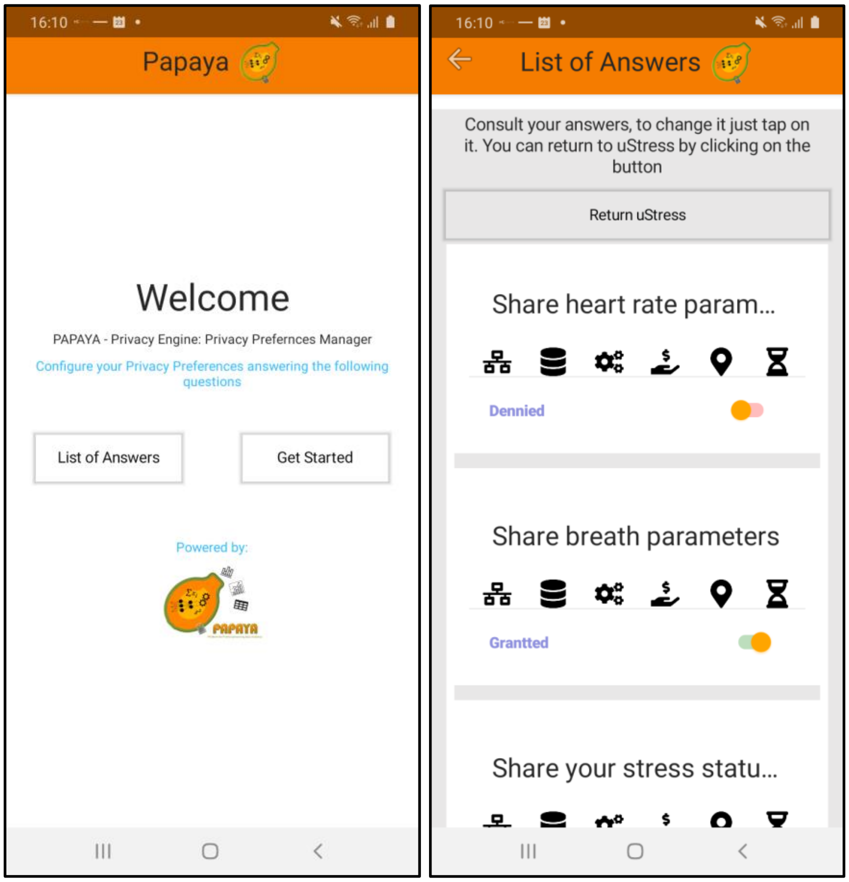


Questions & Answers

- Basics of collaborative learning
- Risks for the individual as a result of collaborative learning
- Risk mitigation by means of differential privacy
- Differential privacy illustrated
- Applying differential privacy to collaborative learning

Figure 26 PAPAYA uStress application: access to DS tools

Notice that (as shown in Figure 26) it is always possible for the worker to access the DS tools to understand how PAPAYA protects his data during processing, and where his data are distributed in trusted and untrusted environments to perform computations. Also, it is possible to access the Privacy Preferences Manager, which is part of the Privacy Engine.



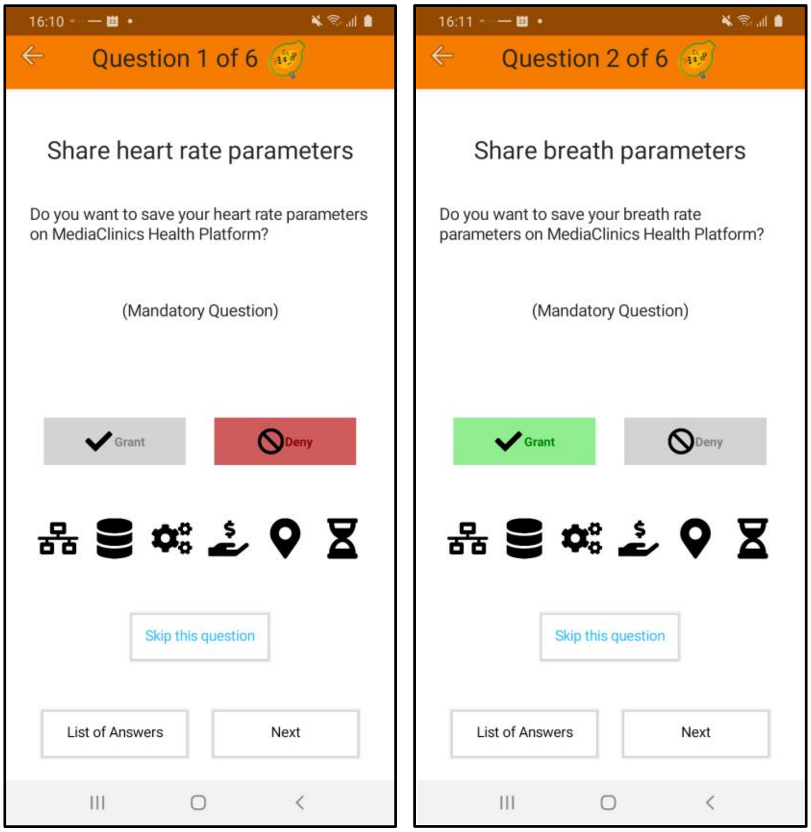
PAPAYA - Privacy Engine: Privacy Prefernces Manager

Configure your Privacy Preferences answering the following questions

Consult your answers, to change it just tap on it. You can return to uStress by clicking on the button

Figure 27 Welcome page and list of preferences from the Privacy Engine

Figure 27 shows the interface of the Privacy Engine, that can be accessed by the worker (from the uStress app menu) to set up his privacy preferences. The first screenshot shows the welcome page of the application: from here the user can visualize the list of preferences, and start with the setup of preferences. The second screenshot shows the current preferences of the worker: for each parameter, the worker can see if sharing that parameter has been allowed or denied.



Share heart rate parameters

Do you want to save your heart rate parameters on MediaClinics Health Platform?

(Mandatory Question)

Share breath parameters

Do you want to save your breath rate parameters on MediaClinics Health Platform?

(Mandatory Question)

Figure 28 Choice of privacy preferences by workers using the Privacy Engine

Figure 28 shows how to set up specific preferences for specific fields. For instance, here the worker decided to deny the processing of the heart rate and allowed the processing of the breath parameters.

3.3 Requirements validation

In this section, we validate the implementation against the requirements specified in Deliverable D2.2.

Table 10 Requirements coverage for Use Case 2

ID	Title	Acceptance Criteria	Validation
UC2.EUR.HCI.1	Inform users about data processing procedures and protection	Information SHOULD be made available to users in the user interface or in another form.	Covered; The data tracing DS tool for UC2 informs about data flows and processing procedures. The DS tool for explaining Differential Privacy for Collaborative Learning informs about the privacy-preserving techniques.

UC2.EUR.HCI.2	Inform user about objectives and incentives for sharing data	The user SHOULD be clearly informed about the objectives and benefits of data sharing when providing their consent. This information should focus primarily on the benefits for the individual, and benefits for the common good should also be mentioned, as this also may be an incentive for some users.	Covered; access to the service is voluntary, as this is a welfare service offered by the company. The user is informed orally about the goals and benefits of the service, and can decide (also via the Privacy Engine) if and how to participate to data sharing
UC2.EUR.HCI.3	Policy options	Consent user interfaces with policy options SHOULD be in place.	Covered; definition and adaptation of data sharing policies is done with the integration with the Privacy Engine, that allows a user to define what to share and when to share it

3.4 Validation by stakeholders

In this section, we validate the implementation by presenting it to the stakeholders.

3.4.1 Evaluation Objectives

To evaluate the data subject tools designed for the medical use cases with stakeholders, we focused on evaluating the data subject tool for explaining differential privacy for UC2. The user interface concepts of the data tracing tool for UC2 is based on the Data Track tool which was already subject to a series of user evaluations and successive improvements that we conducted earlier [15, 16]. For these reasons, it was not subject to this final evaluation with end users.

The results of our evaluation reported in the section are presented in more detail in [27].

The first main research objective of our evaluation has been to explore the suitability of metaphors for differential privacy, including the metaphors that we used in our tool, for effectively communicating the underlying differentially private data analyses to lay users. Metaphors are a means to present new ideas through the use of more familiar ones [17]. In our data subject tool, we focused the on graphical metaphors that are elaborated with short simple accompanying information conveyed as text in its simplest form.

A second research objective has been to analyse the end users' perception of a privacy-preserving data analysis scenario for UC2 and the related data subject tool. In particular, we were interested to analyse how far users value and trust the privacy-preserving data analysis approach using differential privacy and how far they find the information about the privacy-preserving approach to be provided via a data subject tool useful for making informed decisions.

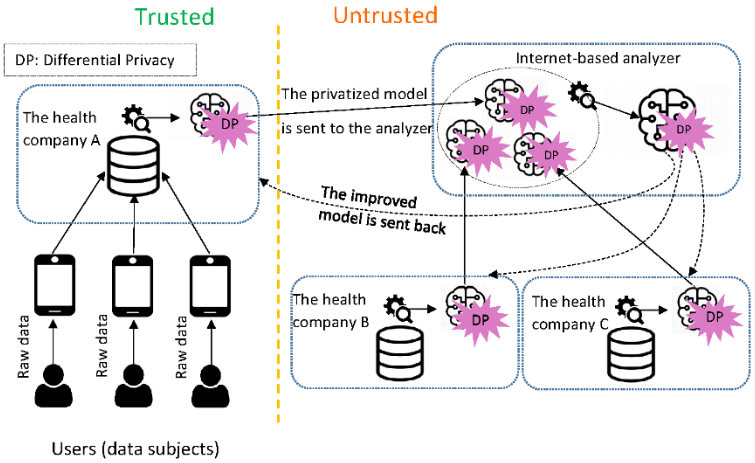


Figure 29 High-level data analysis scenario in UC2 of PAPAYA (shown to our interviewees-see

Figure 29 is a high-level representation of the differentially private data analysis scenario in UC2 of PAPAYA that is an example of centralized differential privacy. In this centralized model in the context of federated learning, the aggregators have access to the actual information of their users who should rely on the trustworthiness of the aggregators. In other types of centralized model, there is one data aggregator. Contrary to centralized models, in local models, the aggregator does not see the actual data of an individual. Note that for addressing our first research objective, we were more generally investigating metaphors for differential privacy both local and centralized differential privacy with one aggregator besides the scenario depicted in Figure 1.

This is also motivated by the layered approach of our DS tool. As shown in the UI displayed on the right side of Figure 26, the start page of the DS tool includes links for explaining the basics of collaborative learning and for illustrating differential privacy with the help of metaphors in general before explaining that differential privacy can also be applied to collaborative learning (see also D3.4 [12]).

3.4.2 Evaluation phases

Our approach to reaching our objectives consists of three phases: 1) metaphor generation, 2) metaphor analytical evaluations based on expert analyses, and 3) metaphor empirical evaluations involving lay users with a focus on metaphors for the UC2 scenario. All three phases are addressing our first research objective of exploring suitable metaphors. The third phase is in addition also addressing our second research objective of analysing the users' perceptions of privacy-preserving data analysis in UC2 and related information to be provided in our data subject tool.

Figure 30 shows a general view of our approach, based on the extended and adapted version of Alty et al.'s framework [17] to fit our objective. Demjaha et al. [18] previously benefited from the framework proposed by Alty et al. [17] to generate and evaluate their explanatory metaphors for E2E encryption.

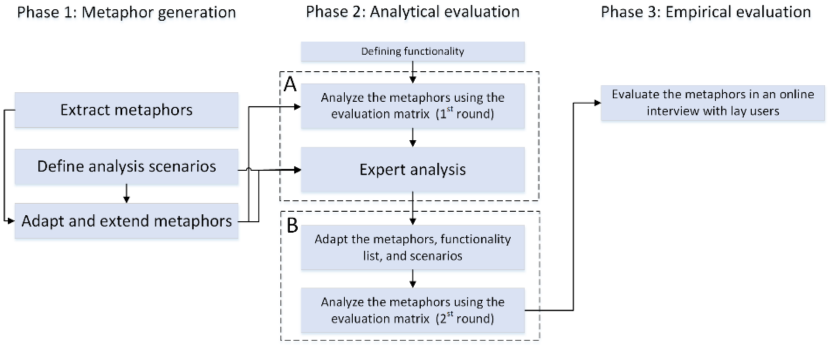

Phase 1: Metaphor generation
Phase 2: Analytical evaluation
Phase 3: Empirical evaluation
Defining functionality
A
Extract metaphors
Analyze the metaphors using the evaluation matrix (1st round)
Evaluate the metaphors in an online interview with lay users
Define analysis scenarios
Expert analysis
Adapt and extend metaphors
B
Adapt the metaphors, functionality list, and scenarios
Analyze the metaphors using the evaluation matrix (2st round)

Figure 30 Our method to reach our objective

To begin with, in the first phase, we reviewed literature and media outlets to see how others conveyed the concept of differential privacy to users using metaphors or analogies. Parts of the metaphors that we derived from this review were also used in our data subject tool. We then used the results of our investigation to extend and adapt the metaphors based on the type of data analysis scenario.

In the second phase, to conduct analytical evaluation, we benefited from the metaphor evaluation matrix in [18] and adapted it that is shown in Appendix 1. However, to evaluate the metaphors using the template table, we needed to identify the general privacy functionality of differentially private data analyses. We made a balance between a functionality list that is detailed enough to cover the main characteristics of differentially private analyses and also sufficiently simple for conducting analytic assessment and finding suitable metaphors. Thus, the list excludes more elaborate features such as post-processing and group privacy. In the results section we present our functionality list. In the second phase, we conducted two rounds of analytical evaluations. After the first round of evaluation, we received feedback from experts, adapted our metaphors, and conducted the second round of evaluation for our adapted metaphors before we evaluate our adapted metaphors empirically in user studies. In total, eight privacy experts both from academia and industry reviewed our materials in step A of phase 2 (see Figure 30), including our description of data analysis scenarios, our original functionality list, the resulted metaphors in phase 1 and the first round of our analytical evaluation.

Finally, in the last phase, we conduct online interviews with lay users. Figure 31 shows the study design of our interviews. The summarised script of the interviews we conducted for UC2 of PAPAYA is provided in Appendix 2, which contains what we conveyed to users and what questions we asked. We conducted a pilot interview and completed ten interviews for the PAPAYA scenario depicted in Figure 1. Participation in the interviews was legitimised by informed consent and the use of a Persona ("Alex", see also Appendix 2) that allowed participants to answer all questions from the perspective of this artificial person allowed us to prevent the collection of any sensitive personal data. We received ethical approval by one of the Ethics Advisor at Karlstad University.

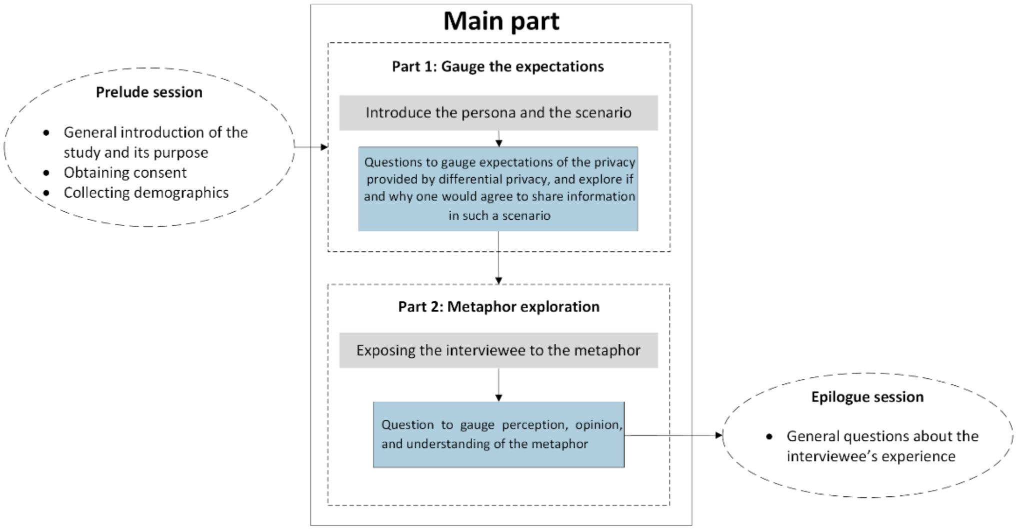

Main part
Part 1: Gauge the expectations
Introduce the persona and the scenario
Questions to gauge expectations of the privacy provided by differential privacy, and explore if and why one would agree to share information in such a scenario
Part 2: Metaphor exploration
Exposing the interviewee to the metaphor
Question to gauge perception, opinion, and understanding of the metaphor
Prelude session
General introduction of the study and its purpose
Obtaining consent
Collecting demographics
Epilogue session
General questions about the interviewee's experience

Figure 31 The study design of the interviews conducted in phase 3

3.4.3 Results

3.4.3.1 Results of Phase 1

3.4.3.1.1 Metaphors derived from reviewing the literature and media outlets.

Warner [19] for the first time proposed randomization of responses by a spinner for improving the reliability of responses to sensitive questions. Our literature review revealed that the spinner metaphor was used by Bullek et al. [20]. The spinner has been also used in media outlets to convey how differential privacy works.[1]

We investigated the media outlets and companies applying differential privacy to see how differential privacy is conveyed to users. We found that differential privacy is conveyed to people using an example of tossing a coin for changing responses to sensitive questions[2], noisy sound waves of radio channels[3], and a noisy portrait[4] from the media outlets. Exploring how companies described differential privacy to their users did not result in any further metaphors.

[1] An example of using spinner by Mark Hansen: https://accuracyandprivacy.substack.com.
[2] Simply Explained: https://www.youtube.com/watch?v=gI0wk1CXlsQ.
[3] National Institute of Standards and Technology: https://www.youtube.com/watch?v=-JRURYTfBXQ.
[4] Nikolas Sartor at Aircloak blog: https://aircloak.com/explaining-differential-privacy.

The metaphors in the forms of a noisy portrait (of the writer Selma Lagerlöf), a variant of the spinner Bullek et al. [10] used, and noisy sound waves of radio channels were also used in our data subject tool.

3.4.3.1.2 Metaphors generated in phase 1

Not all metaphors are necessarily suitable for conveying different differentially private data analyses. A noisy picture may not be suitable for centralized differential privacy because it does not convey that perturbation happens on the aggregate-level. We adapted and extended the metaphor of a noisy picture by adding noise to a picture which is a combination of several portraits to better reflects that noise is not added to the original data collected from users. In addition, randomized response techniques, the coin flip and spinner examples, are only suitable for local differential privacy. We excluded the coin metaphor because we assumed that it would be harder for users to think of a deformed coin that would make it more probable, for example, to have tails rather than heads compared to spinners with different probabilities to convey the trade-off between privacy and security in different cases.

The picker wheel, both variants of noisy pictures, and noisy broadcasts of a radio channel served as the input to our first analytical evaluation (see Appendix 3).

3.4.3.2 Results of Phase 2

3.4.3.2.1 Functionality list

Here we provide the functionality list that we used for our second round of analytical evaluation (step B in Figure 30) that is the adapted list after receiving feedback from experts.

Functionality 1. A differentially private analysis[5] bounds and quantifies the probability of additional privacy risk that any individual would face because of her/his participation in a data analysis.

Functionality 2. The privacy of a differentially private analysis is controlled by tuning a privacy loss parameter.

Functionality 3. The smaller the value of the privacy loss parameter, the better the privacy guarantee for an individual.

Functionality 4. The smaller the value of the privacy loss parameter, on the other hand, the less accurate the results of data analysis are.

Functionality 5. A differentially private analysis randomly perturbs data on an aggregate-level (i.e., the results of the analysis) or individual level (i.e., the input data) dependent on the context.

Functionality 6. The amount of perturbation is controlled by the underlying differentially private analysis.[6]

[5] A differentially private analysis is often called a mechanism.

[6] To have a differentially private data analysis we should know what to perturb and to what extent.

Functionality 7. A differentially private analysis is resistant to privacy attacks based on auxiliary information, i.e., any past, present, and future information that an attacker may have.
Functionality 8. A differentially private analysis does not promise unconditional freedom from privacy risks.[7]

Note that the first feature (F1) in the list can be interpreted in different ways. For example, F1 should convey for the centralized model that the results of a differentially private data analysis do not significantly depend on any particular individual's data so an individual will not be affected, adversely or otherwise, by allowing her data to be used in the analysis. F1 can also be rephrased in terms of plausible deniability. Although a metaphor may not directly convey F1, it may imply different interpretations of F1. Further, to communicate about the underlying differentially private data analysis to lay users we did not focus on the privacy loss parameter but on the role of perturbation in providing privacy and the effects of perturbation on the accuracy of the results. Therefore, if a metaphor conveys that more perturbation leads to better privacy but less accuracy, we assume it covers F3 and F4.

3.4.3.2.2 The results of the 1st round of analytical evaluation and expert analysis

As a result, of the expert analysis, we excluded the metaphor of noisy sound waves of a radio channel because it suffers from a highly undesirable feature (conceptual baggage). The metaphor implies that the original data collected by the aggregator can be heard by anyone who listens to the radio channel at the right frequencies. Nonetheless, those who should receive the sound waves without noise (those who should have access to the original data) are either the data subjects or trusted data aggregators. Further, the metaphor conveys that anyone who receives the data, i.e., listen to the radio channel can decide on the amount of noise that should be added to the data. However, an adversary, as an example, does not decide on how much noise should be added to the results of an analysis. In addition, our experts advised that if we have public information, e.g., radio broadcast through an FM radio channel, it does not make sense to apply differentially private mechanisms. Despite its problems, this metaphor, as also confirmed by our experts, is suitable if we want to highlight the importance of tuning the privacy loss parameter, for example, by an aggregator rather than if want to convey that there is a trade-off in every differentially private system.

Based on our experts' feedback we confirm that the spinner metaphor is only suitable for the local differential privacy. We adapted and extended our preliminary spinner metaphor (see Appendix 4**)** to better communicate F3, F4, and F6 for local differentially private data analysis scenarios. The spinner metaphor may suffer from a rather undesirable feature if not properly introduced to users. Bullek et al. [20] reported that some participants preferred the most truthful spinner because they associated the perturbation, they made on their answers to lying to the entity asking questions. Should the spinner be used to communicate the underlying mechanism to users in local models we recommend conveying a message to users that they do not perturb their data

[7] Note that any useful data analysis carries the risks that it will reveal information about individuals.

but, for example, a device on their phone does the perturbation before sharing the data with a remote server.

Initially, for the metaphors based on noisy figures, we used the portraits of famous people. In the media outlets, the noisy portraits of famous people were used as well. However, based on our experts' comments, we adapted the metaphor and avoided using photos of famous and well-known people, although it does not solve the undesirable features of these metaphors. Unless distorted with a high amount of noise, a noisy picture may still reveal the identity of the person in the picture and specific characteristics of him/her. The metaphors based on the noisy pictures do not necessarily cover F1 and F7. Based on our experts' feedback we confirm that a single noisy picture (see Appendix 4) is only suitable for the local models and a noisy picture that is a combination of several other pictures, as depicted in Figure 32, is more suitable for centralized models. However, the centralized model in UC2 of PAPAYA is a specific case in which the parameters of a model are distorted rather than, for example, the results of doing simple statistical calculations on the input data, i.e., sum, average, median, etc. Therefore, instead of using a noisy picture, we made a new metaphor, and we used a distorted brain figure as a metaphor of a differentially private trained model as depicted in Figure 33 to convey about differential privacy in UC2 of PAPAYA that we then tested in our interviews.

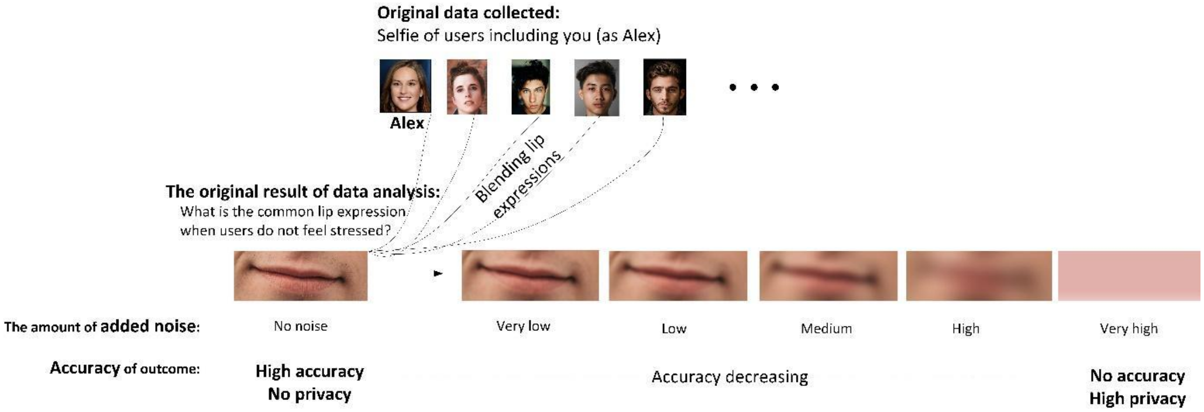


Figure 32 Metaphor to convey about differential privacy for centralized models

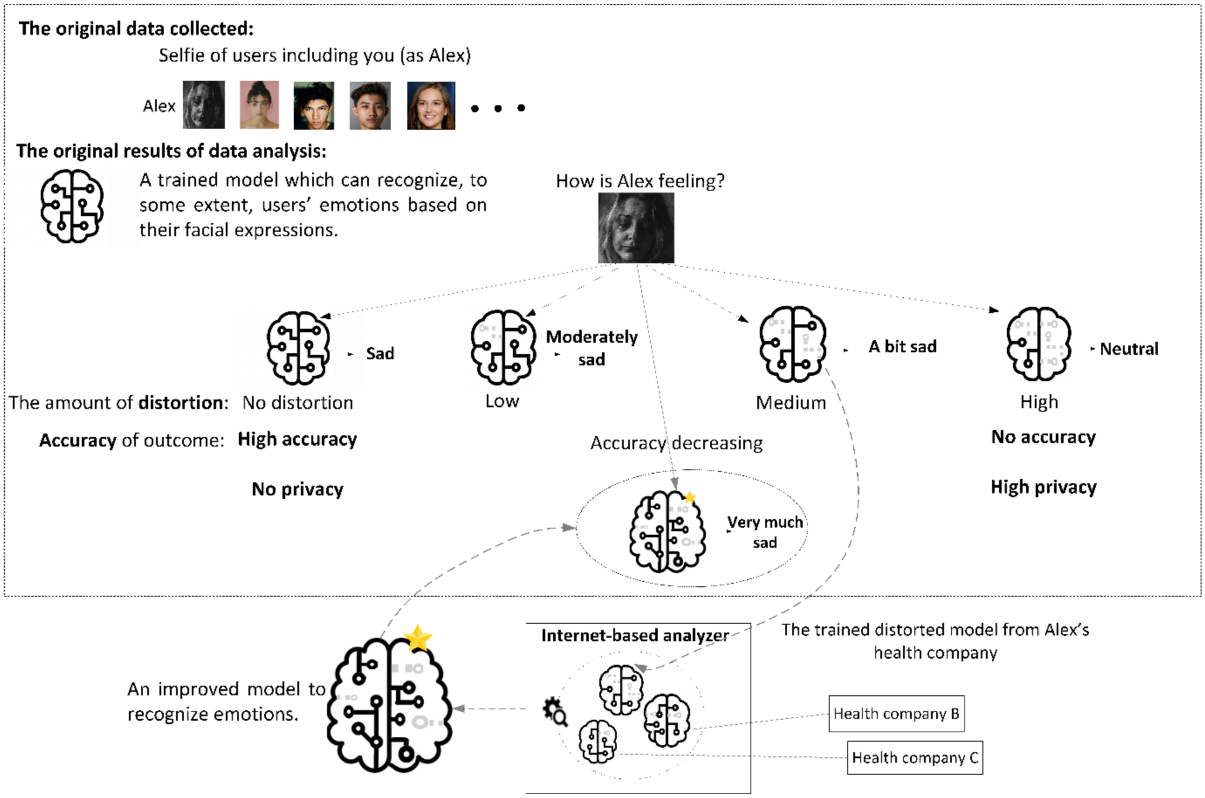

The original data collected:
Selfie of users including you (as Alex)
Alex
The original results of data analysis:
A trained model which can recognize, to some extent, users' emotions based on their facial expressions.
How is Alex feeling?
Sad
Moderately sad
A bit sad
Neutral
The amount of distortion:
No distortion
Low
Medium
High
Accuracy of outcome:
High accuracy
Accuracy decreasing
No accuracy
No privacy
High privacy
Very much sad
An improved model to recognize emotions.
Internet-based analyzer
The trained distorted model from Alex's health company
Health company B
Health company C

Figure 33 Metaphor to convey about differential privacy for centralized models in the context of federated learning

3.4.3.2.3 The results of the 2nd round of analytical evaluation

Table 11 shows whether each of our adapted metaphors (shown in Figure 32, Figure 33, and Appendix 4.) conveys or implies the features in the functionality list, although it is subjected to be validated by users studies, and shows for which scenario it can be used. What we assume could be understood from a metaphor is different from what lay users grasp. The Y (Yes) in Table 11 means that the metaphor has the potential to convey or implies the feature without clarification by further information, for example, in the form of an accompanying text. Features F3 to F6 are conveyed by all four metaphors. F1 and F8 are implied by the spinner metaphor. F8 is implied by the other metaphors as well. The noisy picture metaphor for the local model does not cover F1 and F7. The noisy combined picture metaphor may convey F1 and F7. However, whether it really covers F1 and F7 is pretty much dependent on the combination of all pictures selected for that metaphor. In addition, users' understanding and perception of, for example, how much the aggregate-level picture might be revealing and if and how the added noise can circumvent privacy leakage from a combined/aggregated picture play a significant role. The distorted brain metaphor shown in Figure 32 is very abstract and whether it conveys or implies F1 and F7 is very much dependent on what users know or understand from the concept of the model.

Table 11 Features in functionality list potentially could be covered by the metaphors.

Metaphor /feature	F1	F3	F4	F5	F6	F7	F8	Context
Spinner	Y	Y	Y	Y	Y	Y	Y	Local DP
Noisy picture - single	N	Y	Y	Y	Y	N	Y	Local DP
Noisy picture - combined	Y	Y	Y	Y	Y	Y	Y	Aggregate-level DP
Distorted brain model	Y	Y	Y	Y	Y	Y	Y	Aggregate-level DP (federated learning)

3.4.3.3 Results of Phase 3

In this section, we summarise the results from our interviews based on notes that were taken and evaluated. A more detailed thematic analysis of interview transcriptions will be conducted in our future work.

3.4.3.3.1 Demographics

We announced our online interviews on the Prolific platform and recruited 10 participants for the PAPAYA scenario. We used the pre-screening filters on Prolific to exclude people who had computer science, computing (IT), or engineering as their fields of study. We also used the filters to recruit people whose countries of residence were European Union countries, European Economic Area countries, the UK, or Switzerland. Two of the participants were females and the rest were males. Our participants were quite young with eight of them in the 18-25 age group and two in the 26-35 age group. Our participants had a variety of fields of study or occupation ranging from political science, English language studies, and pastry making and baking to being a cook and a nail technician. To make sure that we interviewed lay users who do not have any knowledge about differential privacy and are not generally knowledgeable about how privacy protection techniques work we asked a question at the beginning of our interview to gauge interviewees' previous knowledge of privacy techniques and differential privacy. The results show that none of the interviewees previously heard about the term differential privacy and none of them was knowledgeable about privacy protection techniques although a few of them could name some technologies that people could use to protect their privacy, for example, VPN, adblockers, Ghostry browser extension tool, and encryption.

3.4.3.3.2 Factors playing a role in users' decisions to share their data

It seems that the mere presence of a privacy technique regardless of how it works can persuade most of the users to share their data in the scenario depicted in Figure 29. In the first part of the interview, after receiving information about the data analysis scenario and being informed about the presence of a privacy technique to protect users' privacy, the majority of our participants voiced their desire to share their data on behalf of Alex (the persona they role-played in the study) and they mentioned that they would not do it if there was no privacy mechanism involved. Besides the presence of a privacy technique to protect users' privacy, the type of data requested to be shared, the trustworthiness and the reputation of the company requesting the information, and the purposes for which the information is used, i.e. beneficial for the individual and common public played a role in the interviewees' decisions to agree to share their data. It was important for the participants that their data would not be used for commercial purposes while they were less concerned if used for the common good. All the interviewees understood that Alex would benefit from agreeing to share her data to be analysed in the way we described in the scenario by receiving better recommendations to cope with stress. This might also have contributed a lot to their positive attitudes to share their data in this scenario. Our participants also referred to some risks for Alex if she shared her data but it did not outweigh the benefits in their opinions as the majority of them decided to share the data. Interestingly, one participant mentioned that he/she would even be more comfortable if company A was collaborating with some medical associations to be sure about the quality of recommendations she/he would receive to cope with stress. Also, a few participants considered Alex's situation and mentioned that in stressful conditions you probably would not think of data privacy risks and you would like to receive some help to cope with the situation.

The main obstacle for participants who refused to agree to share their data (on behalf of Alex) was the presence of other parties involved in the scenario including the PAPAYA platform and the other companies (companies B and C in Figure 29) and scepticism about how the privacy technique applied would work and protect their privacy.

After receiving more information about how the mechanism worked (see Figure 33 and its related description in Appendix 2) some participants changed the decision they previously made about sharing their data to be analysed in the scenario. The information about distortion made some interviewees more willing to share their data or more confident about the decision they previously made to share their data. However, on the other hand, not the way the mechanism worked but the use of selfies in the example we described caused a few interviews to change their decisions from a previous yes for sharing their data to a new no. One participant said that it was not convincing how sending selfies would help the app to provide recommendations for coping with stress. Also, a few participants voiced their concerns about the exact level of distortion when they were asked if they wanted to review and change the decision they made previously. They wanted to know what amount of distortion would protect them properly and they were concerned about the lack of accuracy.

3.4.3.3.3 Understanding of privacy functionality

First, the concept of the model was not well known/understood by the participants based on the short description we provided to them. Apart from the participants who could not describe a model in their own language, others described the model as “a program”, “deep learning AI to collect as much data as it can to get the most optimal result for users”, “selfies users send”, “model of different face scans”, and “what the AI makes out of the input I give to it, a pattern”. Interestingly, one participant also associated the brain icon in Figure 6 with the representation of what goes on in Alex’s mind, given the fact that the model (represented as a brain) would tell how an individual feel. The lack of understanding about what a model is could have contributed to their misunderstanding about the mechanism. Some of our participants referred to distorted selfies (instead of a distorted model) when they were answering the interview questions about their understanding of distortion. In the description of how the mechanism would work, we conveyed that a model learns from its inputs which, in this case, were users’ selfies. Also, in Figure 6, Alex’s selfie is shown as one example of an input to the trained model to see what it predicts. In addition, it was not conveyed how distortion would happen apart from saying that the information the model has learned from its inputs would be distorted. We assume that all of these facts could have contributed to users’ misconception about distorting selfies instead of the model.

It was understandable and acceptable for our participants that distortion could help to protect and enhance users’ privacy. Participants believed that distortion could avoid, for example, the internet-based analyser from inferring if their selfies were also used to train a model and the hackers from inferring their actual stress-related information. However, not all participants believed that in general distorting a model was required to protect their privacy because they thought that a model would not probably leak (any important) information as long as there was no personally identifying information such as names, addresses. Also, not all participants believed that the inclusion or exclusion of their data (selfies) would change the (undistorted) model. Nonetheless, those who mentioned that an undistorted model could leak information about them believed that accessing a distorted model would prevent an entity from inferring their actual stress-related information, as long as the distortion process was not reversible. The participants believed that there would always be some remaining privacy risks even with the protection mechanisms applied.

Participants all understood that there is a trade-off between accuracy and privacy and privacy would be better protected with more distortion. Participants had different opinions about the amount of distortion which should be applied. While some of them preferred to have no distortion or low level of distortion and supported their decisions by highlighting the objective of making improvements, others selected the medium amount of distortion to both protect the privacy and have some level of accuracy.

In summary, the metaphor depicted in Figure 33 (and its description, see Appendix 4) directly conveys the privacy functionality F3, F4, F6, and F8 and users understand them. However, not all will be able to infer F1, F5, and F7. People may not well understand/know the concept of the model and as a consequence, they may be confused that the original data get distorted instead of the model trained based on the input data. Therefore, based on Figure 33 not all users will understand F5 functionality without further clarification. Greying out some parts of the brain model was very abstract, based on our participants' comments, and they required more information, based on specific examples, on what is exactly distorted and how. Such information could also help to convey F5. In addition, our results show that it is not generally easy for everyone to infer F1 but more concrete examples especially based on the exceptionality of one person in a population can help to convey the role of distortion in limiting the effects of an individual's data on the distorted results. The description of the metaphor should also directly communicate F7; otherwise, it would not be easy for people to infer it.

3.4.3.3.4 Information required about the underlying mechanism

Our participants believed that in general understandable and easy to grasp information about the underlying privacy mechanisms used in a system is helpful for them to make their data sharing decisions and it can improve their trust. When asked if they wanted to know more about the privacy mechanism in the scenario our participants voiced their desire to have more information on how the mechanism would protect their data, and to what extent it would be effective to avoid the other parties involved access their data or track them. They specifically wanted to know if the mechanism could help them be anonymous and if information such as names and addresses would be protected. However, they wanted such information to be understandable for “normal people”.

The participants believed that they got the general idea from Figure 33 and it was easy to read and understandable although the results showed some misconceptions about the model and what would be distorted that we reported previously. The interviewees suggested some ways to improve it based on what they thought was missing or not clearly described. They believed that distortion was shown in a very abstract way and greying out parts of the brain model was not accurate enough in their opinion to describe distortion. They needed more concrete example showing how distortion happens, what distortion means in this context, and what exactly is being distorted. For example, they mentioned that what is sent to the internet-based analyzer is not clear for them, knowing the fact that users’ selfies were not shared. They also made a few comments for the information they required to know which were related to the exact data analysis scenario apart from its privacy mechanism. For example, they wanted to know how much it would take to receive the improved model from the time they shared their data with the health company. They also wanted to know if the system could detect whether users were faking their moods.

Our participants did not have any preconception about how the mechanism would work which could be explained because they never heard about differential privacy and they did not know any other privacy protection techniques. Not having preconceptions, they were not surprised by the description provided to them and did not find anything unexpected. Although they expressed their unwillingness to have technical and mathematical information about the underlying mechanism, they appreciated the possibility of having access to more understandable information about the mechanism including the remaining privacy risks in easy language. However, they mentioned that they would not necessarily read it.

3.4.3.3.5 Factors playing a role in users' trust

We asked our interviewees if they, role-playing as Alex, would be concerned to have their actual data be analysed by the internet-based analyser in Figure 29. The majority of the interviewees were concerned about it and referred to several countermeasures which could mitigate their concerns. Transparency of how data would be protected and the flow of data, not requesting too much of data and not requesting sensitive types of data, providing pieces of evidence which could assure them about the safety and suitability of collaborating with the internet-based analyser, and having good purposes for processing the data by the analyser were the mitigating factors mentioned by our interviewees. Interestingly some interviewees did not consider health-related information such as heart rate and sleep cycles as sensitive information and they were more concerned if directly personally identifying information such as names, addresses, and birthdates were shared and processed. Our participants were concerned that the internet-based analyser would use their data for advertising and they would be the target of, for instance, stress-reduction products because they as the customers of health companies in Figure 29 might suffer from stress. The participants wanted to be assured that their data would not be used for other purposes than to promote people's health conditions and improve the technologies involved to help users have better health conditions.

We explicitly requested our participants to elaborate on the factors which would play a role for them to trust the privacy mechanism in the scenario (before revealing more details about how it would work) to protect their privacy. Our participants' responses revealed that the reputation of the company that uses the privacy mechanism and whether they trust that company affects how they trust the mechanism and the protection it provides to them. Further, transparency about the underlying mechanism could improve their trust in the system to protect their data. Also, the reputation of the mechanism itself, for example, in the media outlets, to what extent it has been used by others, if it is standardized, and if it has been confirmed by experts are important for users. The purpose of data analysis also plays a role in users' trust in the privacy mechanism applied. A few participants mentioned that as long as the purpose of data analysis is to make things better, i.e. improve the app to detect stressful conditions, and there is no personally identifying information involved then they would trust the privacy mechanism to protect their privacy.

After being exposed to the metaphor for the differentially private mechanism in the data analysis scenario, we ask our participants if they would trust the mechanism in general to protect their privacy. Their answers revealed that the amount of distortion, the trade-off between accuracy and privacy, and analysing data from different users influence users' trust in a differentially private mechanism to protect their privacy. To trust a differentially private system, our participants wanted to know about the amount of distortion applied and its effects on the privacy of their data and the accuracy of the results of the analysis. Also, they were concerned that distorted data over time could lead to more false results.

3.4.4 Conclusions

Main conclusions in relation to our two research objectives are summarised below.

3.4.4.1 Suitability of metaphors for data subject tool

The analytical evaluation in phase 2 has shown that the metaphors used in our data subject tool in the forms of a noisy portrait, of a variant of the spinner by Bullek et al. [20], and of noisy sound waves of radio channels are not directly suitable for conveying the functionality of differential privacy for federated learning as used in UC2. The metaphor of a distorted brain could convey this functionality. Our first empirical evaluation results with lay users from phase 3 showed that our lay users could understand that distortion would help to protect their privacy and that there was a trade-off between accuracy and privacy based on the amount of distortion. However, our results have also revealed challenges for some lay users to fully comprehend what the metaphor tries to convey. The respective mental models of users and possible improvements need to be followed up by our future research.

Nonetheless, in our data subject tool, we reduced the complexity by explaining collaborative learning and differential privacy in general first separately. For conveying core functionalities of differential privacy in general, including the privacy-accuracy trade-off and the approach for protecting privacy by perturbing the data in a controlled manner, most of the metaphors used in our data subject tool can still work well and thus fulfil this purpose with the following exceptions or modifications: First of all, we recommend excluding the noisy radio channel metaphor due to its conceptual baggage. Secondly, for the noisy picture metaphor we recommend not taking a photo of a famous person and preferably we recommend using the metaphor of a noisy combined picture, which better reflects that noise is added to data aggregates.

When clicking the link “Applying differential privacy to collaborative learning“ on the start page of the data subject tool (see Figure 30), the user interface informs the user only on a high level that the data sets sent to the analyser are differentially private and that the analyser will be unable to learn whether an individual contributes to the data of an organisation, while the organisation still benefits from a larger variety of predictions.

We recommend improving this information by conveying the functionality of differential privacy in the context of federated learning with suitable metaphors such as metaphors based on the distorted brain model shown in Figure 33.

In the next subsection, we also summarise what information should be accompanied by a metaphoric presentation in the form of text, in its simplest representation, or should be conveyed in policy notices for enhancing end user trust and acceptance.

3.4.4.2 Transparency on differential privacy for promoting trust and acceptance

Our first results show that transparency of the underlying privacy mechanism in UC2 can both increase users' trust in the system to protect their privacy and can mitigate their concerns regarding data sharing and third-party access in the scenario. Based on our results, we discuss below what type of information should be provided on a top and on a second layer of multi-layered privacy policies for enhancing end user trust and acceptance of differentially private collaborative learning:

On a top policy layer, based on our empirical evaluation results with lay users, we suggest that users should be informed about: (a) the mere presence of the mechanism to protect their data without further details about the mechanism, (b) the reputation of the company that uses the privacy mechanism (which affects users' trust in the company and consequently users' trust in the mechanism to protect their data), (c) the reputation of the mechanism itself, for example, in the media outlets, (d) to what extent it has been used by others, (e) if it is standardized, and (d) if it has been confirmed by experts. Also, the purpose of data analysis impacts the users' trust in the mechanism and the protection it provides to them which should be prominently conveyed to users.

Our results showed that just information about the mere presence of a privacy technique and functionality to protect users' data in the data analysis scenario for UC2 is seemingly enough for persuading most of the users to use the system and share their data. However, the possibility of accessing the information on how the mechanism would protect users' data and to what extent it would be effective to avoid other parties involved in the data analysis scenario to access their data or track them back can further help users trust the system and can persuade more users to share their data.

Therefore, on the second policy layer, users could be provided with more high-level information about how the underlying mechanism protects their data by using perturbation and its effects on accuracy and privacy. Our results showed that the information about the presence of distortion could make people more willing to share their data and make their decisions with confidence. Their trust in a differentially private mechanism to protect their privacy will also likely be influenced by the amount of distortion, the trade-off between accuracy and privacy, and whether the analysis uses the data from different users. In particular, users would like to know about the amount of distortion applied, whether the amount of distortion is enough to protect their privacy and its effects on the accuracy of the results of the analysis.

4 Conclusions

This deliverable presented the outcome of the Task T5.1 (*"Validation through eHealth UC"*), in which the technologies and tools provided by the PAPAYA framework have been integrated and validated with two eHealth solutions, namely,

- a tool for performing arrhythmia detection in patients (presented in UC1);
- a tool for performing stress detection in employees (presented in UC2).

The two use cases, which were presented in Deliverable D2.1, stem from real-world scenarios where there is the need of performing computation on the cloud on sensitive data (e.g., ECG data) to extract some analytics (i.e., arrhythmia classes for UC1, stress status for UC2).

The validation phase conducted during the project and reported in this deliverable proved that:

- the **integration** with the PAPAYA framework and technologies is accessible and easy to perform, even with pre-existing tools (like CardioPharma, a software produced by MediaClinics and used in UC1). Indeed, having REST API as an entrypoint for the PAPAYA components, the integration with them is totally compatible with the integration of any Web-based microservice, hiding the complexity of PETs behind the PAPAYA component and leaving the programmer performing the integration (in this case, MediaClinics software engineers) completely free of the burden of learning how to apply a PET by hand;
- the **validity** of the offered solution (i.e., the PAPAYA platform and the DS tools) allows companies to enrich their offer with privacy-preserving solutions, and this: i) requires low effort; ii) guarantees high impact on perceived data protection level and high usability.

The work conducted in this task relates to the one reported in Deliverable D5.2 (the twin of D5.1 where mobile and phone usage use cases are validated) and Deliverable D5.3 (which reports recommendations and refinements for the PAPAYA platform). The combination of these three deliverables completes the validation of the PAPAYA framework.

References

[1] Sharma, N., & Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Computer methods and programs in biomedicine, 108(3), 1287-1301.

[2] Can, Y. S., Arnrich, B., & Ersoy, C. (2019). Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. Journal of biomedical informatics, 92, 103139.

[3] He, J., Li, K., Liao, X., Zhang, P., & Jiang, N. (2019). Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal. IEEE Access, 7, 42710-42717.

[4] Cinaz, B., Arnrich, B., La Marca, R., & Tröster, G. (2013). Monitoring of mental workload levels during an everyday life office-work scenario. Personal and ubiquitous computing, 17(2), 229-239.

[5] Yang, H. K., Lee, J. W., Lee, K. H., Lee, Y. J., Kim, K. S., Choi, H. J., & Kim, D. J. (2008, August). Application for the wearable heart activity monitoring system: analysis of the autonomic function of HRV. In 2008 30th annual international conference of the ieee engineering in medicine and biology society (pp. 1258-1261). IEEE.

[6] Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., & Mozos, O. M. (2015, June). Stress detection using wearable physiological sensors. In International work-conference on the interplay between natural and artificial computation (pp. 526-532). Springer, Cham.

[7] Gjoreski, M., Gjoreski, H., Luštrek, M., & Gams, M. (2016, September). Continuous stress detection using a wrist device: in laboratory and real life. In proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct (pp. 1185-1193).

[8] Bakker, J., Holenderski, L., Kocielnik, R., Pechenizkiy, M., & Sidorova, N. (2012, January). Stess@ work: From measuring stress to its understanding, prediction and handling with personalized coaching. In Proceedings of the 2nd ACM SIGHIT International health informatics symposium (pp. 673-678).

[9] Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014, November). The swell knowledge work dataset for stress and user modeling research. In Proceedings of the 16th international conference on multimodal interaction (pp. 291-298).

[10] Wikipedia contributors. (2021, February 25). Electrocardiography. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:15, February 25, 2021, from https://en.wikipedia.org/w/index.php?title=Electrocardiography&oldid=1008866741

[11] Wikipedia contributors. (2021, February 5). Heart rate variability. In Wikipedia, The Free Encyclopedia. Retrieved 13:46, February 26, 2021, from https://en.wikipedia.org/w/index.php?title=Heart_rate_variability&oldid=1005015618

[12] Fischer-Hübner, S. (Ed.). “Transparent Privacy preserving Data Analytics”. PAPAYA Deliverable D3.4, April 2020.

[13] Alaqra, A.S., Ciceri, E., Fischer-Hübner, S., Kane, B., Mosconi, M., & Vicini, S. (2020, July). Using PAPAYA for eHealth-Use Case Analysis and Requirements. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 437-442). IEEE. https://www.diva-portal.org/smash/get/diva2:1543701/FULLTEXT01.pdf

[14] Alaqra A.S., Kane B., & Fischer-Hübner S. Machine Learning Based Analysis of Encrypted Medical Data In The Cloud: A Qualitative Study of Expert Stakeholders’ Perspectives. Journal of Medical Internet Research. JMIR Human Factors. 07/06/2021:21810 (forthcoming/in press). https://preprints.jmir.org/preprint/21810/accepted

[15] Angulo, J., Fischer-Hübner, S., Pulls, T., & Wästlund, E. (2015, April). Usable transparency with the data track: a tool for visualizing data disclosures. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (pp. 1803-1808).

[16] Fischer-Hübner, S., Angulo, J., Karegar, F., & Pulls, T. (2016, July). Transparency, privacy and trust–Technology for tracking and controlling my data disclosures: Does this work?. In IFIP International Conference on Trust Management (pp. 3-14). Springer, Cham.

[17] Alty, James L., Roger P. Knott, Ben Anderson, and Michael Smyth. "A framework for engineering metaphor at the user interface." Interacting with computers 13, no. 2 (2000): 301-322.

[18] Demjaha, Albese, Jonathan M. Spring, Ingolf Becker, Simon Parkin, and M. Angela Sasse. "Metaphors considered harmful? An exploratory study of the effectiveness of functional metaphors for end-to-end encryption." In Proc. USEC, vol. 2018. Internet Society, 2018.

[19] Warner, Stanley L. "Randomized response: A survey technique for eliminating evasive answer bias." Journal of the American Statistical Association 60, no. 309 (1965): 63-69.

[20] Bullek, Brooke, Stephanie Garboski, Darakhshan J. Mir, and Evan M. Peck. "Towards understanding differential privacy: When do people trust randomized response technique?." In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3833-3837. 2017.

[21] PAPAYA, Deliverable D2.1: Use case specification, https://www.papaya-project.eu/node/153

[22] PAPAYA, Deliverable D2.2: Requirements specification, https://www.papaya-project.eu/node/154

[23] PAPAYA, Deliverable D5.1: E-health use case validation, https://www.papaya-project.eu/node/162

[24] PAPAYA, Deliverable D5.2: Telecom use case validation, https://www.papaya-project.eu/node/163

[25] PAPAYA, Deliverable D5.3: Refinement recommendations for the platform, https://www.papaya-project.eu/node/164

[26] PAPAYA, Deliverable D6.4: Intermediate business plan and exploitation report, https://www.papaya-project.eu/node/169

[27] Farzaneh Karegar, Simone Fischer-Hübner. “Vision: A Noisy Picture or a Picker Wheel to Spin? Exploring Suitable Metaphors for Differentially Private Data Analyses", EuroUSEC workshop, October 2021. ACM Digital Library (forthcoming - accepted).

Appendix 1 Evaluation Matrix for Metaphors (adapted from Demjaha et. al.)

	M+	M-
DP+	**(DP+M+)** Desirable: features provided by DP and supported by the metaphor. This leads to correct understanding and informed decision.	**(DP+M-)** Undesirable: features provided by DP and not supported by the metaphor. This leads to misunderstanding and underused features of DP.
DP-	**(DP-M+)** Very undesirable: features implied by the metaphor and not supported by DP: conceptual baggage. This leads to overestimating the privacy of information, unwanted data sharing based on the condition described incompletely, and trust issues.	**(DP-M-)** Not important: features not implied by the metaphor and not supported by DP.
If M is suitable for?	YES	NO
Local DP		
Aggregate-level DP		
Aggregate-level DP in the context of federated learning		

Appendix 2 Interview script

1. Main session

1.1. Part 1

1.1.1. Introduction of the persona and set the scene

The interviewer shares her/his screen and shows the persona and describes it the persona, and then shows the data analysis scenario depicted in Figure 1 and describes it.

Persona: Meet Alex, who is generally healthy but is suffering from stress. Alex is using a wearable device.

Alex

The wearable device measures and collects health-related data:

- The number of steps.
- Sleep cycles.
- Heart rate.
- …

Using the app, Alex can:

- Monitor what has been collected.
- Receive different recommendations.

Description of scenario:

Figure 1 is shared with the interviewee and the interviewer describes it using the text below:

“The app notifies its users, including Alex, that it is possible to receive supportive recommendations to assist them to cope with stressful conditions if they want and agree.

To do so, the health company needs to:

- Receive stress-related information from different users. Stress-related information, for example, may include users’ responses to questions about their moods daily or users’ selfie pictures on different occasions, when they feel stressed or not.
- Analyze the information it collects from different users to make a model for detecting and predicting different stressful conditions and then provide remedies and assistance to cope with stress. *The model works a little bit similar to how human brains process information and learn. Therefore, you can think of the model as an artificial brain that learns from its inputs. The input data to the model is users' stress-related information from which the artificial brain learns how to detect and predict stressful conditions.*

Nonetheless, the amount of data the health company can collect from its users are limited and does not suffice for the model to detect all stressful situations well. Therefore, the health company wants to participate in a collaboration project involving other similar companies who have the same goal and suffer from the same limitation. In this project:

- Each company analyzes its’ users data locally and makes a model for detecting stressful conditions.
- Each company shares its local model with an Internet-based analyzer.
- The Internet-based analyzer receives the local models from different companies and makes a better model for detecting stressful conditions. All the involved companies can then can benefit from the improved model.

In this scenario, Alex trusts the wearable device, her phone, and her health company but not the other companies involved and the internet-based analyzer. Therefore, working collaboratively with other companies and an Internet-based analyzer to make a better model for the detection of stressful conditions can negatively affect Alex’s (and other users’) privacy. So there is a privacy problem. Note that the model may reflect the specific characteristics of its input data including Alex’s data. Thus, it could be leaked, for example, that Alex has a stress problem.

However, to protect user privacy and mitigating the privacy problem, in this scenario, the health company (A) protect its model using a privacy mechanism to satisfy so-called differential privacy. DP is a formal notion of privacy and provides provable privacy assurances. This differentially private mechanism prevents privacy leakage, to a certain extent, about individuals to the internet-based analyzer and to the other health companies involved. Afterwards, the health company (A) shares the differentially private model with the Internet-based analyzer. Other companies involved do the same.”

1.1.2. Questions in part 1

Gauge the participant’s initial predisposition:

Q 1. “Have you heard about any privacy protection techniques (techniques that can be used to guarantee users privacy and to improve it)? Have you ever heard about differential privacy?”

If No:

Continue with gauging their understanding.

If Yes:

Q 2. "In which context did you hear about it?"

Q 3. "Do you know what differential privacy is? Can you explain it in your own language?"

Gauge their understanding and expectations:

Q 4. "Would you, if you were Alex, agree to share your data to be analyzed in the way described? What factors did play a role in the decision for Alex?

Q 5. (if the interviewee did not talk about it ask:)

How did the differential privacy mechanism play a role in your decision? (follow-up questions:) Would it matter if another mechanism were used to protect your privacy instead of differential privacy?"

If they did not agree to share:

Q 6. "What should have been different so you as Alex would agree?

From all regardless of decision they made:

Q 7. "What do you want to know about the mechanism applied (the differentially private mechanism) to protect your privacy? What information would you like to be added in the description of scenario?"

Q 8. As mentioned in the scenario, you as Alex trust your device and the health company but not other companies and the internet-based analyzer. Would you be concerned about it if the analyzer process your data? Why? What can mitigate your concerns?

Q 9. "What would be the benefits for you as Alex if you agreed? What would be the risks for you?"

Q 10. "(pointing to the figure of scenario) In this scenario, from whom do you expect your actual stress-related data to be hidden?

(follow-up:) Could your health app see your actual stress-related data? What about your health company? What about Internet-based analyzer? Or other companies involved?"

Q 11. "In this scenario, it is mentioned that your privacy is protected against potential privacy risks using a specific mechanism. What factors do play a role for you to trust this mechanism to protect your data?"

1.2. Part 2

1.2.1. Metaphor description

The interviewer shows Figure 6to the interviewee and describes it based on the text below:

"Now imagine that your health company wants to create a model that can recognize a user's emotion from his/her facial expression. Again, note that *you can think of a model as an artificial brain that learns from its inputs. In other words, a model can be trained based on the characteristics of its inputs to do a special thing.* The health company requests its users including Alex, to share their selfies and then uses the selfies to train a model so the model can recognize emotions based on facial expressions. The model can, for example, predict if a user is very happy, sad, somehow confused, stressed, furious etc. Here in this figure, you see a trained model based on users' selfies. Now if the trained model receives a user's selfie as its input it can predict the user's emotion.

As I mentioned before, the health company protects the users' privacy by using a differentially private mechanism to train the model as depicted in this figure. Before sharing its locally trained model with the analyzer, the health company perturbs (carefully distorts) the trained model based on the underlying mechanism. It means that the health company distorts the information the model has learned from selfies randomly but in a controlled way, for example, using the medium level of distortion. The purpose of distorting the trained model is to assure users' privacy by limiting the effects of each individual's selfie on what the model has learned from the selfies. Therefore, the mechanism applied guarantees that the likelihood of privacy harm users may face by being identified due to uploading their selfies and having their selfies used among other selfies to train the model is limited and insignificant.

This figure is not a precise representation of the underlying mechanism that distorts a trained model and is only a simple example of what distortion means.

Although each health company deliberately distorts its trained model, the final model is better than each of the locally trained models at recognizing the emotions. The final model made by the analyzer is also a distorted model that protects users' privacy."

1.2.2. Questions in Part 2

Q 12. "Would you change the decision you made on behalf of Alex in th previous step after receiving more information about differenti privacy? Why?"

Q 13. "In general, do you think that receiving information about th underlying privacy techniques a system uses would be helpful for yo in making your decision to use a system? How (in what way) it could b helpful?

Q 14. "Is the description of DP understandable and easy to grasp for you What is not clearly described or missed in the description? How th description could be improved?"

Q 15. "Is there any information surprising to you-- did not expect? Please elaborate"

Q 16. "Would you like to know more about the technical and mathematical details of the underlying differentially private mechanism? why? "

Q 17. "The mechanism perturbs (distorts) the model in a controlled way. Can you explain, in your own language, what is the model and what does it mean to distort the model? How does distortion protect your privacy? (follow- up: what is your idea about the need of distorting the trained model to protect your privacy?)

Q 18. " How would your privacy be better protected; by more distortion of the model or by less distortion? What happens if the model gets completely distorted?"

Q 19. What amount of distortion do you prefer to be applied to the model created by your health company? Why?

Q 20. "Can you explain whether there is a the trade-off between the accuracy of the results of data analysis (accuracy of the model) and the privacy of your data?

Q 21. "How you as Alex would be affected if the model is not accurate? Would you rely on the recommendations the app gives you to cope with stress? Why?"

Q 22. What do you think about the accuracy of the final model compared to this model (pointing to the undistorted model in figure)?

Q 23. "I will name some entities and I want you to tell me whether each of the entities I name would be able to get access to your (Alex's) actual stress-related data in your opinion? (Why do you think so?)"

a) Hackers who access the database of the health company? (follow-up: if they access the database, what do you expect they can access; the distorted model? The undistorted model? The actual stress-related data? Or all?)

b) People who know how the differentially private mechanisms work

if they access the distorted model?
(follow-up: What about if they access undistorted model?)

c) Your internet service provider
d) Law enforcement officials (if they contact the health company)

Q 24. "If your health company did not distort the model, would the analyzer be able to prove that your selfie was also used to train the model? What about your close friend (for example, your friend) if he/she gets access to the undistorted model? (Potential follow-up: Do you think that distorting the model would help to avoid it? How?)

Q 25. "How would the model change if you did not agree to have your stress-related data, in this case, your selfie to be used to train the model to recognize emotions? (Potential follow-up: Do you think that distorting the model would help to limit the change in the model that may happen because of your participation/lack of participation? How?)"

(follow-up: Imagine that you as Alex have a special characteristics that no other user has. For example, you are the only one with a very specific facial expression that the model classifies as stressful emotion. How the model would change if your selfie were not included?

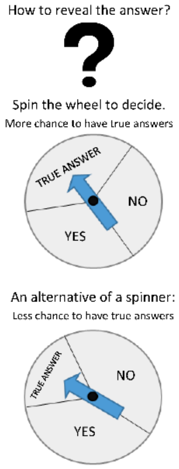
Now imagine we distort this model by distorting what the model learned about the stressful emotions. How the distorted model would change if you were selfie were not used to train it?"

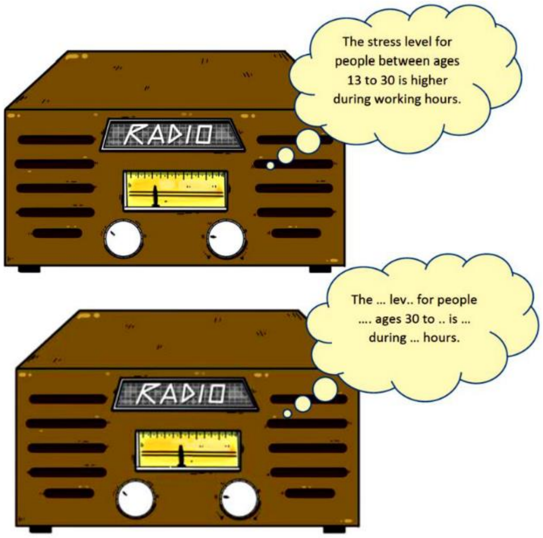
Q 26. "How would you describe the likelihood of remaining privacy risks? Would you accept the remaining risks? Would more information on remaining risks be of your interests for making decision to share your data or not?"

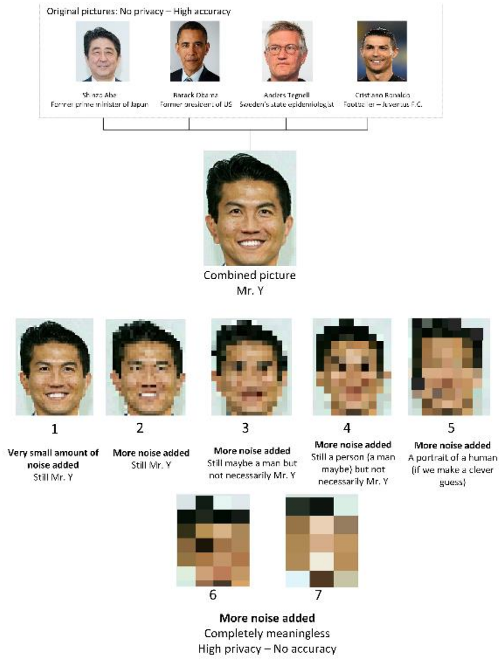
Q 27. "Now that you know more about differential privacy, would you trust this method in general to protect your privacy? Why? (if said NO:) What are your concerns in this regard?"

Q 28. "How would you describe differential privacy to someone who does not know about it? Can you think of any alternative description/example for data perturbation (noise addition/data distortion) rather the one we used for describing the concept of differential privacy?"

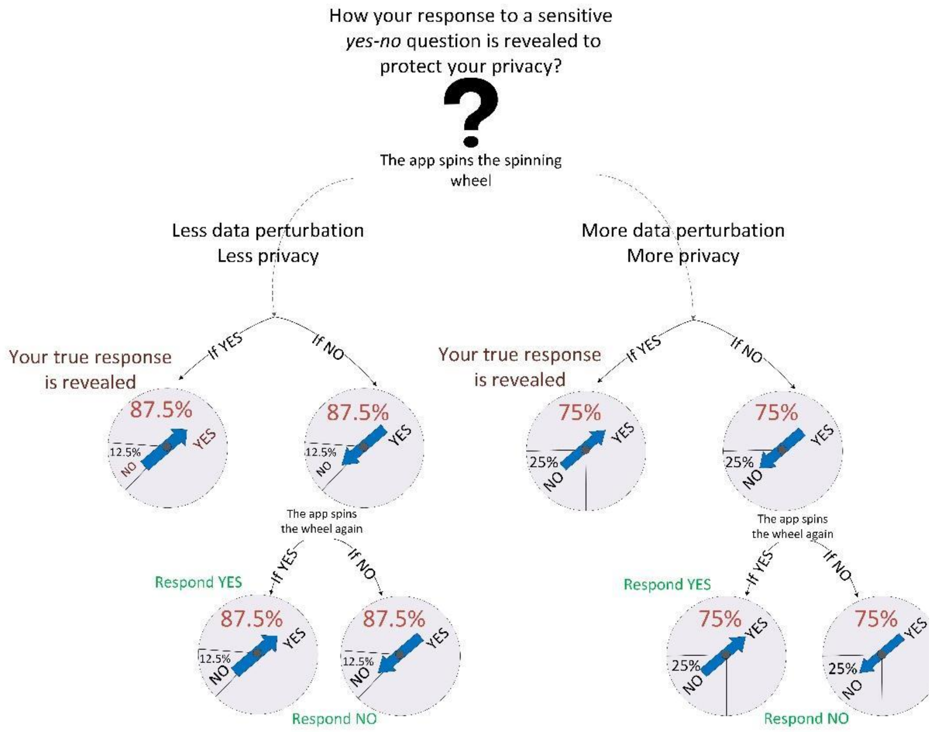
Appendix 3 Metaphors analysed in the first analytical evaluation

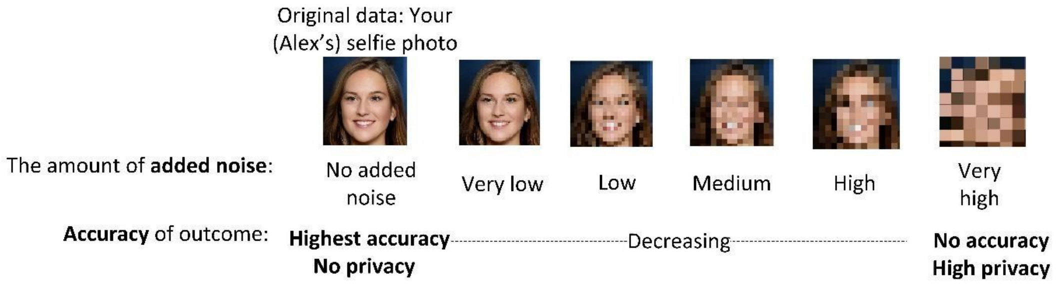

How to reveal the answer?
?
Spin the wheel to decide.
More chance to have true answers
TRUE ANSWER
NO
YES
An alternative of a spinner:
Less chance to have true answers
TRUE ANSWER
NO
YES


The stress level for people between ages 13 to 30 is higher during working hours.
RADIO
The ... lev.. for people .... ages 30 to .. is ... during ... hours.
RADIO


Original pictures: No privacy – High accuracy
Shinzo Abe
Former prime minister of Japan
Barack Obama
Former president of US
Anders Tegnell
Sweden's state epidemiologist
Cristiano Ronaldo
Footballer – Juventus F.C.
Combined picture
Mr. Y
1
Very small amount of noise added
Still Mr. Y
2
More noise added
Still Mr. Y
3
More noise added
Still maybe a man but not necessarily Mr. Y
4
More noise added
Still a person (a man maybe) but not necessarily Mr. Y
5
More noise added
A portrait of a human (if we make a clever guess)
6
7
More noise added
Completely meaningless
High privacy – No accuracy

Appendix 4 Appendix 4: Adapted metaphors resulting from the analytical evaluation


How your response to a sensitive *yes-no* question is revealed to protect your privacy?
?
The app spins the spinning wheel
Less data perturbation
Less privacy
More data perturbation
More privacy
Your true response is revealed
If YES
If NO
87.5%
12.5%
NO
YES
87.5%
12.5%
NO
YES
Your true response is revealed
If YES
If NO
75%
25%
NO
YES
75%
25%
NO
YES
The app spins the wheel again
Respond YES
If YES
If NO
87.5%
12.5%
NO
YES
87.5%
12.5%
NO
YES
Respond NO
The app spins the wheel again
Respond YES
If YES
If NO
75%
25%
NO
YES
75%
25%
NO
YES
Respond NO


Original data: Your (Alex's) selfie photo
The amount of **added noise**:
No added noise
Very low
Low
Medium
High
Very high
Accuracy of outcome:
Highest accuracy
No privacy
Decreasing
No accuracy
High privacy