



HEALTHYCLOUD
Health Research & Innovation Cloud

D7.3 Validation of proposed solutions and gap analysis

Version 1.0

Document Information

| | |
|----------------------|---|
| Contract Number | 965345 |
| Project Website | http://www.healthycloud.eu/ |
| Contractual Deadline | M28, June 2023 |
| Dissemination Level | PU |
| Nature | R |
| Author(s) | Lidia López (BSC), Salvador Capella-Gutierrez (BSC), Juan González (IACS) |
| Contributor(s) | Davit Chokoshvili (PNED G.I.E., WP2), Ilaria Colussi (BBMRI; WP2), Michaela Th. Mayrhofer (BBMRI; WP2), Celia Álvarez (SAS; WP3, WP4), Rossana Castaldo (SDN; WP3, WP4), Shona Cosgrove (Sciensano, WP3), Irini Kessissoglou (Sciensano, WP3), Stefan Klein (ERASMUS MC, WP3, WP4), Silvia Rodríguez (SAS; WP3, WP4), Teresa D'Altri (CRG, WP4), Sina-Victoria Barysch (de.NBI Cloud, WP5), Salvador Capella-Gutierrez (BSC, WP5), Eva Garcia (BBMRI, WP6), Laura Portell (BSC, WP6), Marco Roos (Leiden UMC; WP6), Emilie Cauët (Sciensano, WP7), Esmeralda Ruiz (Universitat de Barcelona, WP7) |
| Reviewer(s) | Harald Wagener (Charité), Anna Niemeyer (TMF) |
| Keywords | Evaluation, Challenges, Proposals |



Notice: The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement N°965345

(c) 2021 HealthyCloud Consortium Partners. All rights reserved.

Change Log

| Version | Author | Date | Description of Change |
|---------|--|------------|--|
| V 0.1 | L. López | 03/03/2023 | Initial version of the outline (TOC) |
| v 0.2 | L. López | 10/05/2023 | List of challenges |
| v 0.3 | J. González | 12/05/2023 | review |
| v 0.4 | L. López | 22/05/2023 | Use case challenges & HealthyCloud proposals |
| v 0.5 | L. López | 22/05/2023 | Results |
| v 0.6 | S. Capella-Gutierrez | 02/06/2023 | Results review |
| v 0.7 | J. González | 07/06/2023 | General review |
| v 0.8 | MT. Mayrhofer | 13/06/2023 | Review, focus on legal and governance |
| v 0.9 | L. López | 19/06/2023 | General review, version sent to reviewers |
| v 0.91 | A. Niemeyer | 11/07/2023 | General review, Focus on understanding and consistency |
| v 0.92 | H. Wagner | 11/07/2023 | General review |
| v 1.0 | L. López, S. Capella-Gutierrez, J. González. | 24/07/2023 | reviewer's comments processed and general review |

Table of contents

| | |
|---|----|
| Executive Summary..... | 4 |
| 1 Introduction | 5 |
| 2 Methodology | 5 |
| 2.1 Gap Analysis..... | 5 |
| 2.2 HealthyCloud evaluation process | 6 |
| 3 Use case challenges (as-is)..... | 7 |
| 3.1 Health-related research project process | 8 |
| 3.2 Challenges..... | 13 |
| 3.2.1 Challenges to define research question..... | 13 |
| 3.2.2 Challenges to Identify datasets | 13 |
| 3.2.3 Challenges to identify computational resources | 14 |
| 3.2.4 Challenges to understand request access procedure | 14 |
| 3.2.5 Challenges to request access | 14 |
| 3.2.6 Challenge to gain access to datasets..... | 15 |
| 3.2.7 Challenges to integrate data | 15 |
| 3.2.8 Challenges to set-up analysis process | 16 |
| 3.2.9 Challenges to perform analysis | 16 |
| 3.2.10 Challenges to understand data submission process | 17 |
| 3.2.11 Challenges to data submission request | 17 |
| 4 HealthyCloud proposal (to-be) | 17 |
| 4.1 Research process including HRIC..... | 17 |
| 4.2 HealthyCloud proposals..... | 19 |
| 5 Results..... | 21 |
| 5.1 Gaps in user profiles | 21 |
| 5.2 Gaps related to identify datasets | 21 |
| 5.3 Gaps related to identifying computational resources..... | 22 |
| 5.4 Gaps related to understand request access procedure | 22 |
| 5.5 Gaps related to request access..... | 23 |
| 5.6 Gaps related to gain access to datasets | 24 |
| 5.7 Gaps related to integrate data | 26 |
| 5.8 Gaps related to set-up analysis process | 26 |
| 5.9 Gaps related to perform analysis | 27 |

| | | |
|------|--|----|
| 5.10 | Gaps related to understand data submission process | 27 |
| 5.11 | Gaps related to data submission request | 27 |
| 6 | Conclusion | 28 |
| 7 | Next steps | 30 |
| | Project Deliverables | 31 |
| | Acronyms and Abbreviations | 33 |

Figures

| | | |
|-----------|---------------------------------------|----|
| Figure 1. | Evaluation process | 6 |
| Figure 2. | BPMN 2.0 constructs..... | 8 |
| Figure 3. | Use cases execution process..... | 9 |
| Figure 4. | Research process including HRIC | 18 |

Tables

| | | |
|-----------|--|----|
| Table 1. | HealthyCloud knowledge outcomes | 19 |
| Table 2. | HealthyCloud functionality outcomes | 20 |
| Table 3. | HealthyCloud legal outcomes | 20 |
| Table 4. | Addressed challenges to identify datasets | 22 |
| Table 5. | Addressed challenges to identify computational resources..... | 22 |
| Table 6. | Addressed challenges to understand data request access procedure | 23 |
| Table 7. | Addressed challenges to understand computational resources request access procedure | 23 |
| Table 8. | Addressed challenges to request data to a data infrastructure | 24 |
| Table 9. | Addressed challenges to request services to a computational resources infrastructure | 24 |
| Table 10. | Addressed challenges to provision access to datasets | 24 |
| Table 11. | Addressed challenges to integrate data | 26 |
| Table 12. | Addressed challenges to set-up analysis process | 26 |
| Table 13. | Addressed challenges to request services to perform analysis..... | 27 |
| Table 14. | Addressed challenges to data submission request..... | 27 |
| Table 15. | Identified gaps during the validation | 29 |

Executive Summary

The final objective of the WP7 is the validation of the HealtyCloud outcomes driven by two use cases on cancer and atrial fibrillation research. We conducted a gap analysis to evaluate if the project outcomes would support the challenges that researchers face during the execution of their research projects. The execution and the results of this validation process are reported in this deliverable (D7.3).

Project use cases have identified a set 23 challenges that researchers often face when they conduct health-related research projects based on the use and reuse of health data. To have a better understanding of the health-related research projects, we characterised this process and classified the 23 challenges according to the activity in the process they are related to.

The health-related research project process has a total of 11 activities, it starts with the definition of the research question and finishes when the researchers submit their results, which may be reused in the future, going through data discovery, access request, data and computational environment preparation, and data provision and use.

Project outcomes have been classified as *knowledge* (guidelines and recommendations), *functionality* (concrete services), and *legal* (addressing the legal and ethical basis for the use and re-use of health-related data). Most of the defined process activities (7 of 9) have all the challenges addressed for the three categories. The activities that do not have all their associated challenges addressed from the three categories are all related to computational resources: “*to understand request access procedure*” and “*to request access*”. Two of them are not addressed for *functionality* and one for *knowledge*. There is also one gap that is not related to any challenge, we found a missing user profile in the set of HealthyCloud users profiles definition (D6.1. - FAIR Health Data Portal expected users' interactions), the one characterising computational resource stakeholder.

Regarding future steps, to be addressed beyond the HealthyCloud project, some future research lines have been identified. Further exploration related to data management is expected to lead to new challenges, for example regarding the turn down of computational resources at the end of analysis as well as data and results retention. In federated scenarios, when the data is analysed in the data holder premises, we judge that data security issues are minimised, but not having access to the data (only to the results) can bring some specific challenges that need further exploration. For instance, involving a trusted third party that vouches, among others, for data quality.

The results of the validation confirm that the access to health-related data is much more regulated than access for computational capabilities. We could anticipate that the broad adoption of Trusted Research Environments (TRE) for processing transnationally sensitive data would represent a change in this situation.

1 Introduction

The HealthyCloud project was funded with the aim of developing the Strategic Agenda for the European Health Research and Innovation Cloud (HRIC) together with relevant stakeholders in the interface of health-related research and cloud technologies. The goals of HealthyCloud's WP7 are to identify HRIC functional requirements by driving two theoretical use cases' and to validate the project outcomes provided by the legal (WP2), knowledge and functionality (WPs 3, 4, 5 and 6).

The use cases functional requirements have been reported in previous deliverables D7.1 (Functional requirement analysis report of Use Case 1 - cancer) and D7.2 (Functional requirement analysis report of Use Case 2 - atrial fibrillation). The deliverables also include the challenges associated with the activities performed in a health-research research project. This deliverable (D7.3) is reporting the execution and the results of the validation process for the project outcomes.

For validation, we followed an agile and iterative methodology involving all the work packages, concluding with a gap analysis to validate if the project outcomes are fitting to the use cases requirements. As the project outcomes are meant to support researchers' activities, we used the challenges behind the use case requirements reported in the previous WP deliverables to lead this gap analysis.

The results reported in this deliverable are meant to be used as an input for the definition of the final HRIC Strategic Agenda (WP8).

2 Methodology

To validate the project outcomes, we need to answer the following question:

Are the HealthyCloud outcomes addressing the challenges that researchers face when they are using and reusing health-related data for their research projects?

The methodology applied for answering this question has been based on a gap analysis between the researchers' challenges and project outcomes.

2.1 Gap Analysis

Gap analysis helps organisations to identify the activities that can be improved (gaps) and then find solutions that can be implemented to improve their performance. In our case, this analysis has been performed to identify the gaps and to validate if the project technological proposals are suitable to fill the identified gaps.

Gap analysis methodology consists of the following steps:

1. *Analyse the current state (as-is).* In this step, we need to understand the current state, in HealthyCloud, this would imply getting to know how researchers are currently performing their research. The use cases have been defined and partially executed in a way that we have been able to characterise a generic process that is followed by researchers when they conduct their activities with the focus on the use and reuse of health-related data for research purposes. As part of this generic process definition, use cases also identified current challenges as open questions to be improved by using HealthyCloud proposed outcomes.
2. *Identify the ideal future state, what we want to achieve (to-be).* In HealthyCloud, the future state is led by the implementation of the work packages outcomes through the Health Research and Innovation Cloud (HRIC).
3. *Find the missing pieces (gaps) between the ideal future state (to-be) and the current state (as-is).* In the case of research projects, the challenges are the gaps. Therefore, we have looked for gaps between use case challenges and project outcomes.
4. *Identify improvements to close the gaps.* We are using gap analysis to evaluate if the project outcomes, the improvements to fill the gaps, are addressing all the use case identified challenges (gaps).

2.2 HealthyCloud evaluation process

We followed an agile and iterative methodology involving all the work packages (see Figure 1).

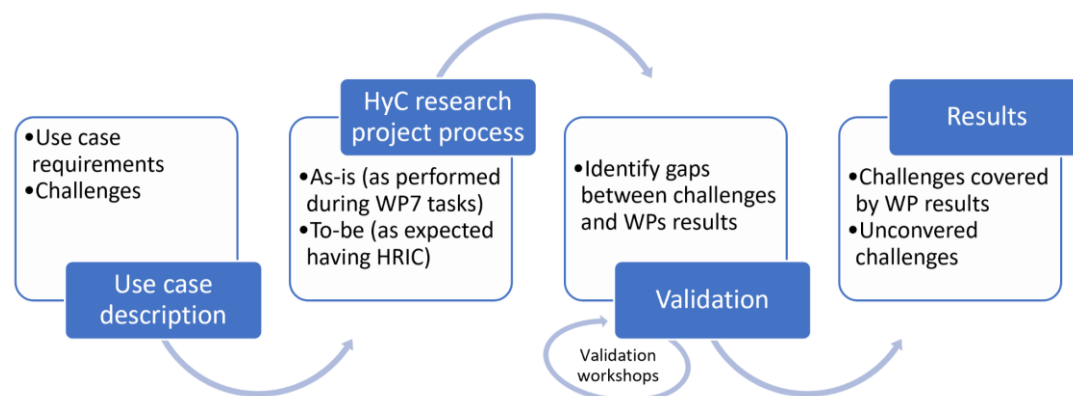


Figure 1. Evaluation process

The evaluation process was divided in the following phases:

- *Use case description.* Use cases defined the use case requirements and identified the associated requirements (see deliverables D7.1 and D7.2).
- *HealthyCloud health-related research project process.* First, in the context of the WP7 team, we conducted several workshops for the definition and

refinement of the expected generic process, including HRIC as an available set of tools, guides, and best practices. One important goal of these workshops was to identify the interaction between researchers and the potential HRIC components, and the other ecosystem stakeholders (i.e., infrastructure providers). The different versions of the generic process have been shared with the other work packages with the aim to present the generic process and gather feedback.

- *Validation.* Once we have the expected generic process identified and the challenges defined, we conducted a total of three dedicated workshops with work packages representatives to collect the set of project outcomes and how these outcomes were addressing the challenges. The series of workshops ended with a final workshop where the identified gaps were shared and discussed by all the members involved. The workshop attendees are included as contributors of this report indicating the institution and the WP they were representing.
- *Results.* This is the last step of the evaluation process. The results and feedback collected during the validation workshops were streamlined and reported in this document (deliverable D7.3). In order to bring together all the information, we contacted WPs to complement the information gathered in the validation workshops.

3 Use case challenges (as-is)

This section describes the process of use cases conducted during the execution of the project. This process has been conducted in collaboration with the other work packages to gather the information that use cases leads need, for example, when the use cases needed to identify the datasets to be (re-)used, they asked WP3 and WP4 to understand where to find the data.

The main aim of the use cases in HealthyCloud was not the evaluation results of the execution of the use case, but exposing the actual challenges that researchers face during the execution of a research project. HealthyCloud project outcomes, e.g. guidelines, recommendations, specifications, best practices, will be reflected as part of HRIC Strategic Agenda, which aims to facilitate systematic trans-European use of health-related data for research purposes.

As part of the characterisation of the current process, we have generated an abstract process model that characterises the high-level activities that are usually performed in a health-related research project based on the use and reuse of health-related data.

3.1 Health-related research project process

The goal of the health-related research project process (depicted in Figure 3) is the execution of a research project based on the secondary use of existing health-related data. It is out of the scope of this process the activities related to the health-related data extraction from the primary sources and its deposition at data hubs and other structures for the management of such datasets.

The process includes the activities performed by the researcher (blue lane in the middle) and the interactions that have been taken to the other work packages and infrastructure providers (data and computational).

The process is described using Business Process Modeling Notation 2.0 (BPMN 2.0¹), Figure 2 describes the subset of the language constructs used in the process model.

For researchers, the process has a total of eleven activities, starting with the definition of the research question, going through data discovery, access request, data and computational environment preparation, data provision and use, and finalising with the publication of results, which may be reused in the future.

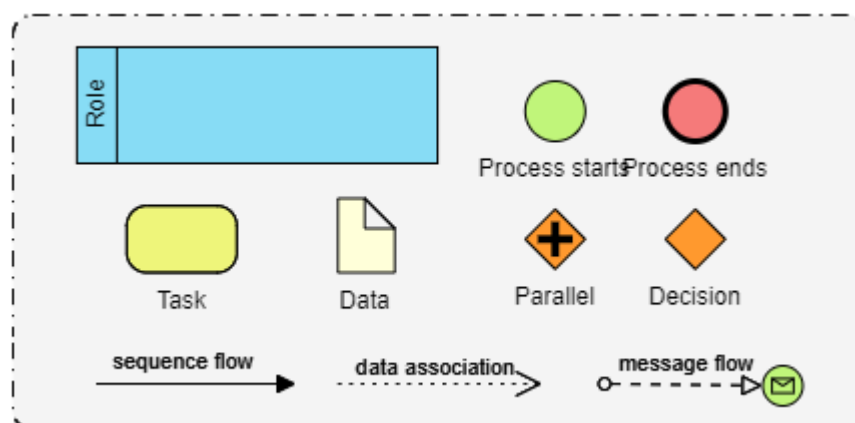


Figure 2. BPMN 2.0 constructs

¹ <https://www.bpmn.org/>

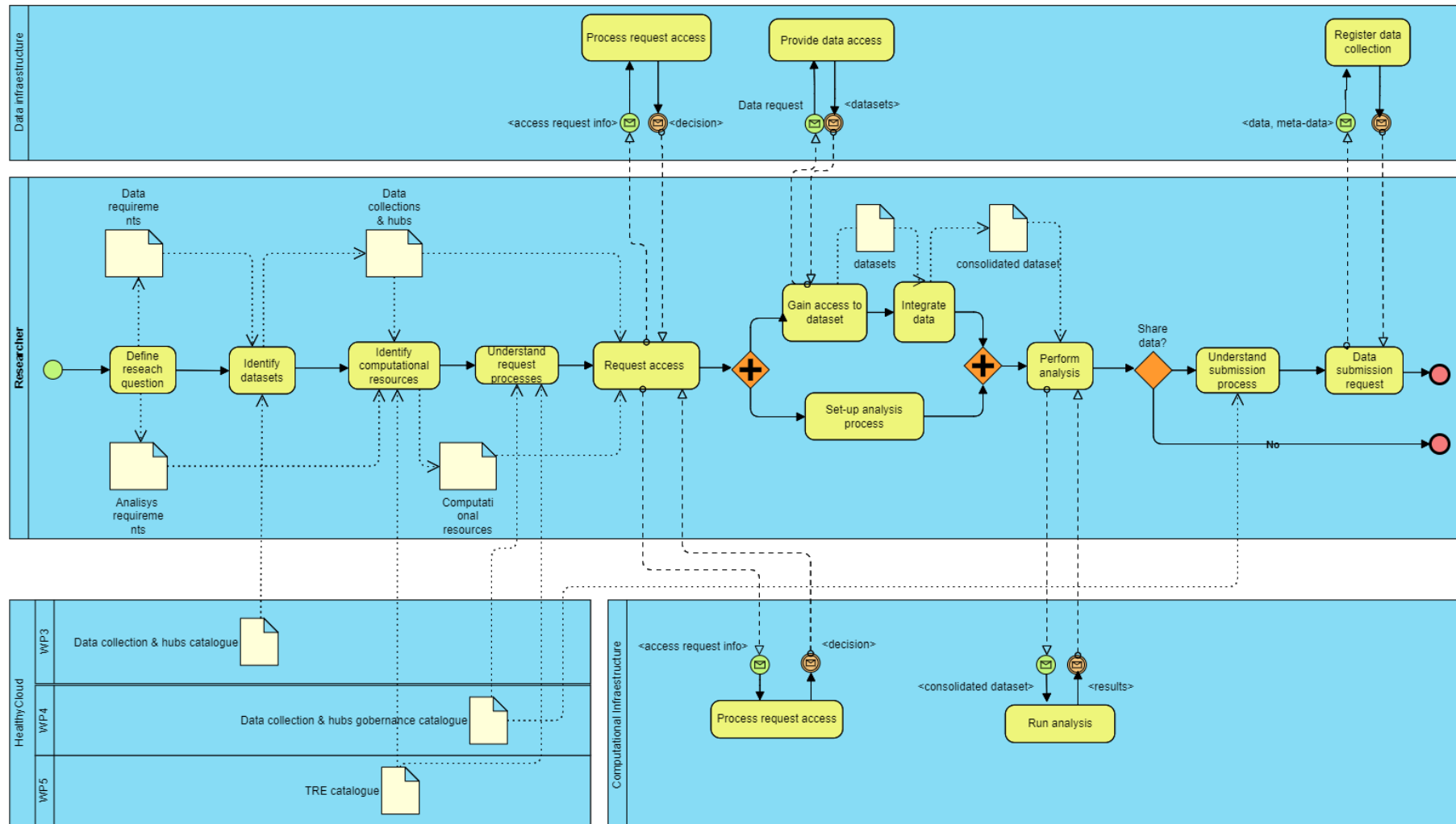


Figure 3. Use cases execution process

The following tables describe the process activities to be performed by the researchers (yellow boxes in the *Researcher* lane in the model). The activities included for other roles, specifically *Data* and *Computational infrastructure providers*, have their own processes associated that depend on the concrete infrastructures.

To keep the process simple, some of the activities cover more than one task. For instance, the activity *Request access* is referring to separated steps for requesting data (to the different data hubs and/or data collections) and computational resources, and the potential iterations until access is finally granted. Each request corresponds to concrete processes and requirements that depend on the infrastructures.

| | |
|--------------------|--|
| Activity | Define research question |
| Description | This activity corresponds to the specification of the scientific question that a given researcher wants to answer using health-related data. |
| Input | None |
| Output | Data requirements and analysis requirements |

| | |
|--------------------|---|
| Activity | Identify datasets |
| Description | Search the suitable set of health-related datasets that are going to be used as input for their research, also known as data discovery. |
| Input | Data requirements |
| Output | Set of <data collection, datahub> |

| | |
|--------------------|---|
| Activity | Identify computational resources |
| Description | Search the suitable set of computational resources needed for the analysis, which can host the requested data |
| Input | Analysis requirements, set of <data collection, datahub> |
| Output | Set of <computational resource> |

| | |
|--------------------|--|
| Activity | Understand request access processes |
| Description | Gather the needed knowledge about the process to request access to data and computational resources from all the infrastructure providers that the researcher needs to use (data infrastructures and computational resources infrastructures) for completing their research project. |
| Input | Set of Infrastructure providers |
| Output | Set of request data requirements and processes to request the data and computational resources |

| | |
|--------------------|---|
| Activity | Request access |
| Description | <p>Preparing the needed information to ask for access to all the data and computational resources. For instance, the project proposal and data management plan.</p> <p>This particular activity has been simplified in the above figure to facilitate its readability because requesting access implies an iterative process between the researcher and the data and/or computational provider until access is granted. Note that separated requests for health-related data and computational access are needed as it is highly unlikely that a single provider can provide both services.</p> |
| Input | Set of <data collection, datahub>, computational resources to be used for the analysis, infrastructures requirements (e.g., data requirements, analysis requirements), and request process. |
| Output | Access accepted (yes/no) |

| | |
|--------------------|--|
| Activity | Gain access to datasets |
| Description | The action of accessing the needed data for conducting by the researcher once permission has been granted. Note this can be done via a Trusted Research Environment where the researcher can utilise the data or provide it as downloadable data. Other mechanisms are emerging, e.g., federated analysis and federated learning, but they are not mature enough to be fully modelled. |
| Input | Data request information (depends on the data infrastructure) |
| Output | Set of datasets (to be download or deposited in a computational resource) or dataset access rights (to be used for the computational resource to be used to get the dataset) |

| | |
|--------------------|--|
| Activity | Integrate data |
| Description | Put all the data in a consolidated dataset if needed. |
| Input | Set of datasets |
| Output | Consolidate dataset to be used in the analysis (1 dataset) |

| | |
|--------------------|--|
| Activity | Set-up analysis process |
| Description | Perform all the activities needed to get the computational resources ready to start the analysis (in-house or getting access to the external computational resources). |
| Input | Analysis requirements, computational infrastructure requirements |
| Output | Computational environment ready to perform the analysis |

| | |
|--------------------|--|
| Activity | Perform analysis |
| Description | Run the analysis, this activity includes (at the end) the turn-down of computational resources at the end of analysis as well as cleaning up the data. |
| Input | Consolidated dataset (if needed), data access rights (if needed), access rights to computational resources |
| Output | Resulting data (output data) |

| | |
|--------------------|---|
| Activity | Understand data submission process |
| Description | Gather the needed knowledge about the process to submit data (secondary-use data) from the data infrastructure where the data is going to be submitted. This data can be papers, resulting models and enriched datasets (re)used as inputs. In the case of the reused datasets, the enriched data is going to be submitted to the same data infrastructure that provides them. Each data infrastructure can have their own submission policies. |
| Input | Set of Infrastructure providers |
| Output | Request data requirements and process to submit the data |

| | |
|--------------------|---|
| Activity | Data submission request (for data sharing) |
| Description | The activities needed to prepare the data to be submitted to the corresponding data infrastructure |
| Input | Dataset used for the analysis, resulting data and/or papers, data infrastructure requirements for submitting data |
| Output | Registered acknowledgment (Yes/No) — ideally returning a permanent identifier |

3.2 Challenges

In this section, we introduce the challenges identified by the use cases across the process presented in the previous sub-section. The challenges are classified according to the activities in the process they are related to.

3.2.1 Challenges to define research question

No challenges associated with this activity.

3.2.2 Challenges to Identify datasets

This activity corresponds to the task to find the datasets that are going to be (re-)used for answering a specific research question (data discovery). The first two challenges are related to the need to access the infrastructures of several data providers in order to obtain all the data needed for the specific research project. The last two challenges are related to the fact that it is difficult to know which data are available.

- *Several data sources.* This challenge arises when several datasets (from different data hubs and/or data collections) are required to answer the research question. Identifying the required data is challenging in many research fields due to the lack of a centralised portal being a difficult task of searching the appropriate data in literature or via contacts.
- *European/National/Regional level.* Depending on the country, you need to inspect either only one national or several regional data hubs/collections. These hubs/collections inform about the legal/ethical requirements for data (re-)use.
- *Uncertainty/unknown data requirements.* The lack of detailed information about datasets tends to prevent the mapping of the research questions to specific data requirements (e.g., specific variables on different datasets). The lack of a common standard for “ELSI Metadata” (i.e., use conditions and restrictions associated with a dataset), also makes it difficult to understand whether an existing dataset can be used legitimately for the intended purpose, further complicating the task of identifying appropriate datasets.
- *Missing common metadata catalogue.* Several data and metadata models are used across the European hubs landscape. A common metadata catalogue should include data governance, the data type and granularity of the data, quality of the datasets and the compliance with the FAIR data principles. As a summary, the lack of standardised metadata leads to the difficulty to find the data in a harmonised way.

3.2.3 Challenges to identify computational resources

This activity corresponds to the task to find the computational resources that are going to be used for the analysis.

- *Minimum computational resources at researcher's premises* (analysis performed at cohort premises in federated scenario). This scenario implies that there is sufficient computational capacity for executing the data analysis at the researcher's side. This might not be always certain due to the lack of computational resources at the researcher's side and/or that data is made available through other mechanisms, e.g., Trusted Research Environments (TRE), that have enough computational power for carrying on the expected analysis.
- *Uncertainty/unknown computation requirements*. Sometimes, the researchers do not know the specific requirements at this early stage of the research.

3.2.4 Challenges to understand request access procedure

This activity corresponds to the task of finding the needed information to understand the requirements and concrete steps needed to request access, including both legal and ethical requirements. The challenges identified for this activity apply to both, data and computational request access procedures.

- *Heterogeneous procedures*. Each infrastructure defines its own request access process, defining different steps and tools to be used. Each infrastructure has locally, regionally, or even nationally different procedures on ethics approvals or governance models. Each country has different legal frameworks that at times can be interpreted differently locally at the hub/collection level. An additional challenge is compliance interdependencies among data controllers.
- *Heterogeneous requirements*. This challenge is directly related to the first one (heterogeneous procedures), each data hub/collection defines a different set of information that needs to be sent in order to consider the access request. For example, the content of the project proposal.
- *Request process information is not easy to find*. The information related to the procedure is not easily findable. Sometimes the researcher needs to contact a helpdesk in the data hub/collection or go to a (hopefully easy to find) concrete page in a website. Interestingly, it might be possible that the information is only accessible in the local language adding a new complication layer.

3.2.5 Challenges to request access

This activity corresponds to the task of requesting access to the data or to computational resources. For this activity, the project use cases identified some challenges that apply to both cases:

- *Detailed study protocol needed.* Some information needed for the access request would not be decided at this stage generally, because it is something that is not needed yet. For example, specific variables, concrete analysis methods.
- *Using sensitive data.* This challenge is directly related to anonymisation/pseudonymisation and the need of moving data into a secure environment. Importantly, secondary use of health-related data for research purposes should be at least pseudonymized.

3.2.6 Challenge to gain access to datasets

This activity corresponds to the task to obtain real access to the requested datasets. The activity contemplates two different scenarios. In the first scenario, the researcher downloads the data. In the second scenario, the researcher obtains the access rights for the data to be transferred to the computational resources. In the second scenario, the data will be accessed directly from the computational resource during the analysis.

- *Data security.* In the case of downloading the data, the researchers need to guarantee security of the downloaded data; thus, using firewalls, avoiding access from non-granted researchers. When using a given computational facility, it is those that require commonly agreed security measures. In the case of downloading the data, the researcher needs to ensure that data processing is in line with security, legal and ethical requirements relevant for the purpose of data use.

A variant of the second scenario (data accessed directly from the computational resources) is when the analytics are performed by the data holder, in this scenario the researcher does not have access to the data, only to the analysis results. We keep this scenario out because the data security challenge does not apply directly, although software should be analysed to ensure that beyond doing what is expected it does not try to leak data, e.g., establish outbound connections and send a copy of local data somewhere else through the Internet. This is different from sharing automatically aggregated results or model gradients trained locally as this type of interaction is expected.

3.2.7 Challenges to integrate data

This activity corresponds to the task to integrate the data corresponding to more than one data model before running the analysis.

- *Data collected with different parameters, for example different time spans with different update periodicity.* In this example, analysis that depends on the follow-up may need work to harmonise the data capture.
- *Structured and/or unstructured individual data level.* Unstructured data will require extra processing effort (AI-based, NLP algorithms, etc.).

- *Linkage limitations with multiple levels of aggregation.* Data from the same individuals is scattered across multiple datasets with same or different identifiers, which should be anyhow pseudonyms to prevent re-identification.
- *Data harmonisation.* When accessing multiple datasets, it is necessary to ensure that all of them are semantically interoperable, which implies using the same ontologies and controlled vocabularies for the content and having data mapped to the same common data model. Data expertise is needed to ensure true comparability about the contents of the data. Specific issues can also arise if different kinds of data should be harmonised, e.g., health records and health variables and questionnaires coming from different cohorts, or the acquisition protocol hardware used to get images. Pre-processing must be performed before integrating datasets.
- *Variable mapping.* Identifying variables associated with the underlying research question and mapping them to the different data hubs and collections varies from site to site with many of them using (semi-)manual processes. As this is an iterative process between the researcher and the data infrastructure, it might take time even to the order of months.

3.2.8 Challenges to set-up analysis process

This activity corresponds to the task to set-up the computational resources, e.g., manage the users that are going to have access and select and upload the software that is going to be used for the analysis.

- *Analysis software quality.* Lack of support for assessing code quality and documentation of the components to be used for the analysis.
- *Analysis software accessibility.* Issues accessing source code and containerization deployment of the components to be used for the analysis.

3.2.9 Challenges to perform analysis

This activity corresponds to the task when the analysis is performed by the computational resources.

- *Cross-border data.* Analysing cross-border data is challenging from the legal and ethical point of view. In the absence of a comprehensive ELSI Metadata (data use conditions and restrictions) model and its mapping onto data processing workflows, it is not easy to demonstrate that the analysis is performed in an ethically and legally compliant manner.
- *Scalability issues.* Computational resources assigned may limit the analysis.
- *Guarantee security/privacy communication to minimise any security/privacy breach.* In a federated scenario, communications security and the data exchange should guarantee the data privacy.
- *Results replicability.* Difficulty to guarantee the provision of same analysis environments when analysing cross-border data or when environments are updated without notice in a single node.

- *No common follow-up patient records.* Missing regulation for the follow-up patient records, i.e., the evolution of the person in the same cohort.

3.2.10 Challenges to understand data submission process

This activity corresponds to the task to find the needed information to understand the requirements and concrete steps needed to submit output data related to the research results.

The use cases do not identify challenges for this activity, use cases report that the submission procedures are not as complex as the data access procedures.

3.2.11 Challenges to data submission request

This activity corresponds to the task to submit the output data related to the research results. This data could be the resulting model for the investigation, or the reused data when this data has been complemented.

- *Sensitive data issue.* Data privacy must be guaranteed using the appropriate data infrastructure and following the original conditions when accessing the original data.

4 HealthyCloud proposal (to-be)

One of the outcomes of the project is the FAIR Health Data Portal specification provided by WP6, including the requirements for the future HRIC². This specification is based on the outcomes from the work packages WP3, WP4, and WP5 and it can be extended to eventually facilitate information about the available computational capabilities for analysing the health-related data available through the meta-catalogue. Figure 4 describes the considered scenario for the gap analysis, which includes the HRIC as a tool to be used during the research project process.

4.1 Research process including HRIC

The activities performed by researchers (blue lane in the middle) are the same as in the current scenario (as-is). The difference between both processes is that in the desired scenario the researcher will interact with some components of the HRIC ecosystem to facilitate the overall activity, speed-up processes and provide clarity on the secondary use of these data for research purposes.

² We refer to the Fair Data Portal specification as HRIC specification in the rest of the document.

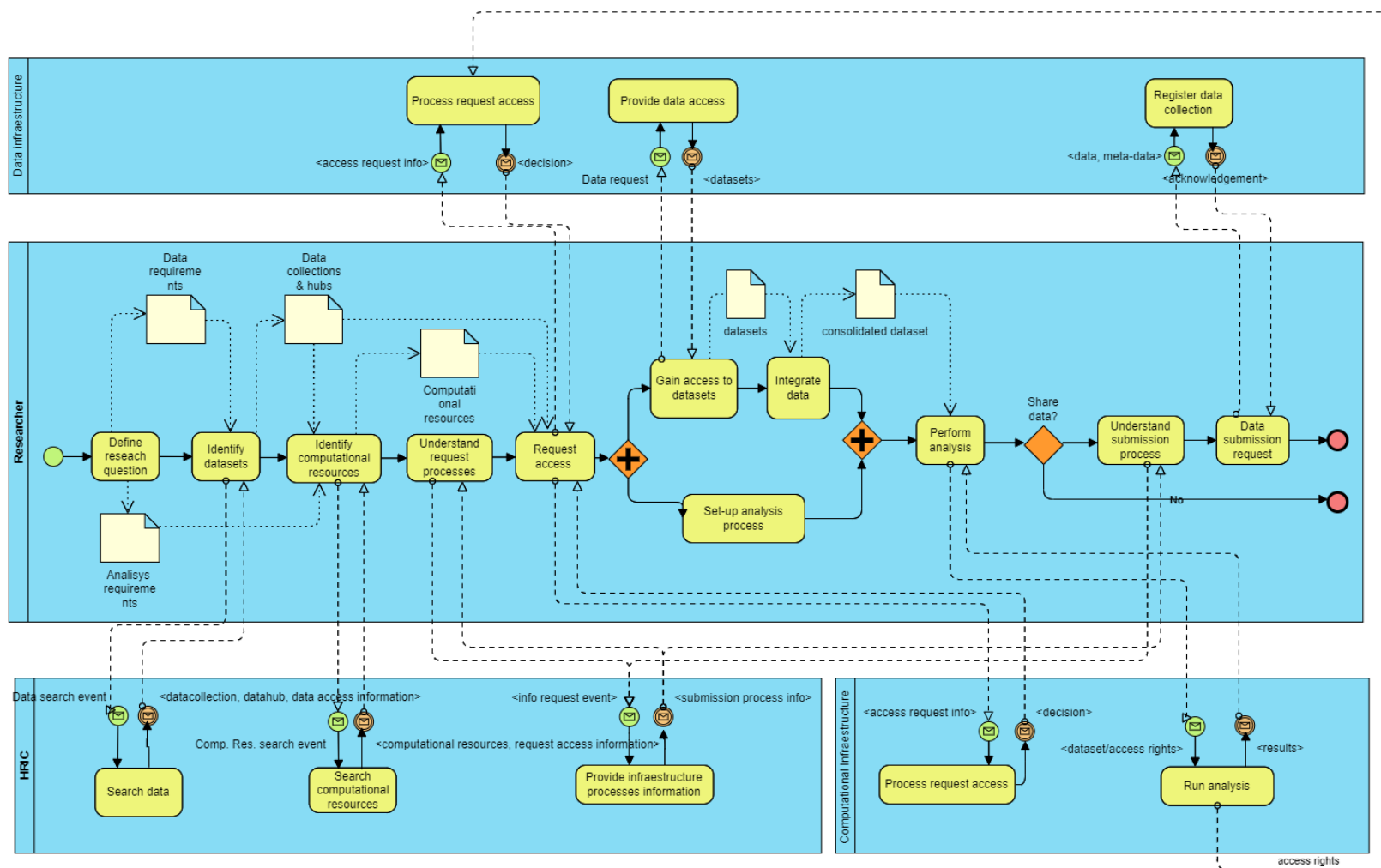


Figure 4. Research process including HRIC

4.2 HealthyCloud proposals

The HealthyCloud proposals should be supporting the use case challenges, filling the gaps between the expected scenario (this section) and current scenario (Section 2.3). We have classified the project outcomes into three categories:

- **Knowledge.** Work packages WP3, WP4, and WP5 have produced two kinds of knowledge, the ones related to the state of the practice in some of the related topics (Table 1, first nine rows) and some recommendations (Table 1, last seven rows).
- **Functionality** (Table 2). WP6 has specified the FAIR Health Data portal as a set of functionalities that should be included in the HRIC infrastructure.
- **Legal** (Table 3). WP2 has produced a set of recommendations to be followed by the HRIC ecosystem stakeholders on governance, modular contract clauses and risk analysis.

Table 1. HealthyCloud knowledge outcomes

| Knowledge recommendations | Deliverable |
|---|-------------|
| Landscape analysis of FAIRness levels | D3.1 |
| Landscape analysis using a health related-data catalogue matrix | D3.3 |
| Report on current discoverability solutions | D4.2 |
| Report on current FAIR adoption level | D4.2 |
| Usage and auditing data hubs indicators | D4.3 |
| Incentive system implemented in the data hubs | D4.3 |
| Landscape analysis of computational resources | D5.1 |
| Secure analysis based on individual research software and workflows managers and orchestrators | D5.2 |
| Security Policies and breaches responses | D5.4 |
| Guidelines to standardise metadata templates | D3.2 |
| Assessment of FAIRness maturity levels | D3.2 |
| Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance | D4.1 |
| Incentive models recommendations | D4.3 |
| Recommendation on technical infrastructure | D5.3 |
| Maturity Model associated with data analysis needs | D5.5 |

Table 2. HealthyCloud functionality outcomes

| Functionality recommendations | Deliverable |
|---|--------------------|
| Expected users' interactions | D6.1 |
| (datasets/datahubs-collections) Metadata catalogue | D6.2 |
| Metadata contribution | D6.2 |
| Computational resources infrastructures catalogue | D6.2 |
| Guidance and knowledge hub | D6.2 |
| Data access scenario 1 (DA-S1). The portal acts as a “one-stop-shop”, gathering all the different data hubs and providing links to them. | D6.3 |
| Data access scenario 2 (DA-S1). In addition to the functionality described DA-S1, the portal will provide a description and comparison of data access conditions of each of the hubs. | D6.3 |
| Data access scenario 3 (DA-S3). This scenario adds a new feature to the DA-S2, making possible dataset search. | D6.3 |
| Data access scenario 4. The portal will provide a single access form that the user fills and goes directly to the hubs, which continue with the rest of the process. | D6.3 |
| Data access scenario 5. In this case the portal, besides providing the form, acts as facilitator during the whole process, being an intermediary between the users and the data hubs. | D6.3 |
| Data access scenario 6. The portal facilitates actual authorization and access to the data, which could only incorporate hubs with controllership over their data. Two options are considered: (a) the portal manages authorizations and access, facilitating/supporting the data access when it is granted; (b) the portal is somehow involved in the access process (e.g., supports expedite access). | D6.3 |

Table 3. HealthyCloud legal outcomes

| Legal recommendations | Deliverable |
|---|--------------------|
| First draft on legal framework for technical safeguards with a focus on cloud usage | D2.1 |
| Framework of modular contract clauses for HRICs | D2.2 |
| Report on DPA and EC Workshop | D2.3 |
| Guideline on ELSI compliant governance models | D2.4 |

5 Results

This section includes the gap analysis results, how the use cases challenges have been covered by the project outcomes. For each challenge, the following subsections include the information about from which category (knowledge, functionality and legal) have been addressed. Not all the points of view apply to all the challenges, these cases are identified as “N/A” (not applicable). The missing gaps will appear as empty cells in the tables in each section.

5.1 Gaps in user profiles

The first gap that has been found was identified when the process was described. In the user profiles defined by WP6 there was no user profile for “*Computational Resource Infrastructure*”. This is one of the roles that we need to include the interaction between the researchers and the other HRIC ecosystem stakeholders.

5.2 Gaps related to identify datasets

All the challenges related to the identification of datasets to be reused have been addressed from all the three outcomes’ categories points of view.

The two challenges related to the need of using several data providers infrastructures (Table 4, first two rows) have been addressed from the functional (WP6) and legal (WP2) perspective. According to the proposed specifications of the FAIR Health Data portal, the future HRIC will provide an initial search-point functionality where researchers can start the process of identifying all the datasets regardless of the number of data hubs and data collections involved. For the legal category, areas of improvement include the implementation of an ELSI Metadata standard, considering capabilities for supporting data subjects in exercising their GDPR rights. Principal compliance challenges, however, lie beyond the specifications of the underlying technical infrastructures, especially when it is unclear under which legal framework the controller operates. Individual data collections (WP3, data collections) and data-hubs (WP4) cannot address these challenges.

The challenges related to the difficulty of understanding the available data are addressed for all categories (Table 4, last two rows). According to the proposed specifications of the FAIR Health Data portal (WP6), the portal will include a metadata meta-catalogue while data collections (WP3) and data hubs (WP4) address these challenges for knowledge category analysing the FAIRness levels and providing guidelines to improve the FAIRness of the data, and for legal category ensuring that requirements for data collection and use are met, obtaining appropriate consent where necessary and anonymizing or pseudonymising data appropriately.

Table 4. Addressed challenges to identify datasets

| | Knowledge | Functionality | Legal |
|--|-----------|---------------|-------|
| Several data sources | N/A | ✓ | ✓ |
| European/National/Regional level | N/A | ✓ | ✓ |
| Missing common metadata catalogue | ✓ | ✓ | ✓ |
| Uncertainty/unknown data requirements | ✓ | ✓ | ✓ |

5.3 Gaps related to identifying computational resources

Both challenges are covered from two points of view (Table 5). WP5 provides analysis and recommendations to improve the findability of the computational resources. According to the proposed specifications of the FAIR Health Data portal (WP6) and its extension, for going beyond data resources, the portal will include a computational infrastructure catalogue. These challenges do not imply legal issues.

Table 5. Addressed challenges to identify computational resources

| | Knowledge | Functionality | Legal |
|---|-----------|---------------|-------|
| Minimum computational resources at client premises (analysis performed at cohort premises in federated scenario) | ✓ | ✓ | N/A |
| Uncertainty/unknown computation requirements | ✓ | ✓ | N/A |

5.4 Gaps related to understand request access procedure

These challenges need to be addressed from two different procedures, the one to request access to data and the second one to request access to computational resources.

All the challenges related to the data request access are addressed completely (Table 6).

The only challenge that is not addressed for the legal category is that the information is not easy to find, it is not addressed because the current legal frame does not oblige to this. WP4 provides recommendations for data hub governance, including request protocols that could be used to make these processes and their requirements more homogeneous and recommending inclusion on the data hubs' website of a section describing the used data governance model among others. WP6 addresses these challenges including some functionality related to data access in some of the considered scenarios, including some suggesting HRIC as a centralised point for requesting data access. For the legal category,

understanding request access to data has legal implications because some activities should include some technical and organisational safeguards because of the sensitivity of the data, WP2 includes the legal requirements related to this activity for the first two challenges.

Table 6. Addressed challenges to understand data request access procedure

| | Knowledge | Functionality | Legal |
|---|-----------|---------------|-------|
| Heterogeneous processes (no common request access protocol) | ✓ | ✓ | ✓ |
| Heterogeneous requirements (for example, the project proposal content or if the data can leave the facility). Each centre has their rules. | ✓ | ✓ | ✓ |
| Request process information is not easy to find. | ✓ | ✓ | N/A |

The challenges related to request access to computational resources are all covered, but only from some of the viewpoints (Table 7). WP5 provides the knowledge to address all of them, and WP6 has included an infrastructure computational resources catalogue that will support the researchers to know where to find the information. Clearly, understanding the protocols on how to request access to computational resources do not have legal implications.

Table 7. Addressed challenges to understand computational resources request access procedure

| | Knowledge | Functionality | Legal |
|---|-----------|---------------|-------|
| Heterogeneous processes (no common request access protocol) | ✓ | | N/A |
| Heterogeneous requirements (for example, the project proposal content or if the data can leave the facility). Each centre has their rules. | ✓ | | N/A |
| Request process information is not easy to find. | ✓ | ✓ | N/A |

5.5 Gaps related to request access

Like for the previous challenges, these ones need to be addressed from two different procedures, the one to request access to data and the second one to request access to computational resources.

The challenges related to request access to data are all covered for the knowledge category (Table 8). WP3 and WP4 include guidelines and recommendations that cover both challenges, three of the data access scenarios defined by WP6 consider HRIC as a single access form that the user fills and goes directly to the hubs addressing the challenges. For the legal category, WP2 provides legal requirements related to this process.

Table 8. Addressed challenges to request data to a data infrastructure

| | Knowledge | Functionality | Legal |
|---------------------------------------|-----------|---------------|-------|
| Detailed study protocol needed | ✓ | ✓ | ✓ |
| Sensitive data issues | ✓ | ✓ | ✓ |

There is one challenge related to request computational resources addressed, only for the knowledge category (Table 9). WP5 addressed the activities behind the sensitive data management (e.g., moving data in a federated scenario). WP6 addresses the data access request, but computational access request is out of the scope of the envisaged functionality. This is the reason HRIC considers the option of being a single access for data request, but not for computational resources.

Table 9. Addressed challenges to request services to a computational resources infrastructure

| | Knowledge | Functionality | Legal |
|---------------------------------------|-----------|---------------|-------|
| Detailed study protocol needed | | | N/A |
| Sensitive data issues | ✓ | N/A | N/A |

5.6 Gaps related to gain access to datasets

This activity is related to downloading the data (once the access is granted) by the researcher or by the trusted research environment before running the analysis. The main challenge is related to the security (and privacy) of the data (Table 10). It is covered for the three categories. From the data collections and data hubs, WP3 provides guidelines to standardising metadata templates and assessing the maturity level of FAIRness, WP4 addresses this challenge in the data hub by studying patterns of governance of selected data hubs and providing data governance recommendations. From computational infrastructure, WP5 provides security analysis. One of the data access scenarios defined by WP6 includes the functionality of HRIC that facilitates actual authorization and access to the data.

Table 10. Addressed challenges to provision access to datasets

| | Knowledge | Functionality | Legal |
|----------------------|-----------|---------------|-------|
| Data security | ✓ | ✓ | ✓ |

5.7 Gaps related to integrate data

Integration of data comes with several challenges, all covered for the knowledge category (Table 11). WP3 addresses all the challenges including guidelines and recommendations for the activities related to the challenges, emphasising the importance of data harmonisation and de-identification. WP6 includes in the HRIC a guidance and knowledge hub to expose this kind of information for some of the challenges (data collected with different properties like time spans and update periodicity; semantic interoperability). Addressing these challenges for the functionality category is out of the scope of HRIC. Challenges directly related to the data integration itself do not have legal implications. WP5 addressed the activities behind these challenges (e.g., mobilising data in a federated scenario) also for the knowledge category in their recommendations.

Table 11. Addressed challenges to integrate data

| | Knowledge | Functionality | Legal |
|--|-----------|---------------|-------|
| Data collected with different properties, like time spans and with different update periodicity | ✓ | ✓ | ✓ |
| Structured and/or unstructured individual data level | ✓ | ✓ | N/A |
| Linkage limitations with multiple levels of aggregation | ✓ | N/A | ✓ |
| Data harmonisation | ✓ | N/A | ✓ |
| Variable mapping (manually) | ✓ | N/A | N/A |

5.8 Gaps related to set-up analysis process

These challenges are related to the software to be used for the analysis, they have been addressed for knowledge and functionality (Table 12). WP5 included recommendations as part of their software analysis and WP6 included software guidelines as part of the knowledge hub, which is part of the proposed specifications of the FAIR Health Data portal. These challenges do not have legal implications.

Table 12. Addressed challenges to set-up analysis process

| | Knowledge | Functionality | Legal |
|---|-----------|---------------|-------|
| Analysis software quality (code quality, documentation) | ✓ | ✓ | N/A |
| Analysis software accessibility (source code, containerization deployment) | ✓ | ✓ | N/A |

5.9 Gaps related to perform analysis

All the challenges are addressed for the knowledge category and most of them for the legal category (Table 13). For the knowledge category, WP3 and WP4 FAIR-related outcomes address them from the data management point of view (metadata), and WP5 with the maturity model address them from the computational point of view. For the functionality category, three of the six data access scenarios, defined by WP6, address the security and privacy communication. Cross-border data should not be addressed for the functionality category because it is related to the legal issues.

Table 13.Addressed challenges to request services to perform analysis

| | Knowledge | Functionality | Legal |
|---|-----------|---------------|-------|
| Cross-border data (legal issues) | ✓ | N/A | ✓ |
| Scalability issues | ✓ | N/A | ✓ |
| Guarantee security and privacy communication | ✓ | ✓ | ✓ |
| Replicability | ✓ | N/A | N/A |
| No common follow-up patient records | ✓ | N/A | N/A |

5.10 Gaps related to understand data submission process

No challenges were identified related to understanding the data submission process.

5.11 Gaps related to data submission request

All WPs have considered the sensitiveness of the data in all their results in the context of submitting it for future (re-)use (Table 14).

Table 14.Addressed challenges to data submission request

| | Knowledge | Functionality | Legal |
|------------------------------|-----------|---------------|-------|
| Sensitive data issues | ✓ | ✓ | ✓ |

6 Conclusion

The final objective of the WP7 is the validation of the HealtyCloud outcomes driven by two use cases (cancer and atrial fibrillation). We conducted a gap analysis to evaluate if the project outcomes would support the challenges that researchers face during the activities performed in the execution of their research projects. We conducted this gap analysis in close collaboration with all the work packages. Indeed, this collaboration was implemented inviting some WPs to the internal WP7 regular meetings and organising a total of three workshops for the evaluation process. The workshops have been complemented with some communication to gain extra details for streamlining the validation workshops results.

Project use cases have identified a set 23 challenges that the researchers often face when they execute projects based on the secondary use and reuse of health-related data for research purposes. To have a better understanding of the health-related research projects, we characterised this process and classified the 23 challenges according to the activity in the process they are related to.

The health-related research project process has a total of 11 activities, it starts with the definition of the research question and finishes when the researchers submit their results, which may be reused in the future, going through data discovery, access request, data and computational environment preparation, and data provision and use.

The results of the gap analysis are detailed in Section 2.5. We have classified how the WPs outcomes are addressing the challenges from three categories: *knowledge* (guidelines and recommendations), *functionality* (concrete services included in the HRIC ecosystem) and *legal* (addressing the legal and ethical requirements for the responsible use of health-related data for research). From the 11 activities, there are two activities without associated challenges, the first one is related to “*defining the research question*”, which is in the scope of the concrete research projects. For the second one, the use cases do not identify challenges “*to understand the data submission process*”. For the rest of activities, most of them (7 of 9) have all the associated challenges covered from the three categories. The activities that do not have all their associated challenges addressed from the three points of view are all related to computational resources: “*to understand request access procedure*” and “*to request access*”. Two of them are not addressed for the functionality category and one for the knowledge category. There is also one gap that is not related to the challenges, we found a missing user profile in the set of user profiles definition during the health-related research project process characterisation, the one characterising the computational infrastructure provider stakeholder.

Table 15 details the identified gaps across the different proposed activities.

Table 15. Identified gaps during the validation

| Issues not addressed | Gap |
|---|---------------|
| Computational infrastructure profile user profile | Missing |
| Heterogeneous processes and requirements when understanding computational request access procedures | Functionality |
| Detailed study protocol needed when the computational resources are requested | Functionality |
| Sensitive data issues when the computational resources are requested | Knowledge |

7 Next steps

This is the last of the three deliverables of WP7, WP7 brings two real world use-cases centred on research questions related to cancer and atrial fibrillation. These use cases requirements, reported in D7.1 and D7.2, have been used by work packages (from WP2 to WP6) to propose specific solutions. This third deliverable reports the results of the WP proposed solutions validation. This deliverable aims to gather useful information regarding the project proposed outcomes, and potential challenges that haven't been addressed in the framework of the project, and, therefore, that will need to be incorporated into future versions of the HRIC Strategic Agenda.

Regarding future steps, to be addressed beyond the HealthyCloud project, this deliverable explicitly provides some known unaddressed challenges. Additionally, to these challenges, we have identified some other topics that were out of the scope of this evaluation but are worth mentioning to be considered as future research lines.

In the health-related process used to articulate the evaluation, the activities associated with the data management plan were not explicitly included. Further exploration related to data management is expected to lead to new challenges, for example regarding the turn down of computational resources at the end of analysis as well as data and results retention.

This report includes some challenges related to federated scenarios, for example *Minimum computational resources at researcher premises* or *Guarantee security/privacy communication to minimise any security/privacy breach* associated with the activities *identifying computational resources* and *performing analysis*. We also considered the federated scenario when the data is analysed in the data holder premises, in this case we judged that the data security issues are minimised, but not having access to the data (only to the results) can bring some specific challenges that need further exploration. For instance, it usually also involves a trusted third party that vouches, among others, for data quality.

The results of the validations confirm that the access to health-related data is much more regulated than access for computational capabilities. We could anticipate that the broad adoption of Trusted Research Environments (TRE) for processing transnationally sensitive data would represent a change in this situation.

Project Deliverables

This section includes the list of deliverables mentioned in this document.

- [D2.1] Adrian Thorogood. (2023). Framework of modular contract clauses for HRICs. Zenodo. <https://doi.org/10.5281/zenodo.7871247>.
- [D2.2] Framework of modular contract clauses for HRICs. Available online: <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5f0b6b7b5&appId=PPGMS> (last access: 24/07/2023).
- [D2.3] Report on DPA and EC Workshop.
- [D2.4] Guideline on ELSI compliant governance models. Available online: <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5fd07ac44&appId=PPGMS> (last access: 24/07/2023).
- [D3.1] Shona Cosgrove, Pascal Derycke, & Irene Kesisoglou. (2023). Landscape analysis of FAIRness levels of health-related data using catalogue matrix. Zenodo. <https://doi.org/10.5281/zenodo.7871308>.
- [D3.2] Celia Alvarez-Romero, & Alicia Martínez-García. (2023). Guidelines to standardise metadata templates and assessment of FAIRness maturity levels. Zenodo. <https://doi.org/10.5281/zenodo.7848253>.
- [D3.3] Landscape analysis using a health related-data catalogue matrix.
- [D4.1] Alicia Martínez-García, & Celia Alvarez Romero. (2023). Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance. Zenodo. <https://doi.org/10.5281/zenodo.7871375>.
- [D4.2] Alicia Martínez-García, Celia Alvarez-Romero, & Silvia Rodríguez Mejías. (2023). Report on current discoverability solutions and FAIR adoption level. Zenodo. <https://doi.org/10.5281/zenodo.7875071>.
- [D4.3] Report on the usage and auditing data hubs indicators and proposal of an incentive system.
- [D5.1] Analysis of existing computational infrastructures models including ELSI.
- [D5.2] Rosa M Badia. (2023). Analysis of existing orchestration mechanisms for distributed computational analyses including a general overview to facilitate new developments. Zenodo. <https://doi.org/10.5281/zenodo.7875243>.
- [D5.3] Guidelines to establish sustainable computational infrastructures for the future HRIC ecosystem. Available online: <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5f47ec036&appId=PPGMS> (last access: 24/07/2023).
- [D5.4] A study of existing site security policies for sensitive data and protocols for responding to breaches.
- [D5.5] Reference guidelines for the establishment of an ethically sound and legal compliant health data research ecosystem.
- [D6.1] Laura Portell-Silva, Salvador Capella-Gutiérrez, & Lidia Lopez. (2023). FAIR Health Data Portal expected users' interactions (1.0). Zenodo. <https://doi.org/10.5281/zenodo.7949977>.

- [D6.2] Specifications for the FAIR data portal. Available online: <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5f47eecb1&appId=PPGMS> (last access: 24/07/2023).
- [D6.3] Specifications for data access. Available online: <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5fc7d3664&appId=PPGMS> (last access: 24/07/2023).
- [D7.1] Functional requirement analysis report of Use Case 1. Available online: https://healthycloud.eu/wp-content/uploads/2022/07/D7.1_FunctionalAnalysis_Structure-Cancer.Submitted.25052027-1-1.pdf (last access: 24/07/2023).
- [D7.2] Functional requirement analysis report of Use Case 2. Available online: https://healthycloud.eu/wp-content/uploads/2023/04/D7.2_FunctionalAnalysis_Structure.pdf (last access: 24/07/2023).

Acronyms and Abbreviations

- BPMN – Business Process Modeling Notation
- D – deliverable
- ELSI - Ethical, Legal and Societal Implications
- FAIR (principles) - Findable, Accessible, Interoperable, Reusable data
- HRIC – Health Research and Innovation Cloud
- HyC - HealthyCloud
- TOC - Table of Content
- TRE - Trusted Research Environment
- WP – Work Package