PHIRI
Population Health Information Research Infrastructure

Report on scalability, sustainability, and rapid cycle analysis requirements

Deliverable 6.6 31.10.2023

Francisco Estupiñán-Romero
Javier González-Galindo
Natalia Martínez-Lizaga
Juan González-García
Carlos Tellería-Orriols
Martin Thissen
Enrique Bernal-Delgado

IACS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101018317

Table of Contents

1. Executive Summary

This report, learning from the implementation of the WP6 use cases, aims to reflect on how scalable the design and implementation of the PHIRI research infrastructure has been, and discuss elements around sustainability, specifically the implementation of rapid cycle analysis.

This document stems from the experience of implementation of the COVID-19 use cases in WP6, highlighting the strengths and limitations of the federated analysis approach to produce evidence responding to relevant population health policy questions; it largely discusses implications on scalability and sustainability, and finally, provides recommendation for an eventual continuation of the PHIRI infrastructure improving the infrastructure.

With regard to scalability, the document assesses the conditions of participation in the use cases as well as provides insight on how to increase data hubs participation, considering technical, organisational, and ethical aspects. We have also tried to identify the challenges and opportunities for scaling up the federated analysis approach, such as data quality, interoperability, security, privacy, governance and coordination, communication, and evaluation of the user experience within PHIRI.

Secondly, we have assessed the capacity and capability barriers of the data hubs to participate in the use cases, and identified key factors for sustainability as resources, infrastructure, skills, incentives, policies, regulation, and stakeholders' engagement.

Thirdly, we have examined the challenges for conducting rapid cycle analysis in terms of the health data life cycle from generation to availability for analysis, such as timeliness, completeness, accuracy, relevance, reliability, comparability, and consistency over time.

Finally, we have summarised the main implications from all the previous reflections, and provided recommendations for future research carried out upon the baseline achievements of the PHIRI federated research infrastructure.

2. Glossary

You can check other definitions across this document in the PHIRI Glossary at https://www.phiri.eu/glossary

Scalability	For the purposes of this report, scalability refers to the ability to handle increasing amounts of health data from varied sources and increasing the number of data hubs participating in a use case without compromising the performance or the quality of the analyses[1].
Sustainability	For the purposes of this report, sustainability refers to the ability to maintain and update the data infrastructure and the analytical framework over time and across different contexts, and to the ability to prototype and implement new use cases upon the demand of new policy relevant questions[2].
Rapid-cycle analysis	For the purposes of this report, rapid cycle analysis refers to the ability to produce timely and actionable results that can inform policy decisions in response to emerging challenges and to the sustained capability to update results of a use case upon data refreshment without compromising the internal validity of the analysis, thus providing continuously updated evidence to guide policy decisions (also known as near real-time analysis)[3].

3. Introduction

Population Health Information Research Infrastructure (PHIRI) aims to facilitate and support open, interconnected, and data-driven research by sharing cross country COVID-19 population health information, and exchanging best practices related to data collection, curation, processing, use, and reuse following Ethical, Legal, and Social (ELSI) Issues and Findability, Accessibility, Interoperability, and Reusability (FAIR) principles. It has the objective to: 1) provide a Health Information Portal/HIP for COVID-19 with FAIR catalogues on health and health care data, 2) provide structured exchange between countries on COVID-19 best practices and expertise, and 3) promote interoperability and tackle health information inequalities.

The aim of PHIRI Work Package (WP) 6 is to demonstrate how a broad variety of data (e.g., administrative and survey data) can be reused in a distributed way across Europe. WP6 looks at COVID-19 impacts in specific subgroups by conducting research through real life use cases of immediate relevance. Furthermore, these use cases represent pilot activities supporting the development of a health data research infrastructure mobilising data from different European countries to inform health policy.

The purpose of the use cases has been to answer specific population research questions relevant to public health or health policy. Each research question was translated to a suitable (common) data model specifying all data and information requirements and defining a semantic framework for the common understanding and comparison of each of the participating partners. Finally, the data model supports the generation of synthetic datasets enabling the development, testing of the methods and algorithms for the analyses and the extraction, transformation, and loading of the required data. In this federated research infrastructure, partner institutions (i.e., research groups) participating in any of the use cases were responsible for managing their data under their own system level security and privacy policies. The federated deployment of the research questions implied that no data leaves the participant partners' instances, following the 'data visiting' principle; therefore, bringing the (local) analyses to each of the partners as a reproducible analytical pipeline (PHIRI app) ready to be deployed and run to reliable produce the expected outputs enabling international comparison.

WP6 has established the governance mechanism and implemented the solutions to respond to four use cases carried out in multiple sites (local nodes at country/region level) plus a pilot study on the continuous surveillance on the pandemic evolution based on the rapid cycle analysis and publication of individual-level public health data:

1. Use case A: Impact of the COVID-19 on health care in more vulnerable populations, completed by Austria (AT), Croatia (HR), Finland (FI), Italy (IT), Aragon (ES), and Wales (UK)
2. Use case B: COVID-19 related delayed care in breast cancer patients, completed by Belgium (BE), Marche (IT), Latvia (LV), Aragon (ES), and Wales (UK)
3. Use case C: Effects of the COVID-19 pandemic on maternal and newborns health completed by Euro-Peristat Network (27 countries and 4 UK nations)
4. Use case D: COVID-19 related changes in mental health care, completed by Austria (AT), Croatia (HR), Finland (FI), Romania (RO), Aragon (ES), and Wales (UK)

The full operative federated research infrastructure has demonstrated the suitability of the federated approach in producing accurate and timely outputs for a rapid policy response on COVID-19.

The implementation of the use cases was formalised in the specification of several common data models, the agreement on data management process and the co-design and development of reproducible analytical pipelines concerning both the local analysis within each participant partners' premises and the general analyses for international comparison.

Brief description of the use cases

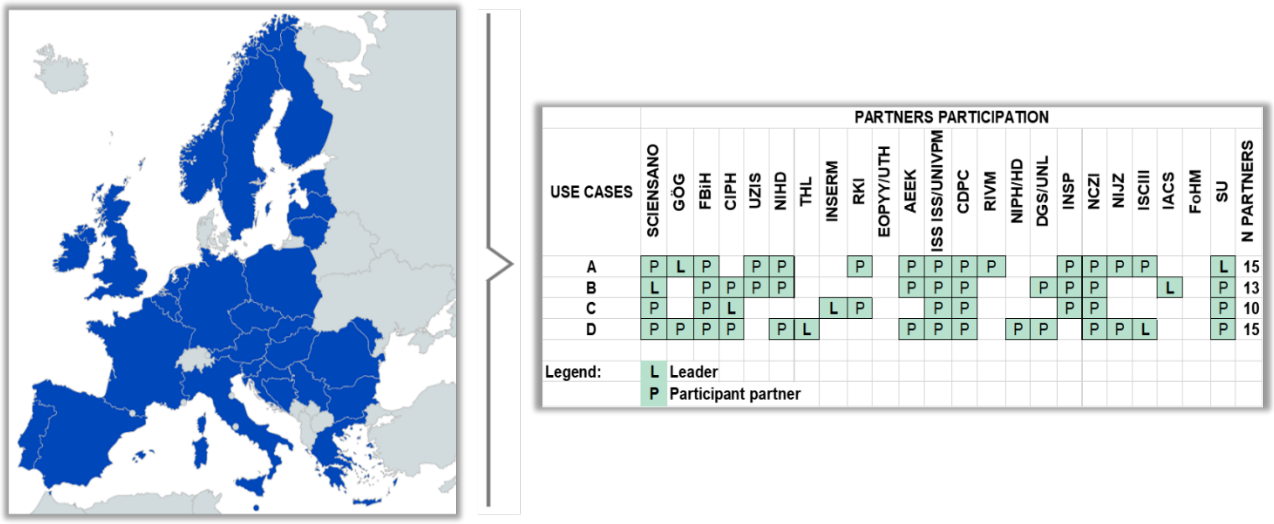
a) Use case A Vulnerable populations, inequalities, and risk factors with direct or indirect impact on healthcare outcomes during the COVID-19 pandemic: To clarify whether healthcare utilization patterns in vulnerable populations vary between settings and over time and are linked to the COVID-19 epidemiological situation, using individual-level health record, administrative and research data combined with ecological/group level contextual data
b) Use case B COVID-19-related delayed care in breast cancer patients: Aimed to demonstrate whether there has been an increase in surgical and co-adjuvant (i.e., radiotherapy, chemotherapy, hormonotherapy, and immunotherapy) treatment delay because of the COVID-19 crisis in eligible women diagnosed with breast cancer using individual-level health records, administrative and research data combined with ecological/group level contextual data.
c) Use case C Effects of the COVID-19 pandemic on maternal and newborns health: Investigating the pandemic's direct (infection by SARS-CoV-2) and indirect effects on perinatal health using routine population birth data and assess whether effects differ by socioeconomic context.
d) Use case D COVID-19 related changes in population mental health: Measured changes in population mental health associated with the COVID-19 pandemic. It aimed to demonstrate whether there has been an increase in healthcare utilization of mental health treatments because of the COVID-19 crisis in eligible patients diagnosed of depression and anxiety using individual-level health record, administrative and research data combined with ecological/group level contextual data.
e) Demonstration pilot for rapid-cycle federated analysis. Piloting the federated analyses of the evolution of several indicators for the monitoring and surveillance of the COVID-19 pandemic (7-day reproductive number, 7- and 14-day incidence rates, regular and ICU admissions, bed occupancy rates) and 7- and 14-day predictions, using individual data from several data sources -COVID-19 monitoring and surveillance registries, hospital data, and administrative data.

Data hubs participation in WP6 use cases (roles and responsibilities)

Depending on the use case, five to twenty-seven countries could produce actual evidence on policy-relevant questions based on the reuse of real-world health data routinely collected by the Health Systems. Their participation involved applying for access to sensitive individual-level health data, securely accessing and managing it, and running the analytical pipelines within the data hubs' premises (i.e., public health and research institutions participating in PHIRI as partners) to produce aggregated data, data quality reports and local analysis reports with valuable information to inform policy. Some institutions – particularly in Use Case C, collaborating with the Euro-Peristat network – were willing and able to implement the federated approach even out of the scope of their participation in PHIRI without having any resource assigned from the project.

The WP6 partners were interested in different use cases at the start of the project. However, only some WP6 partners could participate in the use cases they wanted. The main reason was that they needed more data to answer the research question. Sometimes, the data was not available or accessible in time due to multiple factors (i.e., difficult to overcome procedural or organizational barriers to access the data, lack of technical support within their organizations to extract the data,

data fragmentation across many organizations not participating in the project, or other). Other times, the data was not complete, reliable, or good enough to use. *Figure 1* below shows how many WP6 partners participated in each use case.



	PARTNERS PARTICIPATION																					
USE CASES	SCIENSANO	GÖG	FBiH	CIPH	UZIS	NIHD	THL	INSERM	RKI	EOPYY/UTH	AEEK	ISS ISS/UNIVPM	CDPC	RIVM	NIPH/HD	DGS/UNL	INSP	NCZI	NIJZ	ISCIII	IACS	FoHM	SU	N PARTNERS
A	P	L	P		P	P			P		P	P	P	P			P	P	P	P			L	15
B	L		P	P	P	P					P	P	P			P	P	P			L		P	13
C	P		P	L				L	P			P	P				P	P					P	10
D	P	P	P	P		P	L				P	P	P		P	P		P	P	L			P	15
Legend:	L	Leader																						
	P	Participant partner																						

Figure 1. Partners initial commitment to participate by use case

The design and implementation of each use case was led by a partner responsible for coordinating the actions required to achieve the expected results, collect the local outputs from the participants, and generate the final report on the international comparative analysis. Use cases leaders were supported in the development of their statistical analyses, the implementation of their analytical pipeline by a technical team (WP7) leading the development and implementation of the PHIRI federated research infrastructure (PHIRI app) in collaboration with the Developers Forum in WP7. In addition, the partners participating in each of the different use cases counted with the support and assistance of a Help Desk team dedicated assist them. The Help Desk team clarified partners' doubts about data and information requirements and data model specifications of the use cases, assisted partners in deploying the reproducible analytical pipelines to run the analyses and help solve any technical issue or code bug found in the implementation of the statistical analyses or the development of the PHIRI app.

Use case leaders were responsible for promoting the research question and producing a first specification of the common data model with the definition of the data requirements to respond to the proposed question. Both the common data model and the statistical analyses were co-developed through many iterations by all partners participating in each use case under the coordination of the use case leaders.

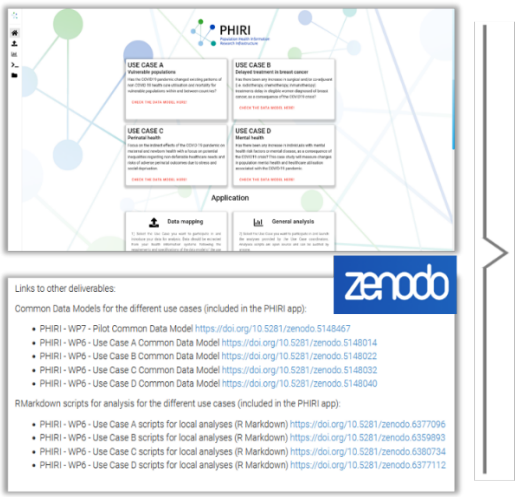
Upon use case leaders made a first version of the common data model specification available to the data hubs - partners participating in each use case; the WP6 coordination launched two surveys in a row. The first survey aimed to check general data availability and IT capabilities in the data hubs. It included questions about persons of contact as domain and IT experts in each data hub, health data access application processes, health data availability and institutional IT environments and tools and technical capabilities and resources. The second survey, tailored to each use case data model specification, asked questions regarding the availability of the required health data complying with the definitions of the common data model. Use case leaders relied on the information provided by partners to guide and coordinate their participation, while partners were able to assess their capabilities and potential participation based on their compliance with the data requirements.

On the other hand, partners or other institutions (i.e., data hubs) participating in each use case were responsible for participating in the improvement of the common data model specification and the

development of the statistical analyses in a collaborative iterative process. Each data hub was also responsible to apply for data access following their system level security and privacy policies and in accordance with their ethical and legal frameworks. Data hubs were also responsible for managing their data and for deploying the PHIRI app to run the analytical pipeline corresponding to the use cases they were participating. Furthermore, the data hubs were responsible for checking the outputs of the local analyses and sharing the aggregated data with the use case leader for them to complete the international comparison analyses, helping to interpret the comparative results once the final analyses were completed.

The PHIRI federated research infrastructure was supported on the development of reproducible analytical pipelines including all technical solutions and components required to reliable produce the outputs required for the international comparison based on the local analyses. The PHIRI app achieved this by providing a Docker solution including all documentation and scripts to manage and analyse the data wrapped around a practical user interface that dealt with all dependencies and technical requirements. All components of the PHIRI app were developed using open-source software solutions and were available to the research community via publication on the Zenodo EU repository.

Although most data hubs were able to test the deployment of the PHIRI app, running the local analyses (*at least 14 data hubs some participating in multiple use cases - see Figure 2 below*); the data hubs could also download the scripts for the statistical analysis and run them in their systems. However, they faced the disadvantages of having to deal with dependency issues derived from the requirement of multiple libraries to implement complex statistical analyses, not being able to use the preferred user interface and having to run separately the data quality and exploratory data analyses accompanying each use case.



USE CASES	SCIENSANO	GÖG	FBiH	CIPH	UZIS	NIHD	THL	INSERM	RKI	EOPYY/UTH	AEEK	ISS ISS/UNIVPM	CDPC	RIVM	NIPH/HD	DGS/UNL	INSP	NCZI	NIJZ	ISCIII	IACS	FoHM	SU	N PARTNERS
	PARTNERS PHIRI-APP DEPLOYMENT																							
A, B, C, D	D	D		D	D	D	D					D	D	D		D			D		D	D	D	14
		*					*	*															*	

Legend:
D PHIRI-app Docker deployed and tested (or RMarkdown scripts tested)
* Partners that have made the commitment to test the solution in their systems (WP7 Developers' Forum)

Figure 2. Relation of data hubs that could deploy and test the PHIRI app.

Other institutions out of the scope of the PHIRI project were also able to download, deploy and test the PHIRI app completing the local analyses of the different use cases using synthetic data provided as part of the documentation of the common data models specification.

The final step on the implementation of each of the use cases consisted in collecting the aggregated data produced as output of the local analysis in each data hub. Those data were then meta-analysed to produce an overall report on the international comparison of participant countries. Those reports aimed to answer the research question with relevant insight informing EU policy on the impact of COVID-19 in population health.

In parallel to the implementation of the use cases in WP6, the IT experts of the data hubs participated in the Developers' Forum in WP7. They also contributed to scripting and reviewing the open-source code for the statistical analyses.

Finally, the Help Desk team offered support to both use case leaders and data hubs during the whole process to ensure the success of the use cases in producing the expected results.

Summary of the Help Desk activity

The Help Desk team solved **45 tickets** during the PHIRI project. Each ticket was a task related to any of the use cases or the PHIRI app (i.e., update the data model, update the scripts, or new version or correction of the PHIRI app, or other) that the team had to complete due to a partners' request for assistance or a bug report. The team had **278 conversations** (each conversation/thread including a few emails totalling more than 900 emails) **and 27 meetings** (at least 1 hour) with the partners.

Outputs of the WP6 use cases

Up to date, WP6 and WP7 have published 13 releases in Zenodo –some with multiple versions- related to both the use cases and the PHIRI app (*see Table 1*).

Release (digital object)	Version	Unique views	Downloads
PHIRI APP - WP7 - PHIRI Federated Research Infrastructure (FRI) - D7.2 Mid-size prototype of PHIRI federated research infrastructure (10 versions published)	v.2.3.0	**2,467**	**265**
PHIRI - WP7 - Pilot Common Data Model (1 version published)	v.1.0.0	**266**	**45**
PHIRI - WP7 - Pilot scripts for local analyses (R Markdown) (1 version published)	v.1.0.0	**66**	**76**
PHIRI - WP7 - Pilot outputs from the local analyses (interactive reports) (1 version published)	v.1.0.0	**35**	**30**
PHIRI - WP6 - Use Case A Common Data Model (2 versions published)	v.3.0.0	**602**	**147**
PHIRI - WP6 - Use Case A scripts for local analyses (R Markdown) (4 versions published)	v.1.1.2	**417**	**494(294*)**
PHIRI - WP6 - Use Case B Common Data Model (2 versions published)	v.2.0.0	**407**	**79**
PHIRI - WP6 - Use Case B scripts for local analyses (R Markdown) (2 versions published)	v.4.0.1	**231**	**240(165*)**
PHIRI - WP6 - Use Case B outputs from the local analyses (interactive reports) (2 versions published)	v.1.0.1	**177**	**53**
PHIRI - WP6 - Use Case C Common Data Model & Study Protocol (3 versions published)	v.2.0.1	**454**	**100**
PHIRI - WP6 - Use Case C scripts for local analyses (R Markdown) (2 versions published)	v.2.0.1	**186**	**207(130*)**
PHIRI - WP6 - Use Case D Common Data Model (2 versions published)	v.2.0.0	**379**	**94**
PHIRI - WP6 - Use Case D scripts for local analyses (R Markdown) (3 versions published)	v.1.1.1	**285**	**343(203*)**

*Table 1. Digital objects published in Zenodo produced by use cases in WP6 and WP7 (stats on unique views and downloads were provided by Zenodo - last checked 20th, September 2023). *Unique downloads for the current version.*

The results of each use case will be available as part of an academic publication as a Special Issue of the European Journal of Public Health at the end of the PHIRI project.

Use cases in WP6 also aimed to demonstrate the implementation of the PHIRI federated research infrastructure, testing, and informing on challenges and possible recommendations to improve the scalability, sustainability and rapid cycle analysis capabilities of the proposed framework based on the effective reuse of sensible health data.

The next section of this report will briefly discuss how these requirements were met and what challenges were encountered in the project.

3. Insights on Scalability, Sustainability and Rapid Analysis capabilities of the Data Hubs

In this chapter we address questions on scalability, sustainability, and rapid analysis capabilities that the PHIRI approach would face in case of continuation. For this purpose, we stem from the information compiled during the project, using several sources; a) participants' responses to WP6 data availability and IT capabilities survey; b) use cases' common data model availability surveys; c) the WP7 IT infrastructure characterization and user experience survey; and, d) WP7 Developers' Forum meetings, while main findings come from the interaction between the data hubs and use case leaders with the Help Desk while implementing their use cases.

For the purposes of this report, **scalability** refers to the ability to handle increasing amounts of health data from varied sources and increasing the number of data hubs participating in a use case without compromising the performance or the quality of the analyses. **Sustainability** refers to the ability to maintain and update the data infrastructure and the analytical framework over time and across different contexts, and to the ability to prototype and implement new use cases upon the demand of new policy relevant questions. **Rapid cycle analysis** refers to the ability to produce timely and actionable results that can inform policy decisions in response to emerging challenges and to the sustained capability to update results of a use case upon data refreshment without compromising the internal validity of the analysis thus providing continuously updated evidence to guide policy decisions.

The next sections of this report will discuss challenges encountered in the implementation of the use cases and how these requirements were met during the project.

Scalability challenges

For the purposes of this report, scalability refers to the ability to handle increasing amounts of health data from varied sources and increasing the number of data hubs participating in a use case without compromising the performance or the quality of the analyses.

Challenges in scaling up the use cases can be classified in terms of interoperability issues following the interoperability dimensions within the European Interoperability Framework (EIF)[4]. Main challenges envisaged during the implementation of the use cases were related to:

a) lack of specification of the research question leading to a poorly or unspecified common data model (*Semantic Interoperability*),
b) lack of reliable sources of the health data required to respond the research question (*Organisational Interoperability*),
c) lack of access to the relevant data within the stewardship of the participant institution (*Organisational Interoperability*),
d) difficulties identifying the institution holding the required data within the participant country (data hub) (*Organisational Interoperability*),

e) lack of quality assurance policies both at the Health System, and at the data stewardship institution managing the reuse of the health data for secondary purposes such as research (*Organisational and Semantic Interoperability*),
f) difficulties accessing relevant information on the data application processes and the governance of health data within the data hubs – *this information being essential to ensure the feasibility, quality and validity of the analyses, as well as the compliance with ethical and legal requirements* (*Legal and Organisational Interoperability*),
g) difficulties accessing the required technical and IT expert human resources within the data hubs to assist in the deployment of the technological solutions (*Organisational and Technological Interoperability*),
h) lack of technical and IT capabilities to contribute to the development or support the deployment of the analytical pipelines (*Organisational and Technological Interoperability*),
i) lack of trust, training and difficulties in accessing and using open-source solutions to develop statistical analysis in a research context (*Organisational and Technological Interoperability*).

In this section, an in depth description of the steps given during the project to enable the PHIRI federated infrastructure to be scalable.

The design and implementation of the PHIRI federated research infrastructure have been developed upon a) the data visiting principle –*data does not move but code moves*; b) the orchestration of the research question throughout a workflow that ensured legal, organisational, semantic, and technological interoperability; and c) a *master-worker* federated computational architecture that supported the development of the use cases. Thus, overcoming the limitation for the reuse of sensitive health data and providing a methodology to achieve interoperability between data hubs and research nodes.

The PHIRI methodology heavily relies on the willing collaboration of both domain and IT experts with sufficient knowledge on the intricacies of their data and their institutional policies to gain responsibility on dealing with overcoming the organisational barriers that could exist ensuring their participation on the use case of interest. Incentives are on the data hubs to gain insight both in their own data provided the relevant research questions and on their comparative situation in relation to their peers also participating in the study via their participation in EU/International research initiatives.

Both WP6 and WP7 collaborated tackling these challenges during the project by hosting multiple use case meetings led by WP6 coordination and use case leaders, both for domain experts (i.e., population health researchers, public health officers, and healthcare organisations) and for IT experts contributing to the development of the infrastructure via the Developers' Forum.

In addition, setting up the Help Desk from the beginning of the project enabled use case leaders to query and discuss on the implementation of the PHIRI workflow and data hubs to resolve their questions regarding the data model and data requirements, the implementation requirements and sorting out bugs or other user or technical issues during the deployment of the analyses.

Furthermore, PHIRI developments based on *privacy and security by design* principles[5] implementing data minimization and data visiting as fundamental building blocks of the governance of the use cases and conditioning participation upon strict compliance with the stepwise workflow enabled partners to apply for access to the relevant data, even from other institutions (acting as data hubs). At the same time, use cases focused on reproducibility, following the literate programming and open-source principles (*i.e., documentation-first*) enabled script auditing and technical interoperability with any system used by the participant institutions lowering technological barriers derived from lack of resources, training, or IT capabilities, and building trust among participant IT experts, incentivizing their contribution to the final implementation of the analytical pipelines.

In addition, the PHIRI app was conceived to serve a variety of users who are not necessarily technologically versed but with expert knowledge of their health data sources and a fair understanding of the research questions posed; therefore, much consideration was given during design to provide a user interface easing the experience of the participants in running the analyses after delegating in them the responsibility on accessing, querying and transforming their data to the common data model for each use case. This user interface was designed to sequentially guide the user into mapping the input files (*CSV, pipe-separated files in UTF-8 format without BOM*), run the analyses, and access the outputs once generated while seamlessly conducting syntactic checks between the input data and the expected data model, launching standardised data quality analyses and validating data checking rules predefined in the common data models, producing local reports both on data quality and on the local results, and producing informative error logs when failing to proceed with any of the steps described.

A demo operational version of the PHIRI app interface can be accessed at https://phiri.iacs.es/ and tested using the synthetic datasets included (*Figure 3*). It can also be accessed via Use Case B demonstrator within the PHIRI Health Information Portal (HIP).

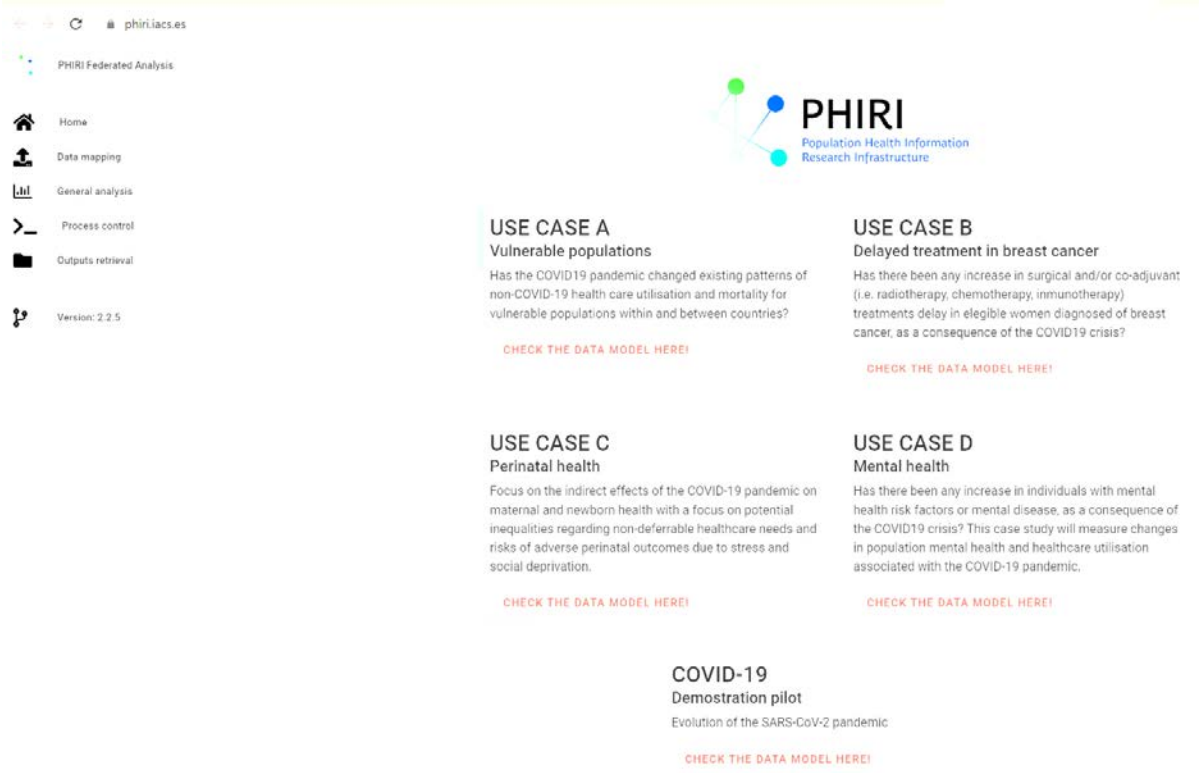


Figure 3. Home screen of the PHIRI app user interface presenting the use cases and providing access to their data models and the contextual menu to the analytical pipeline.

As a result of the reproducibility, privacy and security concerns the PHIRI app is downloadable and can be completely audited. It also can be deployed and run in air-gapped systems without requiring inbound or outbound connections to any external source, thus eliminating the possibility of system breaching or data leakage. In addition, the container technology enabled developers serving all needed components and the environment of the analytical pipeline packaged in a way that minimised systems' requirements considering the efficient use of partners' computer resources. The latest release of the PHIRI app can be accessed in Zenodo at 10.5281/zenodo.5729310.

All these design and development considerations and constraints meant that coordination, evaluation of the performance of the PHIRI app and communication of the outputs were left to human

interaction between the use case leaders, the participant data hubs, and the Help Desk, all that have dedicated resources within the project.

The strategy behind PHIRI federated research infrastructure development followed a pragmatic approach establishing the principles guiding the implementation of the use cases and the technological solutions to easily prototype, test, deploy and run any statistical analysis based on the querying, extraction, and transformation of health data in accordance with the specifications of the data model. This pragmatic approach entailed building up participants' capabilities by enabling them to participate in their preferred use cases, considering they had access to the relevant data, while demonstrating the feasibility of the approach as the first step. This strategy also implied successive capacity and trust building between participant data hubs to facilitate their leadership in prototyping new use cases following the same methodology; thus, introducing the incentive of leveraging the PHIRI federation to respond to their own policy questions.

The aim of this stepwise pragmatic approach was for PHIRI to end up enabling secure programmatic communications between the more technically advanced data hubs after constituting a peer network in which any of the data hubs would be capable of design, promote and develop their own use case to be implemented by PHIRI following the same workflow.

Once legal, organisational, semantic, and technical interoperability issues were solved in the implementation of the use cases, PHIRI efforts focused on progressing on the secure communications ladder toward a federated infrastructure enabling continuous collaboration among peers with similar research, data management and technical capabilities (*Figure 4*).

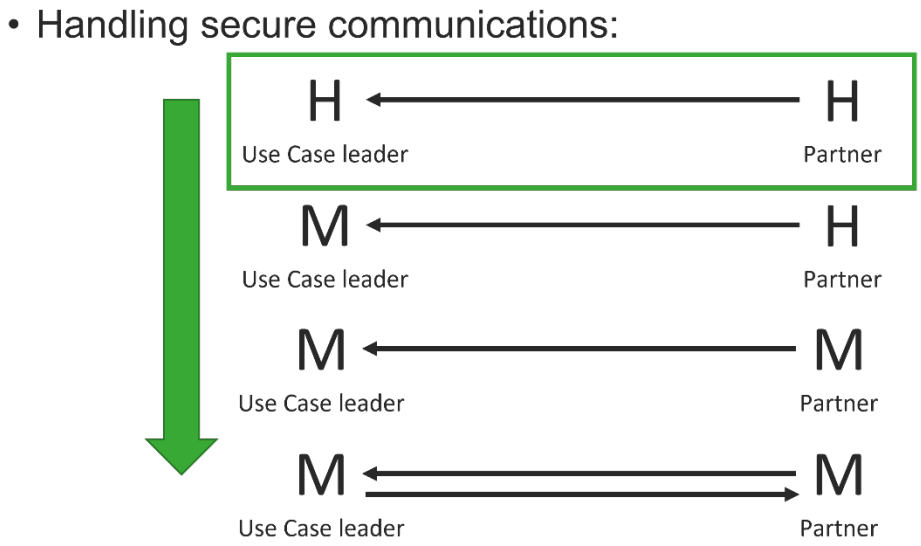


Figure 4. Stepwise approach to a Federated Research Infrastructure: secure communication from Human-to-Human (H-H) interaction (i.e., master-servant), to Machine-to-Machine (M2M) bilateral interaction (i.e., peer-to-peer) should be assured at each step of the ladder. In particular, machine-to-machine interactions requires the technical implementation of secure peer-to-peer communications including AAI and operational systems.

Going "*down*" the ladder entails introducing automation within the workflow between the institution leading or coordinating the use case and participants which can be done under a similar level of supervision of the human-to-human options but that requires securing communications between data hubs.

In order to scale both in terms of number and heterogeneity of the questions being responded, number of data hubs contributing to respond the questions or data hubs continuous and timely respond to research questions in a synchronous way, machine-to-machine secure

communications is required as it enables automation managing via authorization, authentication and identification, and data hubs profiling to manage the complexity behind the coordination of multiple data hubs, responding or contributing in other capacities to multiple research questions in data hubs networks with different structures. Achieving full machine-to-machine communications either synchronously or asynchronously is also a requirement to deploy truly distributed algorithms aimed at analysing available individual-level data at scale following the data visiting principle via differential privacy among the nodes without the need for producing and meta-analysing aggregated data. In time, these algorithms could achieve better performance for certain statistical analyses requiring access to raw data or with high computational requirements that could not be implemented otherwise (i.e., medical image analysis at scale in multiple sites).

Some steps have been advanced during PHIRI in providing Human-to-Machine secure communications to automate the meta-analysis and report of the latest results of a use case. The Use Case B demonstrator within the PHIRI Health Information Portal (HIP) provides a completely reproducible report that automatically updates with the latest aggregated data produced by the local analyses conducted by participant data hubs and uploaded through using either an API (application programming interface) or a simple web user interface. Further steps have been tested within the scope of the WP7 but have remained as development versions.

Sustainability challenges

For the purposes of this report, sustainability refers to the ability to maintain and update the data infrastructure and the analytical framework over time and across different contexts, and to the ability to prototype and implement new use cases upon the demand of new policy relevant questions.

Beyond scalability, a necessary condition for sustainability, the main challenges for the sustainability of a Federated Data Research Infrastructure such as PHIRI can be classified in terms of governance/policy issues, financing, human capacity, and resource capabilities and stakeholders' engagement. Main challenges envisaged during the implementation of the use cases were:

a) misalignment of research and policy incentives to participate in EU research initiatives (Governance/Policy),
b) lack of understanding of EU and national regulation on the basis for legitimate uses of sensitive data such as individual-level health data (Governance/Policy and Stakeholders' engagement),
c) difficulty on financing research projects based on the reuse of health data (i.e., secondary use of data for research instead of primary generation and use of data for research) (Financing and Governance/Policy),
d) uneven data hubs maturity level in terms of the existence, implementation, and documentation of system level information security and privacy policies, data quality assurance systems and researcher capacitation in data management and statistical analysis (Governance/Policy and 'Human capacity and computational resources capabilities)
e) issues with data availability and conditions for data access (Governance/Policy and Stakeholders' engagement),
f) lack of researchers capabilities in terms of data management, data engineering and statistical analysis (Human capabilities),
g) lack of or difficulty to accessing appropriate computational resource and IT experts' availability (Human capacity and computational resource capabilities),
h) lack of research impact or devolution of research results to society (Stakeholders' engagement and Governance/Policy)

Along PHIRI life, most sustainability challenges have been related to misalignment between incentives for the institutions potentially participating as data hubs or research nodes.

Currently, the main incentive to institutions leading access and reuse of health data for research is participating in EU financed research projects aimed at responding to research questions that require international comparison or a large-enough or heterogeneous enough population sample to require access to millions of registries to be feasible to respond. This type of research questions could be classified as population health questions and usually require access to population level data such as electronic health records, medical records, or population registries of an entire population. PHIRI use cases are practical examples of this kind of population health research questions.

Sustaining such a research infrastructure requires standardisation, interoperability and responsiveness nationally and internationally to support urgent, efficient and trustworthy access to data, streamlining the coordination of the multiplicity of governance mechanisms currently in place in the different data hubs, directly linking with EC efforts towards a EU regulatory framework for the secondary use of health data – HealthData@EU[6].

HealthData@EU regulatory proposal aims to tackle several of the sustainability issues commented above via the development of the regulatory proposal through the future approval of both delegating and implementing acts. In this context, the sustainability of PHIRI data hubs and PHIRI functioning will be dependent on their alignment with the HealthData@EU developments and implementation. Furthermore, data hubs can be sustainable and even enhance their capabilities as part of the PHIRI federated health data research infrastructure if they are able to provide services for other data hubs to be able to participate in the HealthData@EU and for researchers to leverage the experience produced by PHIRI in implementing their use cases and achieving and improving science.

Rapid cycle analysis requirements

For the purposes of this report, rapid cycle analysis refers to the ability to produce timely and actionable results that can inform policy decisions in response to emerging challenges and to the sustained capability to update results of a use case upon data refreshment without compromising the internal validity of the analysis thus providing continuously updated evidence to guide policy decisions. Thus, rapid cycle analysis is paramount in terms of sustainability.

Rapid cycle analysis differs from traditional data analysis in terms of timeliness and tightness between the refresh of data and the update of the results or recommendations to decision makers. Rapid cycle analysis aims to produce results in a short timeframe, such as days or weeks, rather than months or years. It also aims at using the most recent and relevant data available, rather than relying on outdated or incomplete data. Finally, rapid cycle analysis is intended to keep stakeholders timely informed based on the best available evidence to help them decide.

Rapid cycle analysis may involve multiple steps such as a) posing a valid research question that can be answered with data and is relevant to decision makers or stakeholders; b) implementing the analytical pipeline to collect/access, process, analyse, and visualise the relevant data; c) testing, refining, and retesting the analytical pipeline to ensure its validity, reliability, and accuracy; d) assessing the results of the analysis and providing feedback and recommendations to the decision makers or stakeholders; and, e) repeating the cycle as new data becomes available or new questions arise.

Then, main requirements for rapid cycle analysis can be considered in two levels. There are requirements linked to developing and testing the implementation of research questions as use cases (i.e., prototyping), and there is another level of requirements to manage the continuous access to the relevant data sources to update results as soon as new information becomes available.

Requirements for prototyping use cases are like those described for the scalability of the use cases and are related to imposing a strict workflow that standardises the specification of the use case definition and sets tight feedback loops both in the design of the use case and during its implementation. These requirements are fulfilled by providing the means and incentives for both domain and technical experts of those data hubs participating with data to close the gaps between decision makers and stakeholders and the evidence they require to inform those decisions. That means setting a framework for the co-development of the use cases with a documentation-first approach, the implementation of open-science practices, and iterative development towards a minimum-viable product that can provide valid evidence in a timely manner while setting the basis for further development and the continuous update of results as new information becomes available.

Requirements for the continuous update of results once a use case is implemented are more like those described for the sustainability of the use cases and are related to the governance of the data and the collaboration of the institutions providing continuous access to up-to-date relevant data and the dedication of their resource to produce timely updated evidence monitoring whichever result of interest while relevant.

Achieving rapid cycle analysis requires a data research infrastructure able to tighten the schedules for a data product (i.e., dashboard, model, etc.) from idea to production-ready. This requires automating any step of the process that can be managed at scale. In the context of a federated data research infrastructure as PHIRI, it means implementing synchronous machine-to-machine secure communications across the data hub networks and building capacity in each node within the network to produce and implement their use cases as peers.

4. Recommendations

The list of recommendations for PHIRI scalability, sustainability and rapid cycle analysis can be organised into four categories: governance and policy, financing, human and computational capacity, and stakeholder engagement.

Governance and policy recommendations

- Establishing clear and consistent regulations and guidelines on the legitimate use of health data for research purposes, respecting the ethical and legal principles of data protection, privacy, and consent – aligning with HealthData@EU developments.
- Aligning the research and policy incentives to participate in EU research initiatives, ensuring the relevance, impact, and dissemination of the research results to inform decision-making and improve health outcomes.
- Developing and implementing data quality assurance policies and standards for health data collection, management, analysis, and reporting, ensuring the validity, reliability, and comparability of the data and the results across different contexts.
- Enhancing data availability and accessibility by establishing common data models, interoperability frameworks, and metadata repositories.

Financing recommendations

- Securing adequate and sustainable funding for the development, maintenance, and operation of the federated data research infrastructure from different sources, such as EU funds, national funds, private funds, or public-private partnerships.

- Exploring innovative financing models and mechanisms that can support the federated data research infrastructure, such as provision of evaluation services supporting performance-based contracts, value-based contracts, or social impact bonds.
- Optimising the use of existing resources and infrastructures by leveraging synergies, complementarities, and economies of scale among different partners and initiatives.
- Monitoring and evaluating the costs and benefits of the federated data research infrastructure using appropriate indicators and methods.

Human and computational capacity recommendations

- Building and strengthening the human capacity for data management, data engineering, statistical analysis, and other relevant skills among researchers, data stewards, IT experts, and other staff involved in the federated data research infrastructure.
- Providing training and support for the use of open source solutions to develop statistical analysis in a research context, as well as for accessing and using the analytical pipelines developed by the federated data research infrastructure.
- Ensuring adequate and reliable computational resources for storing, processing, analysing, visualising, and sharing large amounts of health data in a secure and efficient way.
- Adopting innovative technologies and tools that can enhance the capabilities of the federated data research infrastructure.

Stakeholder engagement recommendations

- Involving relevant stakeholders from different sectors and levels in the design, implementation, evaluation, and dissemination of the federated data research infrastructure, ensuring their needs, expectations, and feedback are taken into account.
- Raising awareness and understanding of the benefits and challenges of using health data for research purposes among different stakeholders, such as researchers, policy makers, health professionals, patients, and the public.
- Promoting trust and collaboration among different stakeholders by ensuring transparency, accountability, and participation in the federated data research infrastructure.
- Communicating and disseminating the research results and impact of the federated data research infrastructure to different audiences using appropriate channels and formats.

5. Conclusions

The extensive and continuous reuse of sensitive health data (e.g., clinical data, electronic health or clinical records, administrative and claims data) could enhance the role of population health research on public policy decisions.

Nowadays, sensitive health data reuse is still scarce. Common arguments to explain this paucity are multiple: data privacy and safety issues; the difficulty to discover data sources of value; complex data application processes and data access rules; uneven data quality; or limited IT experts and computational capacities.

PHIRI has successfully supported the development of use cases overcoming limitations for the reuse of sensitive health data in a federated manner and providing a methodology to achieve interoperability in multiple research nodes. By minimising data movement and keeping data in its original location, the infrastructure has effectively overcome the limitations for the reuse of sensitive health data enabling international comparisons across EU while complying with legal and ethical provision on the matter.

PHIRI has paved the way for a scalable research infrastructure that could eventually close the gap between research outputs and decision-making; large-scale developments of PHIRI will depend on progressively transforming the current *master-worker* architecture in a peer-to-peer network where nodes are able to assume co-leadership and methodological and technological roles.

The sustainability of the achievements so far will depend very much on the data governance, the financing and the human and computational capabilities improvement at network nodes' level, and on the alignment with the development of the future HealthData@EU initiative.

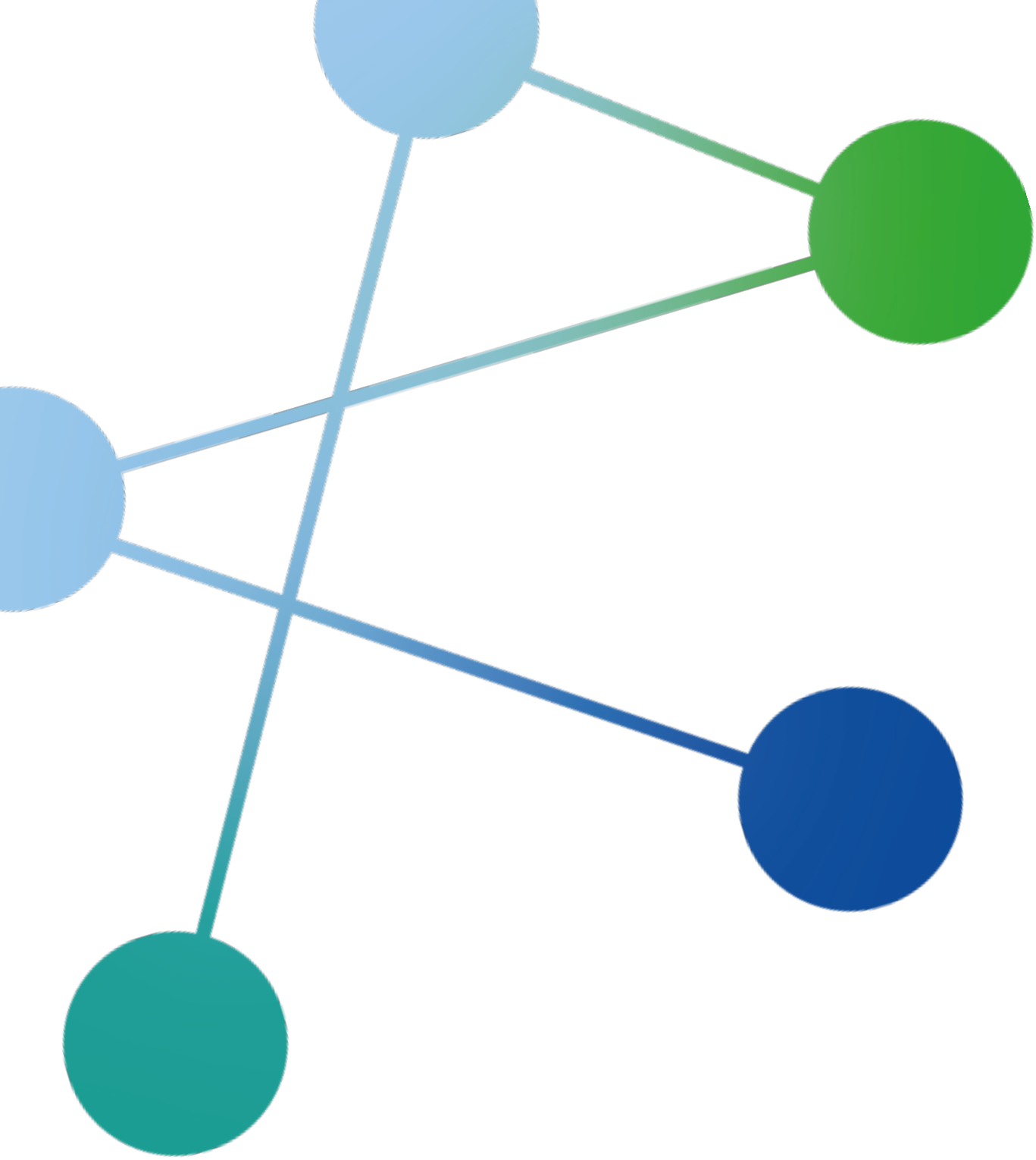
6. References

1. Wikipedia contributors. (2023, August 24). Scalability. Wikipedia. https://en.wikipedia.org/wiki/Scalability
2. Sustainable development. (2023, July 6). Trade. https://policy.trade.ec.europa.eu/development-and-sustainability/sustainable-development_en
3. González-García, J., Estupiñán-Romero, F., Tellería-Orriols, C., González-Galindo, J., Palmieri, L., Faragalli, A., Pristās, I., Vuković, J., Misinš, J., Zile, I., Bernal-Delgado, E., & InfAct Joint Action consortium (2021). Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct. Archives of public health = Archives belges de sante publique, 79(1), 221. https://doi.org/10.1186/s13690-021-00731-z
4. De Ganck, A. (2017, April 21). The new European Interoperability Framework - ISA2 - European Commission. ISA2 - European Commission. https://ec.europa.eu/isa2/eif_en/
5. Wikipedia contributors. (2023b, October 25). Privacy by design. Wikipedia. https://en.wikipedia.org/wiki/Privacy_by_design
6. European Health Data Space. (2023, September 27). Public Health. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

7. Disclaimer

Disclaimer excluding Agency and Commission responsibility.

The content of this document represents the views of the author only and is his/her sole responsibility. The European Research Executive Agency (REA) and the European Commission are not responsible for any use that may be made of the information it contains.



IACS
Centro de Investigación Biomédica de Aragón (CIBA)
Avda. San Juan Bosco, 13, planta 0
50009 Zaragoza, Spain
Ebernal.iacs@aragon.es
www.phiri.eu
@PHIRI4EU

© 2023| published by IACS