PHIRI

Population Health Information Research Infrastructure

DATA MANAGEMENT PLAN

Deliverable 1.1, 30.05.2021

Enrique Bernal Delgado, Juan Gonzalez Garcia, Francisco Estupiñan Romero, Natalia Martinez Lizaga, Ramon Launa Garces

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. INTRODUCTION.

This document provides the PHIRI Data Management Plan (DMP) which according to the Description of Action is to be 1st delivered in Month 6 and then updated periodically. The DMP outlines what kind of data will be collected or generated and how it will be handled, processed and shared. It describes the standards that will be incorporated and the related methodology for data collection, processing and data sharing. This deliverable is based on the template and the guidelines provided by the European Commission (see Section 7 on issues to cover in your Horizon 2020 DMP).

PHIRI will provide large-scale, integrated and sustainable services to population health research community. Thus, PHIRI will provide the validation of different federated research solutions (options for an **e-infrastructure**) upon the experience of a number of **use cases** of immediate relevance in the context of COVID-19 that will focus on selected aspects of (A) vulnerable population groups and risk factors, (B) delayed medical care in cancer, (C) perinatal health outcomes, and (D) mental health outcomes, and a demonstration pilot aiming to test rapid cycle analyses.

Finally, the development of the research solutions out of the use cases and the demonstration pilot (thus, aggregated data and metadata, logic data model, the open source analytical tools, and the actual research outputs) will be included in the PHIRI Portal, a web based **Health Information portal (i.e. PHIRI portal)** which will act as the interface between the research infrastructure and the potential users.

2. DATA SUMMARY.

2.1 Purpose of the data collection/generation and the relation to the objectives of the project

PHIRI for COVID-19 has the following objectives:

- To set up a Health Information portal for SARS-CoV-2 pandemic, including FAIR catalogues on health and health care data for structured information exchange across European countries. This portal will facilitate access to and use of COVID-19 relevant population health (health status and determinants of health) and care data provided by EU countries' public institutions represented in the PHIRI applicant consortium. It also provides the services and tools necessary for researchers to link different data sources and to use Pan-European data in a GDPR compliant, federated way.
- To promote interoperability and tackle health information inequalities.

PHIRI is expected to generate or collect: (1) data from surveys conducted to address health and population health research needs in relation to COVID-19 (WP5, WP8,WP9); (2) meta-reports on existing population health research on COVID-19 in Europe (WP4); (3) metadata from national data sources that could eventually serve as data origins for research purposes (WP4); (4) pseudonymised individual data, aggregated data and metadata, collected and/or produced in the use cases and the demonstration pilot (WP6, WP7); (5) open source analytical tools generated by PHIRI (WP7) or collected from other RIs; (6) the different documents produced as intermediate or final achievements of PHIRI; (7) scientific manuscripts produced as a consequence of the project; and (8) any relevant link to repositories that are collecting data on the SARS-COV2 pandemic.

2.2 Overall data management and data sharing architecture

The two pillars upon which PHIRI builds, the PHIRI Portal and the e-infrastructure, will imply different data management requirements.

The <u>PHIRI FAIR portal</u> will be centrally managed by the PHIRI coordinating institution (Sciensano). The Portal is conceived as an entry point where users will have access to the content described in the previous section (section 2.1). Notably, this includes meta-data describing in an interoperable manner regional, national and international data sources hosted by data institutions within the MSs; likewise meta-data from pan-European domain specific Research Networks. In addition, the portal will include tools (e.g. data models, analytical pipelines…) meant to be reused by the users of the Portal.

When it comes to the <u>e-Infrastructure</u>, PHIRI will propose options for the reuse of massive real-world data (i.e., observational in nature and retrieved from the actual contacts of a patient with the health system) that will develop on a *privacy-by-design* approach. The Governance in a secure-by-design and privacy-by-design approach will rely on both architectural and software decisions that have been taken beforehand.

Architectural elements: As opposed to a centralized architecture, PHIRI is conceived as a federated infrastructure, that will build on a distributed architecture (see figure in section 4 of this DMP) where exchange of individual level data between the **data hubs** (i.e., the partners) does not exist. Thus, data remain in-site, under the responsibility, governance procedures and security assurance of the data hubs, being fully compliant with the national legislations, the administrative provisions within each country, and the data access procedures implemented for research purposes in each participant institution.

In this federated infrastructure, the data hubs (i.e., the partners) will be in charge of implementing the analytical pipeline developed on their own pseudonymised data and producing local results that will be gathered in a central coordination hub (i.e., individual data are not gathered but partial outputs consisting on analytical reports, error logs or aggregated data resulting from the analysis). The coordination hub will be part of the PHIRI portal services hosted in the IACS premises.

These architectural elements will be tested in WP6 via four use cases and in WP7 via a demonstration pilot.

Software elements: The computational solution will consist of a set of analytical pipelines programmed in R and Python, underpinning the aforementioned privacy and safety by design approach. Analytical pipelines will be contained in a Docker container to ease its deployment on partner's premises, following the distributed architecture previously described. The Docker containers will be deployed in a private docker hub available as part of the PHIRI Health Information Portal services. IACS will host the Docker containers and analytical pipelines during their development managing system configuration, version control and coordinating version release. .

2.3 Data sources and size requirements

For the purposes of the data management plan, it is worth differentiating the collection of meta-data catalogues (WP4) from the data exchange produced in the use cases (WP6) and demonstration pilot (WP7).

For WP4 catalogues will contain metadata from data sources from national or regional institutions that collect and/or curate data on the pandemic. Likewise, will also contain metadata from data sources (registries, collections and statistics) that collect and curate data for EU and international comparison. Finally, will add metadata from population and care data sources collected and/or curated by international or national research initiatives with implication in the pandemic and its consequences.

In WP6, the DMP entails the different data motion elements constituting the development of four use cases. So, 1) in premises, extraction-transformation and loading (ETL) of pseudonymized data using the data model and meta-data provisions for the specific use case; 2) storing the subset of data in a server within the premises of the data hub (e.g., national or regional institution); 3) implementation of the analytical pipeline developed elsewhere by the coordination hub in each use case; 4) submission of the output of the analytical pipeline back to the coordination hub (i.e., aggregated data and/or summary results depending on the use case); 5) meta-analysis of those outputs in the coordination hub; and 6) publication of the scripts and outputs in a FAIR way within the Portal in WP4.

The type of data in the previous paragraph will vary from each case to use case being typical data sources administrative data, clinical information from electronic health records, population data, and survey data.

More specifically, the expected four **use cases** and corresponding data sources will be:

Use Case A: Direct and indirect determinants of COVID-19 infection and outcomes in vulnerable population groups with reference to inequalities **Data sources** (preliminary): individual level health record, administrative and research data, ECDC surveillance data. **Participants (preliminary):** SU, GÖG, Sciensano, RKI, AEEK, ISS, INSP, ISCIII/ENS, CDPC, RIVM, UZIS, NIHD, NCZI, NIJZ.

Use Case B: COVID-19 related delayed care in breast cancer patients. **Main sources** (preliminary): individual level data from electronic health records at hospital and primary care, pharmaceutical bills, administrative information. **Participants** (preliminary): IACS, Sciensano, SU, AEEK, DGS/UNL, INSP, CDPC, CIPH, UZIS, NIHD, NCZI, ISS.

Use Case C: Effects of the COVID-19 pandemic on maternal and newborn health. **Data sources** (preliminary): Aggregated data from routine sources on population birth data (birth registers, hospital discharge data, vital statistics) in the 31 countries in the Euro-Peristat network. **Participants** (preliminary): potential PHIRI countries (Sciensano, RKI, ISS, SU, INSP, CDPC, NCZI), potential Euro-Peristat countries (Belgium, Croatia, Cyprus, Estonia, Finland, France, Germany, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Spain, UK-Scotland, United Kingdom).

Use Case D: COVID-19 related changes in population mental health. **Main sources** (preliminary): aggregated data from surveillance systems, hospitals, mental health centres, mortality data, surveys. **Participants** (preliminary): ISCIII, THL, Sciensano, DGS/UNL, HD/NIPH, UNIVPM, SU, AEEK, GÖG, CDPC, CIPH, NIJZ, NIHD, NCZI.

In WP7, the demonstration pilot, as part of the e-infrastructure, will use aggregated data collected daily about COVID-19 cases as part of the surveillance and monitoring of the pandemic. Main sources (preliminary) will be individual-level information collected daily in the COVID-19 registries, and electronic health records as well as the administrative information for those cases in the registry. Participants in this pilot will be (preliminary) SU, GÖG, Sciensano, FBiH, CIPH, ÁEEK, CDPC, THL, IACS.

Size requirements

For the PHIRI catalogues including the description of the metadata as detail above size requirements are low, preliminary estimated below 500 MBs.

In the development of the uses cases in WP6 and the overarching demonstration pilot of a federated infrastructure (WP7) only scripts and the outputs will be shared and thus preliminary estimated size requirements will be below 1 Gb.

When it comes to the options for a federated research e-infrastructure task (Task 7.5) depending on the data hub capacities it might be necessary to install a 2Tb server to deal with ETL procedures, deploy the docker solution and store the outputs.

Finally, high computational and storage capacity will be analysed for each options. To this end, the need of linkage with high performance computing centres or, public or private clouds will be explored, in particular European e-infrastructure (i.e., EGI Federation, EUDATA, GEANT, TESTA-ng, EuroHPC Joint Undertaking, etc.).

2.4 Data description, Data extraction and data storage

Metadata catalogues: PHIRI´s portal will share metadata catalogues on public health and healthcare information and methodologies identified at national and international level. The main tool to provide the necessary information is the network of national nodes, based on JA InfAct framework, and international stakeholders of interest (e.g., ECDC, JRC, EUROSTAT). The metadata catalogue will be created in 2 steps. In a first step, meta templates will be created organising and describing national and international data sources available. In a second step, through a web search, national COVID-19 data portals will be searched to pre-fill the metadata templates.

Use cases: will be using either anonymised data (data for which re-identification is made impossible with current “state of the art” technology), or pseudonymised, partially de-identified, data or aggregated data. For such pseudonymised data, GDPR applies and appropriate compliance must be ensured. ETL processes will depend on the specific data sources of each use case (see Section 2.3) and ETL system organization and procedural at each data hub.

Demonstration pilot (e-Infrastructure): will use pseudonymised individual-level information collected daily in the COVID-19 registries as well as the administrative information for those cases in the registry.

Within the PHIRI federated infrastructure, analytical pipelines will be developed that operate locally on participating data hubs. No individual level data will be transferred between Consortium Partners or to the coordination hub, but only aggregated summary results (e.g. summary coefficients) that are obtained for the specific local hub.

Data storage

Portal All the contents (see above) published in the Health Information Portal for COVID-19 will be stored in Sciensano premises, and will using their own technical support.

e-infrastructure In accordance to the architectural design of the distributed PHIRI e-infrastructure and its technological solutions, individual patient’s data will not be copied to a central repository, but kept stored on partners’ premises where the storage and analyses will take place. Thus, data storage management will follow the partners and data hubs own regulations. The metadata catalogues, aggregated data from analyses and the results out of the use cases and the demonstration pilot will be stored in a secured server of Sciensano with a copy in IACS.

2.5 Data utility

There is a need for a structured European mechanism for COVID-19 data exchange that increases the intelligence on an early response to the pandemic but also allows the analysis of the unintended consequences of the reorganization of the services on non-COVID-19 and chronic patients, and the assessment of the mid-term and long-term impact of the crisis, especially in those most vulnerable populations.

Using the PHIRI infrastructure, researchers will be able to respond to these relevant questions and provide insight of value to the decision-making process made by citizens, clinicians, public health practitioners and policy makers.

3. MAKING DATA FINDABLE, ACCESSIBLE, INTEROPERABLE AND RESUSABLE [FAIR DATA]

The PHIRI portal builds on the FAIR principles. All the contents will be findable, accessible, interoperable and reusable within the PHIRI Portal.

3.1 Data modelling, harmonisation and integration

The data modelling, harmonization and integration mechanisms will be tested and deployed in the formalization of the use cases in WP6 and the demonstration pilot in WP7.

All the use cases and the demonstration pilot aim at first developing a data model that allows the research on the topic of interest (see above use cases details). Data hubs will have to identify those relevant data sources and the attributes and variables of interest and will extract the required data, transform those data in accordance to the data model and store them in their premises. The transformation of data will follow the prescriptions of the data model while the coordination hub will provide assistance when needed to harmonize case definitions and variable selection across different encoding systems. . Finally, the coordination hub will provide data hubs with analytical scripts to run the analyses and submit the outputs back where outputs will be integrated.

The development of solutions to deploy or establish a federated infrastructure will be FAIR and available in the PHIRI Portal, including: a) the development of a common data model for a COVID-19 rapid response; b) the design and deployment of the required data transformation and encoding processes; and, c) the implementation of the distributed analytical solutions, in particular the intermediate processes and final research outputs.

3.2 Data Cataloguing and Persistent Identifiers (PID)

As aforementioned, the PHIRI portal for COVID-19 will catalogue metadata from data sources from national or regional institutions that collect and/or curate data on the pandemic, metadata from data sources (registries, collections and statistics) that collect and curate data for EU and international comparison, and metadata from population and care data sources collected and/or curated by international or national research initiatives with implication in the pandemic and its consequences.

PHIRI will develop a catalogue service indexing available datasets with a persistent, unique identifier or PID. . PHIRI portal will use international standards of publication to facilitate findability and easy browsing and retrieval. The resulting catalogue will, in any case, be browsable by advanced semantic-enabled engines and interfaces.

3.3 Accessibility and Data sharing

The PHIRI portal will act as the single access point for all the contents produced in the project (see above). Sciensano, as PHIRI coordinator and Portal hosting institution, will establish procedures for the inclusion, maintenance and access to the contents.

In any case, PHIRI portal will be complying with the FAIR principles and Open Data provisions, e.g. released under open licenses such as CC-BY 4.0. Source code of the PHIRI portal architecture will be publicly released in a code repository, e.g. GitHub. In turn, the data, metadata or tools collected from other institutions, research networks or data infrastructures will have a link to the original sites and be respectful with the data management provisions of the owners. In the case of the meta-data catalogues, PHIRI will provide with the re-use, sharing and correct citation/crediting of specific subsets of the meta-data catalogues ensuring compliance with Open Science.

In the case of use cases, datasets will be available as per the policies of the respective data curators. Due to data sensibility (health domain) and data granularity, curators might have policies including opt-out clauses from open access repositories or practices.. The software solutions developed in the use cases (for the Common Data Model, the ETL processes and analytical pipelines, and reporting solutions) will be published in public open access repository (e.g. ZENODO, Github). Finally, aggregated datasets will be included in research data repositories and open access infrastructures such as OpenAIRE and ZENODO.

4 ALLOCATION OF RESOURCES, RESPONSABILITIES

During PHIRI, the costs of curation and preservation are related to the secure storage of data and results produced in the surveys and use cases, the associated costs and actual costs of maintenance of the portal and of the project website.

Data and results produced as a consequence of surveys or uses cases will be stored in a secured server of Sciensano with a copy in IACS. Both are covered as overheads in their budgets.

The portal technical set-up and ICT support for the Health Information portal for COVID-19 (task. 2.3) is led by Sciensano. Running costs for this activity are budgeted in WP2 until the end of the action.

When it comes to the project website, running costs (uploading and maintenance of the information included) are covered by the budget until the end of the action.

As part of the PHIRI federated research infrastructure, WP7 will establish the upgrade options on how and what data will be kept in the coordination hub and for how long as part of the RI after the project finalization. As detailed in WP7, it is expected that the federation will delegate the storage, curation and maintenance costs on the hubs in the federation, usually, public institutions that have their own budgets devoted to these operations. These governance aspects will be further explored as part of WP4 on the legal and ethical requirements of population health data sharing, and in WP6 where use cases will provide specific insight.

Sustainability

The proposed Population Health Information Research Infrastructure (PHIRI) on COVID-19 is a practical use case and lays the foundation for ultimately developing a permanent distributed infrastructure on population health (DIPoH). The intent is to support research across Europe through the identification, access, assessment and reuse of population health and non-health data to underpin (public health) policy decisions on this and upcoming pandemics. PHIRI can be set-up to accomplish this in the context of COVID-19. PHIRI offers a European mechanism for structured exchanges and research using population health data and information in current and future epidemics or crises. The aim is to share data and expertise between countries through a Health Information portal on population health in close interaction with key stakeholders in the health information landscape, in particular with ECDC, EUROSTAT, JRC, OECD, and WHO.

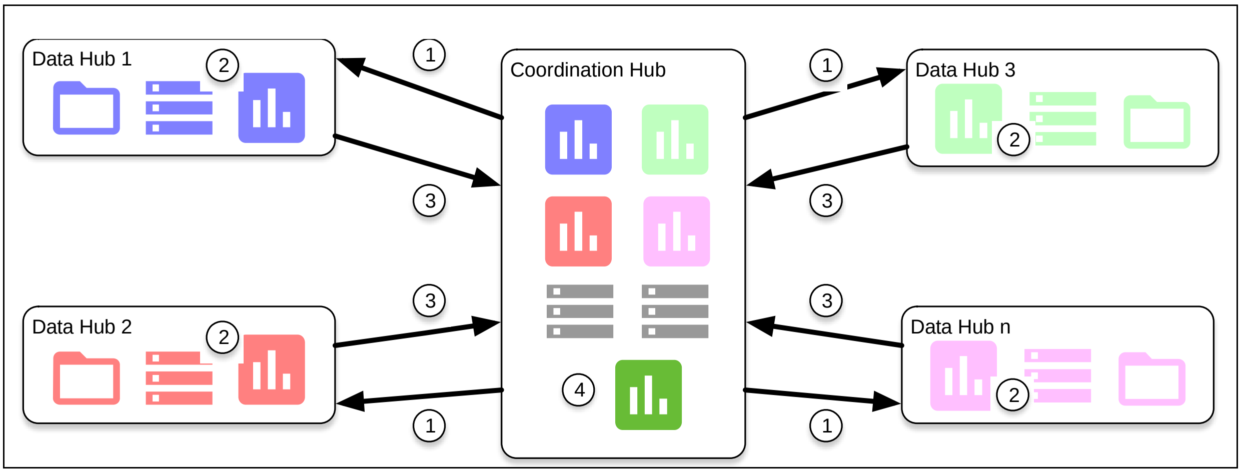
The sustainability plan for further work will be prepared to create a permanent building block of DIPoH in close collaboration with WP3. This task will coordinate and find synergies with the Members of the National Coordination Teams of the COVID-19 Data Platform.

5. DATA SECURITY

Data security and protection of privacy are crucial in PHIRI.

In PHIRI, only use cases (WP6) and the demonstration pilot (WP7) deal with -secondary use- of personal health data. Data management and data analyses will be conducted on data hubs premises' using pseudonymized or anonymised data.

PHIRI computational infrastructure is envisaged as a federated infrastructure (see figure below) where the motion of raw data between the data hubs and a central repository does not longer exist, but a constellation of data hubs that agree on mapping their data in a common data model (CDM), attached to a coordination hub. Only the analytical techniques (scripts) move from a coordination hub to partner data hubs (step 1 in the figure) where they are executed at local level (step 2). In the hubs, partial results are computed and then gathered in a coordination hub (step 3) that combines them to get an overall solution to the research questions (step 4).



A major advantage of this approach lies in the fact that all the analyses with individual-level data are performed in the data hub premises following their own governance rules and regulatory restrictions and avoiding the potential security risks of having sensible data in a single point and the legal restrictions of moving massive data outside regions or countries.

Software packages´ security

The analytical pipelines developed using R and Python will be systematically tested using synthetic data on IACS premises to minimise the impact of the security issues. The Docker images where analytical pipelines will be containerized will be based on secure Linux distributions (e.g. Alpine Linux, Tails, etc.). A solution of public key cypher for package access will be implemented so as to guarantee the execution of properly authenticated Docker images in Partner's premises, to avoid a malicious modification of the images.

6. ETHICAL ASPECTS

Ethical aspects are discussed in section 5 of the DoA. All Consortium Partners are committed to comply with relevant international and EU level fundamental ethical, privacy and security legislation and regulations:

- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

- The Declaration of Helsinki, “Ethical Principles for Medical Research Involving Human Subjects”, that sets out the ethical standards.

- The Oviedo Convention on Human Rights and Biomedicine.

- The Belmont Report “Ethical Principles and Guidelines for the Protection of Human Subjects of Research”.

- The Charter of Fundamental Rights of the European Union (7 December 2000) that sets out civil, political, economic and social rights of European citizens and all persons resident in the EU.

- The Recommendation of the Committee of Ministers No. R (90) 3 concerning medical research on human beings.

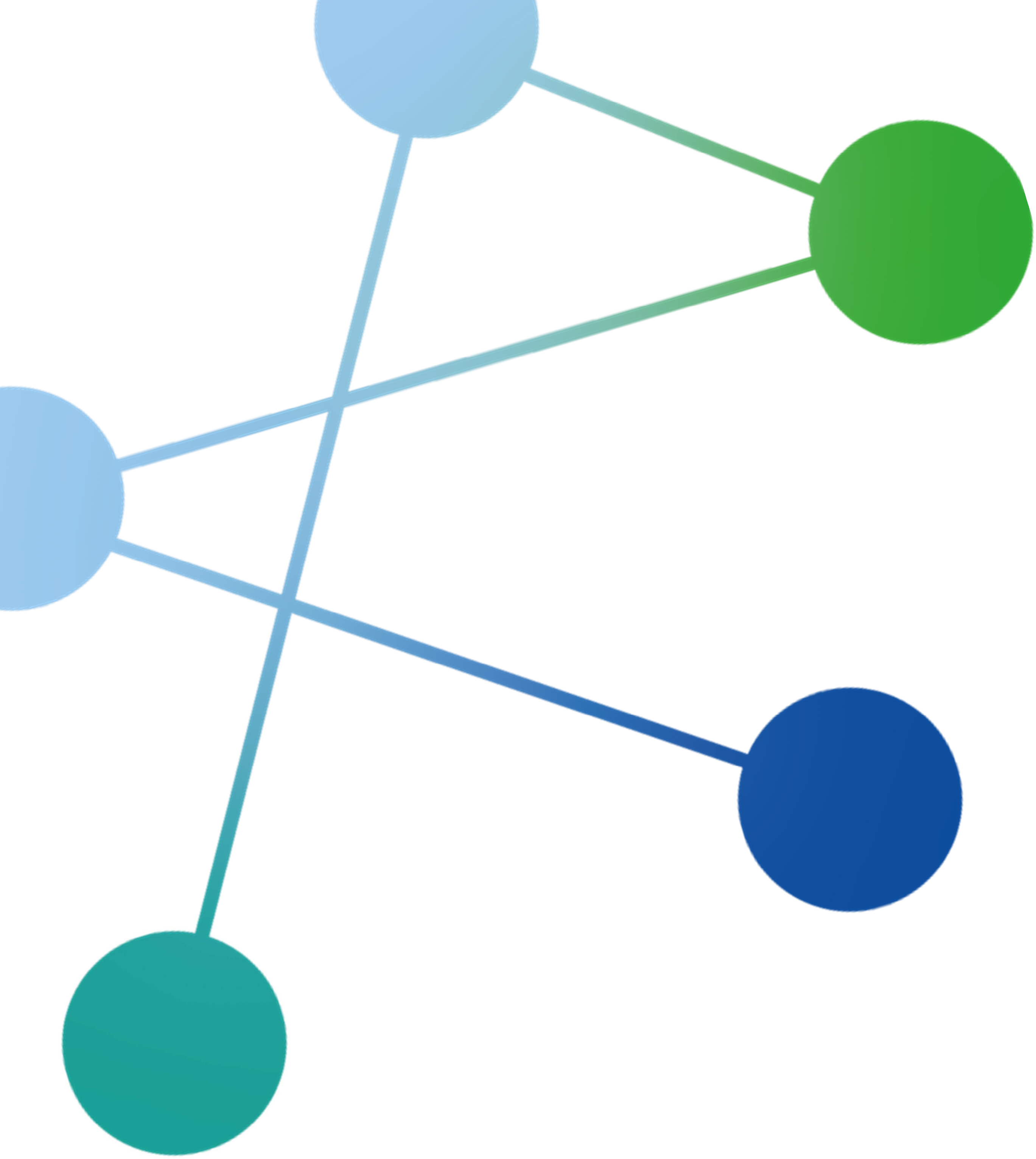
For WP6 use cases using pseudonymised data, GDPR applies and appropriate compliance will be ensured. The Data Protection Officers (DPO) of the Consortium Partners will be involved, where relevant, in use cases in WP6 and the demonstration pilot in WP7 as well as in the surveys (WP5, WP8, WP9.)

The participants in the surveys gathering input for the use cases and the demonstration pilot will receive the necessary information (in language and terms intelligible to the participants) prior to their consent to its participation. The personal data categories captured will be minimized and will be specified in in the Data Management Plan.

The participants in events (conferences, workshops, stakeholder forums, etc.) will also receive the necessary information about the handling of their personal data prior to their consent.

Finally, data and results produced as a consequence of surveys or uses cases will be stored in a secured server of Sciensano with a copy in IACS. In any case, stored results will not contain personal data.

-



Instituto Aragonés de Ciencias de la Salud (IACS)
Centro de Investigación Biomédica de Aragón (CIBA)
Avda. San Juan Bosco, 13, planta 0
50009 Zaragoza
www.phiri.eu
@PHIRI4EU