HEALTHYCLOUD
Health Research & Innovation Cloud

D5.5 Reference guidelines for the establishment of an ethically sound and legally compliant health data research ecosystem

Version 1.0

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M26, April 2023 (extended to M28, June, 2023)
Dissemination Level	Public
Nature	Report
Author(s)	Salvador Capella-Gutierrez (BSC) Harald Wagener (de.NBI-Cloud) Sina-Victoria Barysch (de.NBI-Cloud)
Contributor(s)	Alba Jene (BSC) Gergely Sipos (EGI)
Reviewer(s)	Juan Gonzalez Garcia (IACS) Josep Lluis Gelpi (UB)
Keywords	Health-related data, Capabilities-based Maturity Model, secure processing,

Notice: *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement Nº965345*

Change log

Version	Author	Date	Description of Change
0.1	Salvador Capella-Gutierrez	01/02/2023	Table of contents
0.3	Salvador Capella-Gutierrez Alba Jené	15/05/2023	Initial draft
0.9	Salvador Capella-Gutierrez Harald Wagener Sina-Victoria Barysch	24/11/2023	Cuasi-final draft
1.0	Salvador Capella-Gutierrez Harald Wagener	30/11/2023	Final version of the document answering reviewers comments.

Table of contents

Executive summary

The HealthyCloud project, as a manifestation of the Health Research and Innovation Cloud (HRIC), plays an essential role in pushing forward a strategic agenda for using and reusing health-related data across Europe for research, innovation, and policy-making. Positioned as an interconnected system, HRIC will align with pivotal initiatives such as the European Open Science Cloud (EOSC) and Gaia-X. This interconnectedness demands compliance with existing European regulations, e.g. the General Data Protection Regulation (GDPR), and forthcoming ones, the European Health Data Space (EHDS), while ensuring alignment with national laws dictating health-related data use and processing.

Deliverable D5.5 addresses the reference guidelines for establishing an ethically sound and legally compliant network of computational nodes within HRIC. Those guidelines build in previous deliverables from the same work package and underscore the importance of aligning with maturity levels for processing health-related data of varying granularities. Striking a balance between intricacy and compliance, computational nodes are envisioned to form a distributed infrastructure, adapting to the availability of data and the potential for orchestrating distributed analyses. As the HRIC evolves beyond a technical blueprint, there's a recognised need to mature the ethical soundness of health-related data use. This ensures that HRIC transcends being a technical infrastructure, becoming a comprehensive framework for federated research infrastructures.

The imminent EHDS regulation provides a common framework for research infrastructures and data holders, setting the stage for effective partnerships within an EHDS-based health-related data ecosystem. Beyond EHDS, upcoming regulations such as the EU Data Governance Act and AI Act exert significant influence, establishing boundaries to mitigate risks and safeguarding against potential pitfalls associated with data governance and adopting machine learning. Embracing these regulations is crucial for instilling ethical and legal compliance in the dynamic landscape of health-related data infrastructures.

Beyond the collaboration with EOSC and Gaia-X, HRIC also intersects with the European '1+ Million Genomes' initiative, positioning itself as the technological substrate for enabling the (federated) analysis of genomic data across Europe. Additionally, Gaia-X's architecture blueprints offer structured insights for the health data ecosystem's development, prompting exploration and adoption of these ideas to address unexplored aspects in other approaches, fostering interoperability among data holders and users.

1 Introduction

The HealthyCloud project, as the incarnation of the Health Research and Innovation Cloud (HRIC)[1], elaborates on different aspects of using and re-using health-related data across Europe for research purposes. As such, HealthyCloud is not operating in isolation but should align with major initiatives and efforts to facilitate the use of health-related data, especially the European Open Science Cloud (EOSC) and the public-private initiative for setting up a sovereign European cloud infrastructure, Gaia-X. Both initiatives might include different stakeholders but should align with the existing regulatory efforts at the European level, e.g., the General Data Protection Regulation (GDPR) and the still-under-discussion regulation of the European Health Data Space (EHDS) and national laws dictating the use of health-related data for research purposes.

According to the HealthyCloud Glossary[2], "Data-centric health research computational infrastructure" is a technological infrastructure that provides data as a service and facilitates its processing for research, innovation and policy-making purposes. Understanding these infrastructures and how they constitute the foundations of the Health Research and Innovation Cloud (HRIC) is the main focus of work package five (WP5). WP5 aims to explore existing solutions and propose new ones, whenever necessary, both in terms of physical (hardware) and logical (software) infrastructures, including the different aspects of managing health-related data and its use and re-use across Europe. Within this WP, task 5.5 is in charge of bringing together the outcomes of previous ones in WP5 and connecting their outputs with other parts of the project, including the ethical and legal aspects (WP2), the work on data management (WPs 3 and 4), the end-users' perspectives (WP6) and the connection with real-world use-cases (WP7).

1.1 Building in HealthyCloud's findings towards an ethically sound and legally compliant health data research ecosystem.

HealthyCloud WP5 has produced four deliverables, listed and discussed below, bringing the experience of real-world installations that can guide the establishment of an ethically sound and legally compliant health data research ecosystem regarding its underlying infrastructure. These deliverables, considered the foundations for the current one, also cover other relevant aspects, such as the software needed to orchestrate the analysis of health-related data and how to respond when a security breach occurs.

[1] Aarestrup FM, Albeyatti A, Armitage WJ et al. Towards a European health research and innovation cloud (HRIC). Genome Med 12, 18 (2020). https://doi.org/10.1186/s13073-020-0713-z

[2] Kesisoglou I, Cosgrove S, Derycke P, et al. Glossary of commonly used terms in the field of Health Data Research. Zenodo, (2022). https://doi.org/10.5281/zenodo.6787119

- D5.1: **Analysis of existing computational infrastructure models, including ELSI.**
- D5.2: **Analysis of existing orchestration mechanisms for distributed computational analyses, including a general overview to facilitate new developments**.
- D5.3: **Guidelines to establish sustainable computational infrastructures for the future HRIC ecosystem**.
- D5.4: **A study of existing site security policies for sensitive data and protocols for responding to breaches**.

Deliverable D5.1 discusses different aspects of managing sensitive data and enabling its use for research. These conclusions came after examining 13 physical and logical infrastructures, including one functioning as a health-related data lake. The overall organisation of these infrastructures vary widely, as it is influenced by i) funding models, ranging from national to private initiatives with specific scopes, and ii) the existing legal frameworks for handling and managing sensitive data, which is dedicated by European, National and, in some cases, local regulations. When considering the establishment of the European Health Research and Innovation Cloud (HRIC), this deliverable outlines the need to consider the inclusion of local and centralised **Secure Processing Environments (SPEs)** - also known as Trusted Research Environments (TREs) - providing capabilities for data hosting and analysis while compliant with existing and upcoming regulations. These SPEs would also encompass trusted repositories, domain-specific registries, and secure facilities for data processing, necessitating the development of secure data governance tools for an interoperable and cross-border health-related data ecosystem.

Deliverable D5.1 strongly relies on the "**Five-Safes**" model[3], which ensures the secure handling and analysis of sensitive data through five dimensions: safe projects, people, data, settings, and outputs. This model acts as an architectural approach and as a set of specific requirements for systems dealing with sensitive data. The Five-Safes model addresses crucial questions about data access and emphasises the need for appropriate use, trustworthy researchers, data security, limited access, and non-disclosive results. The report aligns aspects of secure data processing within these dimensions and outlines specific requirements SPEs might need to meet to ensure data safety.

Deliverable D5.2, stemming from task 5.2, outlines the challenges facing **distributed computational analysis across Europe** from a software orchestration perspective. Orchestration systems, like workflow management systems (WMS), facilitate secure and reproducible data analysis by encapsulating runtime environments using software container technologies. WMS generally ensure traceability, reproducibility, and provenance tracking of analysis steps seamlessly handle different storage classes and dependencies. Additionally, emerging approaches like JupyterHub, Binder, and Google Colab offer browser-based

[3] Ritchie, F. Secure access to confidential microdata: four years of the Virtual Microdata Laboratory. Econ Lab Market Rev 2, 29–34 (2008). https://doi.org/10.1057/elmr.2008.73

environments using containerisation for reproducible data analysis, enabling the capture and replay of computational notebooks and fostering collaboration and portability in research communities. Thus, task 5.2 began by identifying key areas for investigation within existing workflow and orchestrator systems, focusing on data and computation distribution, support for sensitive data, and ensuring data provenance and reproducibility. In this work, information about various systems was gathered from project partners, system owners, and other sources, creating an initial catalogue of systems to be further investigated. Promising systems from this catalogue were then selected for further investigation. D5.2 encapsulates insights gained after presenting a hypothetical scenario to various workflow and orchestration systems. The responses highlighted the strategies these systems could employ in scenarios involving multiple data holders, each housing health-related data, usually of a sensitive nature. Solutions varied from orchestrating computations with data exchange to leveraging distributed or centralised approaches while considering data sensitivity and deployment across the hypothetical network.

The subsequent analysis concluded that existing workflow systems offer basic support for distributed computing, cloud environments, reproducibility, and provenance tracking but lack comprehensive support for handling systematically sensitive health-related data, i.e. in their current implementation, they are unsuitable as SPEs without further development. While these systems employ some security measures like encryption or third-party solutions, fully defined processes addressing complex health-related data handling for research, innovation, and policy-making purposes remain an open topic. This deliverable suggests leveraging mature workflow environments for HRIC applications, emphasising the need to reinforce and extend health-related data support within these frameworks. Collaboration with stakeholders in designing and implementing **a software stack that allows analysing data available within HRIC is proposed**, acknowledging the absence of a one-size-fits-all solution for handling sensitive data within workflow execution systems.

Deliverable D5.3 put forward some **recommendations when setting up a computational infrastructure compatible with HRIC**. This deliverable is essential to understand the assumptions made in the context of deliverable D5.5. Among those recommendations, the first elaborates on authentication and authorisation infrastructure (AAI) systems. Indeed, European research computational infrastructures rely on robust AAI systems like the Life Science Login, allowing diverse institutions to access resources while safeguarding privacy. Beyond authentication, verification via principal investigators (PIs) bearing responsibility for research teams and federated identities, exemplified by Global Alliance for Genomics and Health (GA4GH) passports, LS Login ensures secure access to sensitive data, including health-related data. Standardised tagging of datasets using GA4GH Data Usage Ontology (DUO) terms streamlines access requests, although challenges persist with legacy data, where access conditions haven’t been established. Developing comprehensive training, including synthetic datasets for practice, is crucial for researchers and technical staff. Adopting the Five Safes, or similar models, to ensure security and privacy by design when assembling computational

installations dealing with health-related data would ensure effective processes regarding data confidentiality, project approval, researcher training, secure environments, and non-disclosive outputs. It is important to differentiate the design from the operation of such infrastructures as requirements, including specialised staff, might differ. Importantly, continuous monitoring for improvement remains essential for adapting to evolving technical and regulatory landscapes.

Beyond managing health-related data for research purposes, the basic building blocks of **computational infrastructures encompass crucial elements like network, storage, compute, and orchestration**. Networks must ensure logical separation among users, especially in Virtual-Machine (VM)-based or software container orchestration services, while limiting access to the minimum necessary in analysis workflows. Implementation involves techniques such as V(X)LAN separating project-specific networks, avoiding shared networks for administrative and end users, as well setting read-only access for global shared file systems. Storage should prioritise security through encryption at rest, project isolation, and fine-grained access control, distinguishing between persistent and temporary storage. Compute services may employ VM provisioned for distinct project resources, enhancing user autonomy while emphasising encapsulation and avoiding direct access to bare-metal machines. Additional measures should be strongly considered, like homomorphic encryption and/or secure processing in secure enclaves.

An important aspect to consider when assembling a computational infrastructure is its sustainability. It is important to ensure that those infrastructures are fit for purpose to have users, as investments in infrastructures without users are difficult to justify. Indeed, long-term reliability for researchers tends to influence their choice of platforms. Preservation and FAIRness of health-related data necessitate sustainable data hubs. Thus, **sustainability needs to consider operational, strategic and environmental dimensions to ensure the long-term availability** of the infrastructure while adapting to a constantly evolving technical and regulatory landscape.

When considering deliverable D5.4, associated with task 5.4, it focused on **surveying multiple European computational infrastructures, specifically exploring security protocols, breach response strategies, and handling health-related data**. The survey aimed to understand how these infrastructures manage user access to sensitive health-related data and its remote computational analysis, which is essential for building a secure local computational infrastructure that can interoperate with similar ones across Europe.

Following previously identified trends, D5.4 presents an overview of current practices regarding user identification, access control, data processing, environment management, and organisational policies. Interestingly, these practices include, first, the widespread use of federated authentication mechanisms and the recognition of multi-factor authentication (MFA) as a fundamental element in robust identity and access management policies. Second, infrastructures emphasise the necessity of dedicated project environments, logically and technically isolated from others, highlighting the importance of separating environments per

project rather than per user. Third, surveyed infrastructures exhibited practices of automatically locking user environments after data permits expire, along with regular checks to ensure ongoing validity and appropriateness of access granted. Also, pseudonymisation and encrypted transfer are common approaches when handling health data, with audits conducted before data leaves the infrastructure. While security and privacy responsibility rest on data users in generic infrastructures, those dedicated to health-related data actively manage security through variable technical and organisational measures guided by institutional policies in federated settings. Moreover, **certified sites display more comprehensive documentation and emphasise staff training as a common best practice in both internal and external policies**. Indeed, certifications serve as verifiable proof of compliance, ensuring adherence to recognised standards in handling sensitive data. Regarding managing health-related data, cybersecurity measures, certifications like ISO 27001, BSI C5, and CSA STAR, and contractual agreements between data controllers and processors are essential for secure data processing and GDPR compliance. We expect additional guidance developing from the upcoming EHDS regulation.

Deliverable D5.5, associated with task 5.5, builds on all these previous efforts and proposes **a capabilities-based maturity model that guides** the aspects that any infrastructure should consider when joining an ethically sound and legally compliant health data research ecosystem. Such infrastructures might interact directly with researchers, e.g. providing access to secure environments for analysing health-related data, or indirectly, e.g. acting as data hubs following the mandate of the data controllers.

1.2 The Role of Open Science in establishing an ethically sound and legally compliant innovative ecosystem for health data research.

Data-driven health research across Europe includes data generated within research projects, e.g. clinical trials or translation research, and the ones obtained for its secondary use from healthcare settings, e.g. Electronic Health Records, Laboratory tests or clinical imaging. Thus, it represents a multi-factorial scenario where data tends to be distributed and likely siloed within each data source, especially when considering data from healthcare settings. In recent years, there has been an important effort led by the medical informatics community to adopt different standards to facilitate syntactic and semantic interoperability across different data sources. Within those efforts, the open science principles represent a great opportunity to foster innovation and collaboration and the dangers of being vendor-locked for a particular system, in terms of data acquisition and management as well as software solutions for using, re-using and processing such data. Within the Open Science principles, the most relevant ones related to the adoption and use of Open Standards to facilitate the interoperability across

sources, the Open Data principles, intimately connected with the FAIR Data principles[4] to facilitate the reusability of data from its generation by using known standards to represent the data itself and the associated metadata. Importantly, taking into account the FAIR data principles, which further develop the premises of the Open Data, health-related data should be as open as possible and as closed as necessary, reflecting the sensitive nature of such data and the common practice - following the existing EU regulations and national legislations - of keeping it under strict control access mechanisms.

Another important aspect of Open Science to ensure the long-term sustainability of an **ethically sound and legally compliant ecosystem** for health data research is adopting Open-Source licences for software development. This aspect is proving challenging, especially for legacy code, given the interest of different stakeholders in developing those solutions. However, there is a strong push from the different communities to adopt Open-Source licence models that allow the collaborative development of software solutions oriented to adopting other Open Sciences practices, especially the ones related to Open Standards and Open Data. Adoption of Open-Source licences does not necessarily limit innovation as different business models around it can be developed in terms of deployment and maintainability, faster bug-fix and earlier access to new features, and prioritised user support, among other possibilities. However, some stakeholders, e.g., companies and public funders, are reluctant to embrace software developed using open-source licences due to the potential lack of support and low software quality, among other aspects that may negatively impact software development. Adopting a three-tier model[5] for software development, distinguishing between analysis code, prototype tools, and infrastructure-oriented software might help mitigate such reluctance and provide clear indicators on what level of quality is expected depending on the type of software being developed.

[4] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[5] Australian Research Data Commons. A National Agenda for Research Software. Zenodo (2022). https://doi.org/10.5281/zenodo.6378082.

2 Methods

Establishing an ethically sound and legally compliant health data research ecosystem implies an iterative adoption of recommendations and guidelines. Such a process can be assimilated into the development of a Capabilities-Based Maturity Model[6] (CBMM), which facilitates that different stakeholders can take part in the ecosystem at different levels of development and have clear indications of what services can be provided and used. **What services can be provided and used directly affects different profiles within the ecosystem**, especially for the infrastructure provider, researchers and data management profiles. This brings together the data curator, data steward and data manager[7] profiles. Implications have different meanings for each of those profiles. For the infrastructure provider, especially for the computational infrastructure provider, implications related to the different services and their associated configurations that can be set up within its physical infrastructure to serve direct users, e.g., a trusted research environment, or indirectly, e.g., facilitating the management of health-related data. For the researchers, the implications relate to the expected service in a given computational facility and how those services can be used for research, innovation and policy-making purposes. Finally, for data management profiles, implications related to which services are available at the computational infrastructure that can be used for setting up and maintaining a health-related data hub, curating the deposited data and supporting the scientific activities carried on by researchers. Thus, **different maturity levels directly impact the ecosystem's provision and use of the expected services**.

The **Capabilities-Based Maturity Model (CBMM)** is a framework used to assess and improve an organisation's technological capabilities and maturity throughout the technology development process. It evaluates and enhances an organisation's capacity to develop and deploy technology effectively. Considering the distributed nature of the Health Research and Innovation Cloud (HRIC), the CBMM can be considered as a tool for self-evaluation and alignment with the overall strategy. Indeed, when assuming a CCBMM, those are the important aspects to consider:

1. **Focus on Capabilities**: CBMM focuses on evaluating an organisation's capabilities rather than just assessing the maturity of technologies themselves. It examines the organisation's ability to manage, develop, and deploy technologies.
2. **Multi-Dimensional Assessment**: CBMM typically encompasses multiple dimensions. These could include technical expertise, project management, organisational culture, innovation processes, infrastructure, and resource management.

[6] Paulk M, Curtis W, Chrissis M.B & Weber C. Capability Maturity Model for Software. Software Engineering Institute (1993). https://insights.sei.cmu.edu/library/capability-maturity-model-for-software-version-11

[7] Portell-Silva L, Capella-Gutiérrez S & Lopez L. FAIR Health Data Portal Expected Users' Interactions. Zenodo (2023). https://doi.org/10.5281/zenodo.7949977.

3. **Maturity Levels**: Like other maturity models, CBMM often uses a staged approach with maturity levels or stages. Each level signifies a stage of development and sophistication in the organisation's capabilities related to technology development.
4. **Continuous Improvement**: CBMM emphasises continuous improvement. It doesn't just assess the current state but aims to progressively guide organisations toward enhancing their capabilities.
5. **Tailored Assessments**: CBMM allows for tailoring assessments to suit specific organisational contexts and technological domains. Its application is flexible, adapting to different industries or organisations' unique needs and characteristics.
6. **Strategic Alignment**: It emphasises aligning technological capabilities with organisational strategies and goals. This ensures that the technology development efforts align with broader business objectives.
7. **Benefits**: CBMM provides a structured approach to evaluate an organisation's readiness and capacity for technology development. It helps identify strengths and weaknesses, guide investment decisions, prioritise improvements, and foster a culture of continuous learning and advancement.
8. **Implementation and Adoption**: Organisations implementing CBMM often undergo an assessment phase to evaluate their capabilities. Based on the assessment, they devise strategies and initiatives to enhance capabilities in various areas identified as critical for technological development.
9. **Industry Applicability**: CBMM widely applies to industries and sectors that rely on technology development. It helps organisations in areas such as research and development, innovation, product development, and project management.

In essence, CBMM provides a structured framework for organisations to evaluate, improve, and align their technological capabilities with their strategic objectives. It supports the evolution and optimisation of an organisation's capacity to develop and leverage technology effectively. **In the case of HRIC, the CBMM provides a framework to incorporate different computational infrastructures at different maturity levels into the ecosystem, providing them with guidance on reaching more mature stages.**

There are different models to assess the maturity level of technological developments. The best-known one is the **Technology Readiness Level (TRL)**, a method used to assess the maturity of a technology, particularly in the context of research and development. Originally developed by NASA in the 1970s, TRL aimed to evaluate technologies for space missions. However, it has since been widely adopted across various industries beyond aerospace. Similar efforts have been made to assess the readiness of European Research Infrastructures (RIs) in the context of ESFRI, the European Strategy Forum on Research Infrastructures. This

report[8] emphasises the need for a structured approach integrating readiness levels to ensure the long-term sustainability of these infrastructures. They propose a staged funding strategy, establishing checkpoints for progress verification and recommending an independent expert panel to evaluate each phase, advocating that RIs on the ESFRI roadmap must demonstrate progress before seeking further EU funding support. The focus lies on implementing a robust lifecycle approach that assesses and guides the development of RIs, promoting their readiness and alignment with the European Open Science Cloud (EOSC) to enhance coherence and effectiveness in funding allocation across European Research Framework programmes.

The primary goal of TRL is to provide **a standardised and systematic way to gauge the maturity of a technology**. It helps assess the risks associated with implementing new technologies, guiding decision-making, and estimating the readiness for real-world application, which is quite relevant when guiding computational research infrastructures joining a distributed data-driven health research ecosystem across Europe. The TRL scale typically ranges from 1 to 9, each representing a specific stage in a technology's development:

- **TRL 1 - Basic Principles Observed**: This marks the lowest level where scientific research begins, and basic principles are formulated.
- **TRL 2 - Technology Concept Formulated**: The application of basic research starts to be defined, and the feasibility of a concept begins to be assessed.
- **TRL 3 - Experimental Proof of Concept**: Active research and development demonstrate a concept's practicality.
- **TRL 4 - Technology Validated in Lab**: The technology is tested in a laboratory environment to validate its functionality.
- **TRL 5 - Technology Validated in Relevant Environment**: The technology is tested in a relevant environment that simulates operational conditions.
- **TRL 6 - Technology Demonstrated in Relevant Environment**: The technology prototype or model is demonstrated in a relevant operational environment.
- **TRL 7 - Technology Demonstrated in Operational Environment**: The technology is tested and proven in an operational environment, close to its final form.
- **TRL 8 - Actual System Completed and Qualified**: The technology is complete and qualified through testing and is ready for its intended mission or purpose.
- **TRL 9 - Actual System Proven in Operational Environment**: The technology has been successfully deployed and is operational in its intended environment.

[8] European Commission, Directorate-General for Research and Innovation, Supporting the transformative impact of research infrastructures on European research: report of the High-Level Expert Group to assess the progress of ESFRI and other world-class research infrastructures towards implementation and long-term sustainability. Publications Office of the European Union; 2020. https://doi.org/10.2777/3423

Organisations, especially those involved in research, development, and innovation, have adopted the TRL approach as a valuable tool for managing and for assessing technological maturity. Significantly, **it aids decision-making by providing a common language to communicate technology readiness**. This enables better resource allocation, risk management, and project prioritisation based on technological maturity, which is important in a distributed infrastructure with multi-stakeholders.

Beyond the CCBM and the TRL as frameworks, **organisations should consider another axis related to the organisation of their computational services, especially the cloud-based ones**. Indeed, there are four models (as depicted in Figure 1) representing various approaches. The first one - and probably less popular nowadays for its implications - implies having a private cloud system under the direct responsibility of the institution. The other three imply varying levels of management and offer the possibility to have them on-premises or remotely: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The level of control and responsibility varies significantly across these models:

- **Private**: Users have complete control and responsibility for the physical and logical installations. Considering the needs of specialised staff, specific services management procedures, and associated legal responsibilities, this model is found rarely available for end-users.
- **Infrastructure as a Service (IaaS)**: Users have the most control and responsibility. They manage virtualised infrastructure resources such as servers, storage, and networking. Users are responsible for maintaining the operating systems, applications, and data hosted on these resources. Real-world examples include Amazon Web Services (AWS) EC2, Microsoft Azure Virtual Machines, and Google Compute Engine.
- **Platform as a Service (PaaS)**: Users have less control over the underlying infrastructure. They focus more on developing, deploying, and managing applications. The cloud provider manages the underlying infrastructure (servers, storage, and networking), while users handle applications and data. In this case, real-world examples include Heroku, Google App Engine, and Microsoft Azure App Service.
- **Software as a Service (SaaS)**: SaaS gives users the least control and responsibility. Here, users' access and use software applications hosted by a third-party provider over the Internet. The provider manages everything from the infrastructure to the application itself, including updates, security, and maintenance. Examples: Gmail, Office 365, Salesforce, Dropbox.

In summary, the control and responsibility gradually shift from IaaS users to the SaaS cloud provider. Users have more flexibility and control over their infrastructure and applications in IaaS, while in SaaS, they have minimal control as they simply use the provided software. PaaS falls in between balancing control and abstraction of the underlying infrastructure. However, **a careful examination should be conducted when managing health-related data given its sensitive nature** and the existing EU regulations, e.g., GDPR, and national legislation. Indeed,

commercial cloud providers must be certified to ensure they comply with the different regulations.

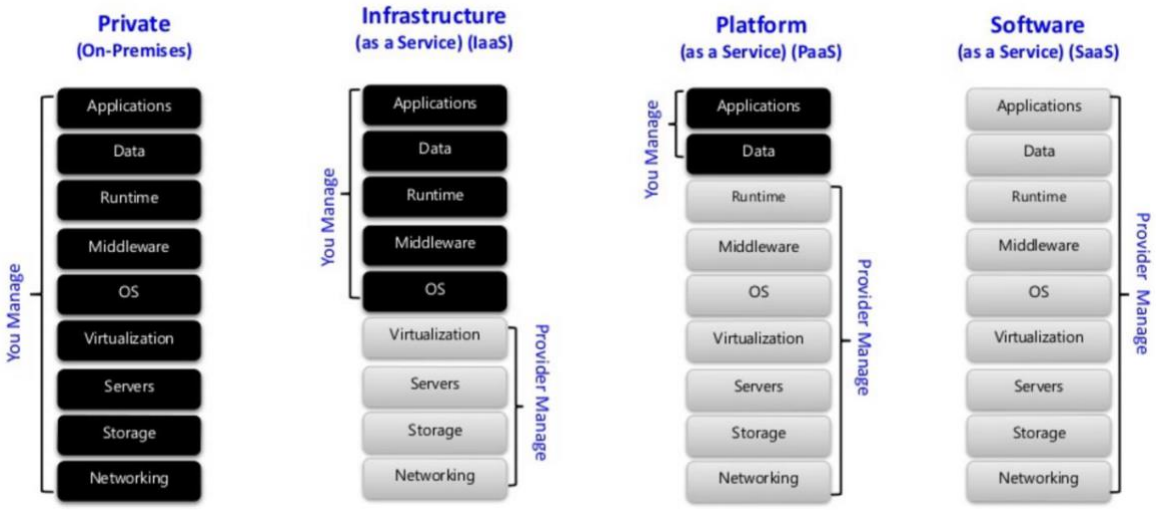

Figure 1. Stack under the user's control on the different cloud service models.

3 A framework for establishing an ethically sound and legally compliant data-driven health research ecosystem.

The Health Research and Innovation Cloud (HRIC) strongly depends **on technological solutions to facilitate the management of sensitive health-related data and its secure processing while observing the existing EU regulations and national legislation**. Several factors, including legal, organisational, human and technological resources, might limit the ability to have fully mature technological solutions to provide those services. Moreover, the heterogeneous nature of health-related data, including its granularity, associated consents and origin, also conditions how data can be managed and analysed locally and across borders within the European Union. Thus, it is necessary to have a framework that incorporates those aspects together with the level of readiness of the different computational installations. This approach would maximise the number of participating infrastructures, contribute towards its long-term sustainability and serve researchers with health-related data and processing capabilities across Europe.

3.1 A utopian vision for establishing a federated data-driven health research ecosystem.

We foresee a five-level maturity model for establishing a federated data-driven health research ecosystem reflecting on the technological readiness of each infrastructure, the possibilities to manage health-related data at different granularity levels according to existing regulations, and the capabilities available at each site. The foreseen five levels are as follows:

- **Level 1:** Researchers find and gain access to relevant data for their studies through different data hubs and repositories and use it to perform additional analysis using their own computational facilities.
- **Level 2**: Researchers can find and gain access to relevant FAIR health-related data through any relevant catalogue and use it to perform additional analysis using their own computational facilities. Minimum data and metadata FAIRification are required to facilitate reusability.
- **Level 3**: Researchers can find and gain access to relevant FAIR health-related data using metadata-based catalogues, which can be domain-specific or transversal, and use it to perform additional analysis. Portals containing information about the catalogues might also indicate where data can be analysed.
- **Level 4**: Researchers can find and gain access to relevant FAIR health-related data through a common portal, but need to bring data independently to each computational infrastructure. Access can be automatically granted if conditions are specified using machine-readable mechanisms, e.g., controlled vocabularies.

- **Level 5**: Researchers can find, request and gain access to relevant health-related FAIR data through appropriate cataloguing mechanisms and analyse them using a distributed/federated computational infrastructure - being transparent to their organisation and interconnection.

Beyond the different maturity levels of readiness for managing and processing health-related data, at least two transversal general areas need to be considered independently of the maturity level of any participating organisation. The first transversal area relates to the use of software:

- **Software stack** supporting the operation of the installations, either to provide direct service to researchers for analysing their data or indirectly by facilitating health-related data management. The software stack should have a high TRL, at least level 6 (see previous section for available definitions), be well-documented, have a strong user base, and ideally use open-source licences that allow further modifications. Being unable to modify software may both preclude reaching higher TRL levels, or maintaining them, as best practices and expectations of how software works tend to change over time. For these tasks, on-premises solutions directly managed by the host organisation, IaaS and PaaS, are the expected approaches to carry them on.
- **Research software** used to analyse available data through individual tools or complex analytical workflows should follow software best practices to contribute to the reproducibility of results. Software development best practices recommendations[9] include having version-controlled repositories for the source code, adoption of software licences - ideally, open-source ones, registry in known registries depending on the nature of the software, e.g., individual tools vs. analytical workflows, and the set-up of contribution guidelines to favour interactions and contributions by the broad community. It is common practice to have PaaS and SaaS as the preferred models to enable the use of research software. PaaS is often used with orchestration managers, e.g., Galaxy and Nextflow, to facilitate the local and/or remote execution of software. At the same time, SaaS is often associated with controlled environments, e.g., SPEs and TREs, where software and health-related data are made available to researchers for their use.
- **Software certification** is an important element contributing towards deploying long-term sustainable computational infrastructures as it guarantees that the software is stable given that the established operation conditions are met. However, even though this is a desirable element, it should not be a strict requirement to manage and process health-related data for research, innovation and policy-making purposes, due to the complexity of the certification process. An important aspect to consider for any software at any maturity level is the existence of detailed and up-to-date documentation. Appropriate documentation is at the basis of the software

[9] Jiménez RC, Kuzak M, Alhamdoosh M et al. Four simple recommendations to encourage best practices in research software. F1000Research 6:876, (2017). https://doi.org/10.12688/f1000research.11407.1

certification process, as it provides the architecture design of the software and its expected operating conditions.

The other transversal area relates to how users can access and process health-related data while ensuring the privacy and secure management of those operations. We also include elements relevant to the computational facilities regarding how to serve users best.

- **Secure access** relates to those mechanisms that enable access to health-related data and processing facilities to process such data. Those mechanisms are based on authentication mechanisms to recognise users accessing any service in the ecosystem. Ideally, those services should be federated to facilitate transparent access across distributed services using single sign-on approaches, with trusted partners managing the authentication service. Multi-factor authentication (MFA) mechanisms are envisioned to strengthen the authentication of users into the system, reducing the risk of misidentification and misuse of identities in the system. Importantly, the current practice when accessing sensitive data across computational installations sets the responsibility to the research project's principal investigator. It is important to ensure that regular checks are performed to ensure that granted access to data and facilities are still valid, as research projects usually have specific timelines for their completion. Beyond managing how users access data and computational facilities, it is also vital **to perform penetration tests at the computational installations to ensure that appropriate cybersecurity measures are implemented and up-to-date** in an always-evolving technological landscape.
- **Secure data processing and analysis** of health-related data for research, innovation and policy-making purposes once the researcher has gained access to it is another essential aspect to consider within the ecosystem. The first consideration is that **data should be at least pseudonymised, preventing the re-identification of individuals with a reasonable technical effort** and, ideally, anonymised. Pseudonymised data allows reporting any incidental finding that was found during the research project to data controllers. However, it must be guaranteed that researchers cannot break those pseudonyms. Other mechanisms like data minimization[10] should also be observed to reduce the re-identification risk by combining different datasets. Moreover, data should be transferred using encryption mechanisms, preventing the risk of exposing sensitive data unintentionally to third parties. Following aspects previously discussed in other deliverables (D5.1, D5.3), each project should have dedicated environments that prevent other users of the same facility from accessing non-authorised data. Finally, considering the 5-Safes recommendations, results must be audited before leaving the infrastructure.
- **Organisational policies and procedures** for service provision, managing and processing health-related data following existing best practices and regulations. To

[10] Article 5(1)(c) of the GDPR.

make this possible, the organisation should ensure that all relevant processes are well-documented, and action plans are established, ideally allowing the certification of these processes following well-established mechanisms, e.g., ISO 27001, BSI C5, and CSA STAR. It implies that staff receive periodic training in common best practices to ensure they are up-to-date with the latest technological developments.

3.2 Capabilities-Based Maturity Model for computational infrastructures joining the data-driven health research ecosystem.

Based on the collected background information across different deliverables and previous sections' findings, we propose **Ω**five-level capabilities**-based maturity model** (CBMM) to guide any new or existing computational infrastructure joining the data-driven health research ecosystem envisioned within HRIC. The CBMM should contribute to an incremental adoption of services to serve research communities better while complying with existing EU regulations and national legislation. CCBM levels go from level 1, representing an isolated computational infrastructure with limited services, to level 5, representing a mature infrastructure integrated into the European HRIC, which can provide services to researchers, innovators and policy-makers.

- **Level 1:** Researchers find and gain access to relevant data for their studies through different data hubs and repositories and use it to perform additional analysis using their computational facilities.
    - **FAIR data use**: Adopting the FAIR principles for the data and metadata is possible, but not mandated or controlled by the hosting facility. It might not be possible to know beforehand about any necessary data transformation to common data models, nor can these transformations persist beyond the research project.
    - **Data Granularity**: It depends on third-party providers, with varying data availability from public repositories to access control ones. The responsibility of dealing with the data correctly is assigned to the research project's principal investigator.
    - **Cross-border data use**: The infrastructure is not expected to facilitate health-related data use across EU member states at this level of maturity.
    - **AAI/Single sign-on**: The infrastructure maintains Authentication and Authorization mechanisms without the possibility of interconnecting it beyond the own infrastructure. Different sign-on mechanisms may be used for different services at the hosting infrastructure.

- **Containerized/Multi-Cloud Applications**: Users are bound to the applications provided by the computational facility. Research software might be available through software containers, but can also be installed directly into the infrastructure. The software could not be available through known public repositories for its deployment anywhere.
- **Infrastructure/platform certification**: Infrastructures self-report according to self-selected standards and criteria.
- **Software/workflow Certification**: Research software and workflow might be validated against requirements for control of data inputs and outputs, but this is not a requirement.

- **Level 2**: Researchers can find and gain access to relevant FAIR health-related data through any relevant catalogue and use it to perform additional analysis using their own computational facilities. Minimum data and metadata FAIRification are required to facilitate reusability.
  - **FAIR data use**: A minimum adoption of the FAIR principles for the data and metadata is required to favour data reusability and ease data discoverability using its associated metadata. Regarding the further analysis of health-related data, it is possible to bring in FAIRified data provided by external services.
  - **Data Granularity**: It depends on third-party providers, with varying data availability from public repositories to access control ones. The responsibility to deal with the data correctly is assigned to the research project's principal investigator. Aggregated and de-identified data is expected to be used at this level of maturity.
  - **Cross-border data use**: It is possible to use health-related data across EU member states, but the infrastructure does not expect nor support it.
  - **AAI/Single sign-on**: The hosting infrastructure provides its own single sign-on service to be used across all provided services by that infrastructure.
  - **Containerized/Multi-Cloud Applications**: Users are bound to the applications provided by the computational facility. Research software might be available through software containers, but can also be installed directly into the infrastructure. At this level, it is possible to run software containers-based analysis workflows.
  - **Infrastructure/platform certification**: Infrastructures is proven to work with FAIR data in hackathons/integration workshops. Processes are well-documented, and action plans are designed.
  - **Software/workflow Certification**: Research software and workflows are proven to work with FAIR data in hackathons/integration workshops.

- **Level 3**: Researchers can find and gain access to relevant FAIR health-related data using metadata-based catalogues, which can be domain-specific or transversal, and use it to perform additional analysis. Portals containing information about the catalogues might also indicate where data can be analysed.
    - **FAIR data use**: FAIR data and metadata is mandatory at this level. Each site sets the minimum requirements for the FAIRification of data. Metadata facilitates the cataloguing and, therefore, the discoverability of FAIR data.
    - **Data Granularity**: Data holders can label and qualify the granularity of the data they use. Incoming data should be at the level of aggregated data.
    - **Cross-border data use**: Data-oriented services strive to implement cross-border compliant data formats and structures, making it possible to access data transnationally.
    - **AAI/Single sign-on**: A federated single sign-on mechanism is available to some of the provided services by the hosting infrastructure.
    - **Containerized/Multi-Cloud Applications**: Most of the software used is available using container technologies for their deployment anywhere.
    - **Infrastructure/platform certification**: Infrastructures have standardised and structured documentation about their processes and procedures for infrastructure operations, including security. A continuous improvement plan is in place and periodically reviewed.
    - **Software/workflow Certification**: Infrastructures have standardised and structured documentation about their processes and procedures for software deployment. A continuous improvement plan is in place and periodically reviewed.
- **Level 4**: Researchers can find and gain access to relevant FAIR health-related data through a common portal but need to bring data independently to each computational infrastructure. Access can be automatically granted if conditions are specified using machine-readable mechanisms, e.g., controlled vocabularies.
    - **FAIR data use**: FAIR data and metadata is mandatory at this level. Each site sets the minimum requirements for the FAIRification of data. Metadata facilitates the cataloguing and, therefore, the discoverability of available FAIR data.
    - **Data Granularity**: Health-related data services differentiate between the various granularities. Access to data is controlled via controls such as Data Access Agreements. Data can be at the individual level with appropriate anonymity or, if pseudo-anonymized, there is no possibility to reverse it.

- **Cross-border data use**: All interfaces to data use are maintained and developed with the goal of cross-border use by design. Cross-European access might not always be allowed.
- **AAI/Single sign-on**: Single Sign-on is supported across all hosting infrastructure services, and integration with third-party identity providers is possible.
- **Containerized/Multi-Cloud Applications**: Hosting Infrastructure has services in place that allow the building and deployment of containerized services for data analytics. Health-related data can be packaged and delivered to federated infrastructures for consumption.
- **Infrastructure/platform certification**: Certification is planned, and appropriate processes and documentation are defined and executed.
- **Software/workflow Certification**: Software is at least examined (and signed by trusted partners) before its deployment. Certification is planned, and appropriate processes and documentation are defined and executed.

- **Level 5**: Researchers can find, request and gain access to relevant health-related FAIR data through appropriate cataloguing mechanisms and analyse them using a distributed/federated computational infrastructure - being transparent to their organisation and interconnection.
  - **FAIR data use**: All data and metadata generated/used implement the FAIR data principles via fully automated means. Considering the potentially sensitive nature of health-related data, metadata is essential to facilitate the findability of available data and understand its access conditions. Using ontological terms to describe such access conditions, e.g. GA4GH Data Usage Ontology (DUO), might contribute to automatically gaining access to those datasets.
  - **Data Granularity**: Health-related data services differentiate between the various granularities. Access to data is controlled via controls such as Data Access Agreements. Data can be at the individual level with appropriate anonymity or, if pseudo-anonymized, there is no possibility to reverse it.
  - **Cross-border data use**: Health-related data is presented in interoperable formats, and infrastructures provide interfaces so that they satisfy the requirements outlined in the different existing regulations at European and national levels, which might include safeguards to control data at the individual level, protecting it with appropriate anonymity and, if pseudo-anonymized, there is no simple way to reverse it.
  - **AAI/Single sign-on**: Single Sign-On supports integration with a federated AAI solution, such as LS AAI, facilitating access to different computational infrastructures and the different data Hubs and Collections.

- **Containerized/Multi-Cloud Applications**: Applications are built in a way that they are infrastructure neutral, and infrastructures implement standards that allow for continuous deployment of sufficiently mature services.
- **Infrastructure/platform certification**: Appropriate Certification proves that processes are in place and under continuous improvement regimens so that services are run in a privacy and safety-preserving manner whenever necessary. This level includes the necessary mechanisms for logging in the different user actions within the system.
- **Software/workflow Certification**: Research software and workflow can be validated against requirements for control of data inputs and outputs continuously. This level includes capturing and describing the execution provenance to facilitate analysis reproducibility and replicability without disclosing relevant aspects of the underlying computational infrastructure, which might compromise its security.

3.3 Ethical Soundness for data-driven health research.

Ethical integrity is a multifaceted aspect influencing data utilisation in health-related research, innovation, and policy-making endeavours. It encompasses the ethical foundation of scientific inquiries made by researchers, innovators, and policy-makers and the ethical considerations applicable to those attempting to execute, replicate or reproduce such research.

Considering the complexity of such a topic, we propose the following rough mapping of Ethical soundness to the maturity levels above. It is worth considering that this mapping is incomplete, as we are not specifying levels 2 and 4.

- **Level 1:** Each research query addresses ethical concerns independently, with a manual assessment conducted by a panel of local ethics experts.
- **Level 3:** The HRIC establishes EU-level applicable ethics guidelines, ensuring consistent implementation across all participating health-related data research infrastructures.
- **Level 5:** Automated categorization of research inquiries based on ethical considerations occurs at the EU level and is applicable to data from diverse sources within the HRIC. Ethics approvals encompass reproducibility in accordance with FAIR principles.

3.4 Legal compliance of a data-driven health research ecosystem.

While legal requirements under existing EU regulation and national legislation typically lead to binary compliance decisions, distinguishing between compliance and non-compliance, the nuanced nature of health-related data granularity introduces complexities. We contend that different levels of data granularity align with maturity levels, particularly as aggregated data and datasets reflecting general trends may not necessarily fall within the special categories

established by the GDPR and elaborated upon by the currently under-discussion EHDS regulation. However, it is crucial to acknowledge the dynamic legal landscape, where additional regulations such as the EU Data Governance Act and the EU Artificial Intelligence Act may influence the legal compliance landscape of the proposed distributed infrastructure. In exploring these intersections, it becomes imperative to address potential challenges and intricacies associated with mapping data granularity to maturity levels while ensuring alignment with evolving regulatory frameworks.

Following the previous section, we make an attempt to make a rough mapping of the legally compliant elements to the proposed technical maturity level from the perspective of the granularity of health-related data being used for research, innovation and policy-making.

- **Level 1:** At this stage, utilising exclusively aggregated data aligns with legal compliance measures, allowing the mobilisation of publicly available data across diverse EU jurisdictions.
- **Level 3**: The ability to mobilise aggregated data across secure health-related data infrastructures reflects a more advanced technical maturity. Including locally record-level pseudonymised and, ideally, anonymised data showcases higher technical sophistication and legal compliance. Emphasising adherence to regulatory guidelines regarding pseudonymisation and anonymisation techniques within this level could further strengthen the alignment with legal requirements.
- **Level 5:** It encompasses handling health-related data at an individual level while ensuring appropriate pseudonymity/anonymity. Highlighting the use of encryption mechanisms and stringent safeguards in data exchange across distributed secure computational infrastructures would reinforce the adherence to legal frameworks governing privacy and data protection.

3.5 Combining Ethical Soundness, Legal Compliance, and Technical Maturity.

The authors understand that technical solutions cannot guarantee legal compliance or ethical soundness. That said, we do believe that the maturity of tooling also implies how well they support achieving legal compliance or prove the ethical soundness of a given research question. Therefore, we propose a holistic view of the maturity levels on all three axes, further divided by the data granularities outlined above. Additionally, detailing specific examples or case studies illustrating how these axes intersect and evolve across different levels of data granularity would further enrich the understanding of this combined maturity view and serve to guide the different stakeholders providing and consuming health-related data.

4 Conclusions

HealthyCloud, as the incarnation of the Health Research and Innovation Cloud (HRIC), elaborates on different aspects of using and re-using health-related data across Europe for research purposes. As such, HealthyCloud is not operating in isolation but should align with major initiatives and efforts to facilitate the use of health-related data, especially the European Open Science Cloud (EOSC) and the public-private initiative for setting up a sovereign European cloud infrastructure, Gaia-X. Both initiatives might include different stakeholders but should align with the existing regulatory efforts at the European level, e.g., the General Data Protection Regulation (GDPR) and the still-under-discussion regulation of the European Health Data Space, and national laws dictating on the use of health-related data for research, innovation and policy-making purposes.

Deliverable D5.5 focuses on establishing ethically sound and legally compliant computational nodes within HRIC, accommodating various maturity levels for processing health-related datasets of different granularities. The higher the maturity level of a node, the more intricate data it can handle, possibly even at an individual record level. However, adherence to regulations mandates processing only pseudonymised and anonymised data while adhering to data minimisation principles.

The computational nodes supporting the foreseen ecosystem for the use and reuse of health-related data for research, innovation and policy-making purposes are expected to form a distributed infrastructure, which will work either in isolation or federally depending on the availability of the data to be processed and the possibility to orchestrate distributed analysis. Thus, the different maturity levels of distributed infrastructures will imply that certain data granularities require minimum maturity before a site or distributed infrastructure can be deemed eligible to process said data. Building upon previous deliverables, e.g. D5.4, security aspects and mechanisms to react to data breaches differ across those maturity levels.

We think there is a need to discuss and expand on how to mature the Ethical Soundness of health-related data use within the HRIC to ensure that the HRIC goes beyond yet another technical blueprint for federated research infrastructures.

4.1 European Health Data Space and other upcoming regulations.

The upcoming EHDS regulation strives to provide a common framework in which research infrastructures and data holders must operate. The publicly available draft of EHDS provides a good baseline for developing approaches to provide secondary use services. Health-related data infrastructures must identify their role within EHDS and develop these roles to be effective partners in an EHDS-based health data ecosystem.

In addition to the EHDS, upcoming regulations substantially influence the lawful operations of health data research infrastructures. The Data Governance Act serves as a critical mechanism, setting boundaries to mitigate potential risks associated with data custodians

abusing their position of privilege. Simultaneously, the AI Act is geared towards safeguarding against potential pitfalls arising from the widespread adoption of Machine Learning. Specifically, it aims to maintain control over Data Provenance and reproducibility, aligning with the overarching FAIR data principles. Such regulations are essential in ensuring ethical and legally compliant practices within the dynamic landscape of health data research infrastructures.

4.2 Integration with other key stakeholders: EOSC, Gaia-X and 1+MG.

The Health Research and Innovation Cloud (HRIC) stands at the intersection of various stakeholders within health-related data initiatives. Interactions and integration with key entities like the European Open Science Cloud (EOSC), Gaia-X, and the European '1+ Million Genomes' initiative for genomics data foster a collaborative, interconnected health-related data ecosystem. HRIC's synergy with EOSC, a vast initiative promoting open science and sharing research data across disciplines, is fundamental in aligning health-related data within a broader scientific context. This partnership enables the seamless integration of health data into the wider scientific domain, encouraging interdisciplinary research collaborations and facilitating access to diverse datasets. HRIC complements the efforts put forward by EU member states and the European Commission on using genomics to advance national precision medicine programmes, including using genomic data for research, innovation, and policy-making purposes. Therefore, it is a natural connection between HRIC and 1+MG and its associated implementation project, the European Genomic Data Infrastructure (GDI), as HRIC might become the underlying technological infrastructure to facilitate the access, mobilisation and analysis of that data.

Gaia-X has developed architecture blueprints and operational models for data spaces since 2020, and the documentation available[11] allows for the structured design of services facilitating interoperability of data holders and data users. We recommend investigating how to adopt and develop the Gaia-X ideas to support the development of the health data ecosystem since it addresses some aspects that have not been covered in depth in other approaches.

11 Gaia-X Architecture Document - 23.10 Release; Gaia-X Policy Rules Conformity Document; Gaia-X - Data Exchange - 23.11 Release