Milestone 2.2

Mechanism for data access and use of routine real-world data

Authors

Author	Partner
Francisco Estupiñán-Romero	Institute for Health Sciences in Aragon, Spain
Ramón Launa-Garcés	Institute for Health Sciences in Aragon, Spain
Iris Van Dam	Sciensano
Shona Cosgrove	Sciensano
Nina Van Goethem	Sciensano
Enrique Bernal-Delgado	Institute for Health Sciences in Aragon, Spain

Keywords

Keywords	Real-World Data, Routine data, Personal Data, Sensitive Data, Research, GDPR, Legislative proposal European Health Data Space

Document history

Date	Version	Editor	Change	Status
2023/02/06	1.0	Enrique Bernal-Delgado	Submission	Final Draft

Scope

As referred in BY-COVID Work Package (WP) description, WP2 brings together data resources and catalogues across domains, captures data governance and access procedures. It aims at aligning metadata descriptions and other relevant semantic information first within domains (e.g., biomolecular and imaging, clinical and health, survey, etc.) and at a second stage (in alignment with WP3 developments) expose a reference catalogue with harmonised metadata descriptions across domains.

Deliverable (D) 2.2, part of BY-COVID WP2, will develop a common sense of the preferred mechanisms already available for accessing real-world data and transfer it across research domains and jurisdictions, describing technical, legal and organisational barriers across data sources and identified solutions for real-world data access including the procedures to manage users permissions for controlled access data.

This milestone, which focuses on patient clinical and health data, starts paving the way of this Deliverable (D) 2.2 providing insight on issues affecting access and use of sensitive real-world data. In this milestone we do refer to routine health data reuse not to the reuse of data primarily collected for research purposes.

The milestone starts with a definition of key concepts, shedding light on the legal bases for access and reuse when the purpose is research, and provides a template for the description and assessment of RWD access and use policies and procedures using, as examples, three real-world data initiatives collected in milestone 2.1.

Concepts

What is Health Data?

Data are understood as health-related when they provide information on health status or health determinants of given individuals or populations, including data from health care provision.

What is Real-World Data (RWD)?

Real-world data are the data relating to patient health status and/or the delivery of health care that is routinely collected by the interaction of a given individual or population with the health system through a variety of perspectives and sources, mostly health and social care information systems, product and disease specific registries, patient-generated data and data gathered from other sources that can inform on health status such as socioeconomic or behavioural data (see FDA definition[1])

RWD in healthcare is observational data by nature. In healthcare, RWD is collected in electronic health records (EHRs), claims and billing activities, product and disease registries, administrative information systems, etc.

Sources of RWD are designed generally with an operational purpose to plan, monitor, manage, or deliver health care, or to ensure the continuity of care across multiple providers. Therefore, those health data sources are mainly designed and implemented with a primary purpose of delivering or supporting the delivery of healthcare. They are rarely interoperable as they are driven by the information requirements of the institutions governing health insurance and healthcare provision and administration, e.g. National or Regional Health Systems.

As a consequence, RWD reuse in health will be affected by a) the data generation mechanism (i.e., routinely collected from the interaction between the population and the health system - mainly healthcare, but also social care, insurers and public administration), b) the data collection purpose (i.e., with a primary purpose to govern, manage and provide health care); and, c) by the digital nature of those data (i.e., usually referred to as electronic health data, in general structured).

Secondary use of RWD

The use of health data from RWD data sources for a purpose that differs from the original intention for which data was collected is defined as secondary use. This definition of secondary use had generated some confusion as health data can also be actively collected with a primary purpose of research and then re-used at a later stage for a different purpose other than research such as regulatory purposes (i.e. health technology assessment), public health surveillance or health monitoring or healthcare planning which are also considered secondary uses. In this milestone we do refer to routine health data reuse not to health research data reuse.

RWD typology

RWD: typical data sources

Almost every interaction with the health system leaves a digital trace; for example, administrative information, data diagnoses, treatments, lab tests, timestamps or health outcomes.

[1] https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

Generally, health data sources where routine RWD can be found are:

- Electronic Health Records (EHRs),
- Healthcare claims and billing activities,
- Disease-specific or patient-based registries
- Health products registries,
- Insurance, census and statistical records
- Population-based health registries (i.e. public health monitoring and surveillance);
- Population-based Health Examination Surveys (i.e., including bio-parameters)
- Population-based Socio-Economic surveys
- Patient-generated data[2] as patient/population experience or opinion surveys (e.g., health barometers) or health data gathered from the internet, wearable and mobile devices used in the process of care

Those typical health data sources may make data available for research purposes under specific circumstances and regularly through restricted data access request processes depending on data availability and the type of data, and conditioned upon a specific research project.

RWD: data sensitivity

The General Data Protection Regulation[3] (GDPR) includes the requirement for the data holders to classify their data depending on their level of sensitivity, as high, medium and low level. Each sensitivity level is rated according to the potential impact that data may have for an individual if confidentiality and privacy were breached.

This translates into different levels of restriction that have to be reflected as part of the data holders' data security and privacy policy. Thus,

- High sensitivity: reserved for data that may produce a major impact in the life of an individual, such as personal data (i.e., some health data, such as certain medical diagnoses, genetic data, etc.). In this case, a data breach would likely cause harm to both the individual and the organisation hosting the data, so it should be processed and maintained within strict cybersecurity controls. This data should also have strict authorization controls, auditing procedures to detect access requests, as well as encryption mechanisms applied to data storage and transfer.

[2] *Although, most Health Systems do not routinely collect patient-generated data as part of their health information systems, apart from the data generated by medical devices including medical imaging during healthcare - usually recorded structured or semi-structured within the medical records or as a piece of the health information systems -, it is still generally considered RWD.*

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

- Medium sensitivity: characterising data that would not likely harm individuals, but it still is considered sensitive information that may describe operational details (i.e., medical appointments, surgical history). These files could be deemed medium sensitive.
- Low sensitivity: Data intended for public consumption or open publication (i.e. health statistics, information on healthcare resources, etc.) could be considered low sensitivity and would not need any strict control.

Data sensitivity classification is thought to guide the requirement for data controllers (i.e. data holders or stewards) to categorise data (i.e. at variable level) according to its level of sensitivity as restricted, confidential, and public data.

Following this classification, the GDPR classifies Personal Identifiable Information (PII) of European Union residents as highly sensitive, and therefore restricted or confidential.

Specifically, GDPR defines personal data as any information that can identify a natural person, directly or indirectly, such as: names, identification numbers, location data, online identifiers, and one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of a person (GDPR article 4.1)

To comply with the GDPR, organisations must classify data within a data inventory structure including information on a) the type of data, the basis for data protection, the categories of individuals involved in the data (i.e. general population, patients, customers, etc.), and the categories of the data users, including researchers in case the data is made available somehow for re-use with research purposes.

In addition to these considerations, at a higher level of abstraction (i.e. dataset or data source level) data sensitivity can be assessed in terms of the likelihood for a particular individual to be identified even when no personal data (PII) is stored in the data set. This is dependent on: a) data granularity (e.g. individual level data vs. provider or area level data), b) data volume and variety in terms of number of data points and attributes collected for each registry; and, c) linkage capabilities between different data sources that could increase the overall information available for re-identification. For instance, high volume, high variety data at individual level that can be linked across multiple data sets offers a higher opportunity to identify a particular individual even when no personal data (PII) is stored within the dataset.

This imposes **further requirements when data are reused for research**; for example, a) pseudonymisation and minimisation, or anonymisation depending on the target of research; and, b) processing and analysing data within secure processing environments (SPEs, also known as Trusted Research Environments - TREs)

Issues accessing sensitive real-world data

A picture of the current landscape shows that real-world health data of interest for research purposes is fragmented within functional or technological silos, in the institutions or organisations where data is collected for primary purposes.

From 25 May 2018 the General Data Protection Regulation (GDPR) is directly applicable in all EU member States. Although this Regulation provides a privileged regime for the access and use of sensitive data for research purposes, the uneven interpretation of the GDPR legal provisions across data holders makes it difficult to have a common understanding on access and use issues, particularly when research requires cross-border data exchange ().

Interestingly, there is an increasing setup of public institutions that are created with a view to be a single entry point for sensitive data, mitigating the heterogeneity of procedures, enhancing legal security and reducing the burden for users to get access to sensitive data. These institutions collect RWD from multiple primary sources, link the data sources, harmonise the data, and provide secure access for research purposes. Some examples can be FINDATA in Finland (https://findata.fi/en/), Health Data Hub in France (https://www.health-data-hub.fr/) Research Data Centre at the BfArM, Germany (https://www.bfarm.de/EN/BfArM/Tasks/Research/_node.html), Health Research Data UK (https://www.hdruk.ac.uk/) or BIGAN in Aragon,Spain (https://bigan.iacs.es/en/home).

Currently, the legislative proposal on the European Health Data Space[4] (EHDS) Regulation, in the provisions of secondary use, provides additional regulatory basis to enhance cross-border access and use of sensitive health RWD for scientific research related to health or care sectors to improve the quality, increase the quantity and the speed of the research studies in the European Research Area.

Personal data access and reuse

The forthcoming EHDS Regulation builds upon the GDPR by foreseeing that data holders have the duty to make available certain categories of personal health data for secondary use [1], when the requested purpose fits within a list of public interest purposes, such as research related to health or care sectors [2].

On many occasions health researchers will need to access and use personal data from patients as the purpose of the data processing cannot be achieved otherwise. When the data to be processed cannot be anonymised, the application of de-identification techniques reduces the risk for privacy while keeping this data useful for research. One of the safeguards most relevant to health sector research is pseudonymisation. In that case, minimisation efforts [3] are to be described in the research protocol while health data

[4] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

access bodies shall provide access to electronic health data in pseudonymised format [4], following GDPR rules as well.[5]

In order to further mitigate the risk for privacy when using pseudonymised data, the use of a Secured Processing Environment (SPE) is recommendable. This has become a requirement in the forthcoming EHDS Regulation [6] in a way that only non-personal electronic health data could be transferred out or extracted from such a secure processing environment [7]. Instead of accessing pseudonymised databases, it may be prescribed accessing synthetic data (ie, simulated data) where research can be performed with such type of data (i.e., algorithm training for artificial intelligence).

Other strategies to reduce privacy risks in the reuse of sensitive data might be the use of a federated approach to data mobilisation where data is kept at holders' premises, code moves to run the analysis locally, using the output (aggregated data) to perform the overall analyses. Some examples on pseudonymised data can be found in PHIRI [https://www.healthinformationportal.eu/services/phiri-demonstrators] or the Baseline use case in WP5 in ByCOVID [https://doi.org/10.5281/zenodo.7560731].

Data access and use applications

In a majority of EU countries, access to health RWD is granted by the authorising body after the evaluation of a research protocol including a detailed data management plan. Usually the authorising body is a research ethics committee (REC) or a data protection agency (DPA).

Once authorisation is provided, researchers sign contractual arrangements with the access body in particular when commercial entities or commercial interests are at stake. In addition to a data access agreement, a principal researcher can sign a self-declaration committing not to re-identify individuals based on combining shared data with other public or non-public data sources, sub-processing agreements or confidentiality statements. In turn, Access bodies should ensure that access is only provided to requested electronic health data relevant for the purpose of processing indicated in the data access application by the data user and in line with the data permit granted [5].

There is no standard process for these data access applications. This is why, in the forthcoming EHDS legislation, there is a provision for the coordination of data access applications at EU level that might include the use of standard model with contract clauses on data access and use [2]. It might also be the case that, in order to reduce complexity of cross-country data requests, there is the possibility of having a data permit issued by one

[5] *The Article 29 Data Protection Working Party, noted that pseudonymised data cannot be considered equivalent to anonymised data "as they continue to allow an individual data subject to be singled out and linkable across different data sets. Pseudonymity is likely to allow for identifiability, and therefore stayed inside the scope of the legal regime of data protection. (Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 2014)*

concerned health data access body benefiting from mutual recognition by the other concerned health data access bodies [3].

Template for the description and assessment of real-world data access and reuse policies and procedures

In this milestone, we have developed a template for the description and assessment of real-world data access procedures. This template has been developed to formally collect the information on access and use policies and procedures on real-world routine data but may well serve to collect such information in the case of human and clinical data collections in the Deliverable 2.2.

The form contains three sections; the first one containing descriptive high-level meta-data of the data institution and the data source; a second one, providing basic information on access procedures; and, the third one, collecting information on the actual use of data once the user has granted access (see table 1). We have tested this information using three examples, different by nature: LINK-VACC a Belgium registry on COVID vaccination that is built on the linkage of multiple population-based registries; BIGAN, a regional data space containing routine data linked data from EHR and population-based registries; and, HEALTH RI, a national data space aiming the linkage of any type of health data, including research data.

Table 1 Access and user procedures and policies

General elements	
Name	
Url	
Country	
Data source type	
Data type	
Information on data provenance provided?	
Information on quality assurance/quality control mechanisms provided?	
Data access	

Who grants access	
Access type	
Access specifications	
Data use	
Information provided on the preparation of the dataset?	
Trusted Research Environment	
Analytical tools and/or computing sources provided?	
Mechanisms for finalisation and devolution?	

General elements	
Name	**Linking of registries for COVID-19 vaccine surveillance (LINK-VACC)**
Url	https://www.sciensano.be/en/projects/linking-registers-covid-19-vaccine-surveillance
Country	Belgium
Data source type	Population health registry
Data type	Pseudonymous data
Information on data provenance provided?	Yes
Information on quality assurance/quality control mechanisms provided?	No
Data access	
Who grants access	Belgian Information Security Committee Social Security & Health
Access type	Application and access to TRE after approval
Access specifications	Dedicated scientists at Sciensano involved in surveillance activities have access to the linked individual-level pseudonymized data hosted at Healthdata.be's secured research environment (access rights granted ad

	nominatum). Pseudonymized demographical data (only sex and age), data on the vaccinator and data concerning the administered vaccine in breakthrough cases, are shared with the Federal Agency for Medicines and Health Products (FAMHP), the competent agency for the monitoring of vaccine safety. External investigators with a request for selected data should fill in the data request form (https://epistat.wiv-isp.be/datarequest). Depending on the type of desired data (anonymous or pseudonymized), the provision of data will have to be assessed by the Belgian Information Security Committee Social Security & Health based on legal and ethical regulations, and is outlined in a data transfer agreement with the data owner (**Sciensano**, Belgian institute for health). Sciensano can share anonymous or pseudonymized data with other scientists in the framework of national, European and international collaborations
Data use	
Information provided on the preparation of the dataset?	No
Trusted Research Environment	Data from the vaccine registry (VACCINNET+), the TestResult database (COVID-19 Healthdata Database) and the COVID-19 clinical database (COVID-19 Clinical Hospital Survey), are hosted in the secured environment of **Healthdata.be** (a service of Sciensano) where personal patient data are available. Links are organised with the databases external to Healthdata.be (COBRHA, IMA and STATBEL) using the national registry number. The Trusted Third Party (TTP) service of the eHealth platform pseudonymizes the patient's identifier and the data from the six databases are stored in Healthdata.be’s pseudonymised environment (separation of secured environment and research environment). A link between the individual-level data in each of them takes place thanks to the use of a pseudonymized national reference number managed by Healthdata.be under a project mandate.
Analytical tools and/or computing sources provided?	Yes, within the secured research environment of Healthdata.be
Mechanisms for finalisation and devolution?	Yes, Sciensano is responsible for the processing of the data. The processing is based on the grounds of public interest (art. 6.1 (e) of the General Data Protection Regulation (GDPR)) and in particular for data concerning health, for

	reasons of public interest in the areas of public health (art. 9.2 (i), of the GDPR). Reports with the aggregated results of the surveillance will be made public and will be shared with partners and federal and regional ministries of public health. The pseudonymized data will be stored for 10 years, in accordance with the approval of the Sectoral Committee of Social Security and Health.

General elements	
Name	**COVID cohort BIGAN Aragon - ES**
Url	https://bigan.iacs.es/es/inicio
Country	Spain (Aragon)
Data source type	Regional Health Data Space: Public health surveillance cohort / Population health registry
Data type	Pseudonymised data
Information on data provenance provided?	Yes
Information on quality assurance/quality control mechanisms provided?	Yes
Data access	
Who grants access	Health Science Institute in Aragon (IACS)
Access type	Data request application and access through TRE upon project's approval.
Access specifications	Researchers can request data from BIGAN by applying for data access through a standardised procedure upon the approval of their research project by an Ethics Research Committee and commitment to the system level information security policies of the Health Science Institute in Aragon (IACS). Data access is restricted to minimum data set required for completing the specified research purposes and for the duration of the project. Link to BIGAN - data access https://bigan.iacs.es/en/services/data-access
Data use	
Information provided on the preparation of the dataset?	Yes. BIGAN provides help desk support on data access through feasibility assessment, and data preparation

	facilitating research including linkage between available data sources upon request.
Trusted Research Environment	A de-identified, minimised and pseudonymised data set is processed following the data model specifications provided in the data request and served compressed and encrypted through a TRE within the internal network provided by the Goverment of Aragon. Researchers can access the data using their credentials and decompres/decrypt the data using a private key.
Analytical tools and/or computing sources provided?	Yes
Mechanisms for finalisation and devolution?	Data request application implies researchers commitment with BIGAN and IACS information system level security and privacy policies including deleting accessed data upon projects completion and mandatory open publication of research findings, as well as advertising key findings, in order to demonstrate to citizens the use and benefits associated with the analysis of their health data information

General elements	
Name	**Health-RI - enabling data driven health**
Url	https://www.health-ri.nl/
Country	Netherlands
Data source type	NationalHealth Data Space: Population health registry
Data type	Pseudonymised data
Information on data provenance provided?	Yes
Information on quality assurance/quality control mechanisms provided?	Yes (https://www.health-ri.nl/quality-and-security-services-provided-health-ri)
Data access	
Who grants access	Health-RI. Health-RI is a non-profit foundation registered under Dutch law (https://www.health-ri.nl/about-health-ri/governance)
Access type	Data request application and access through TRE upon project's approval

Access specifications	Researchers can request data from Health-RI by registering for a valid account and applying for data access through a standardised procedure upon the approval of their data request by an Ethics Research Committee and commitment to the Terms and Conditions for Health-RI (https://www.health-ri.nl/about-health-ri/action-lines-themes-communities/architecture; https://covid19initiatives.health-ri.nl/p/Dashboard) Request account: TraIT servicedesk@health-ri.nl or ELSI elsiservicedesk@health-ri.nl
Data use	
Information provided on the preparation of the dataset?	Yes (https://www.health-ri.nl/pricing-model-health-ri-services)
Trusted Research Environment	Yes (https://www.health-ri.nl/services/health-ri-service-catalogue)
Analytical tools and/or computing sources provided?	Yes, Health-RI services are detailed and updated in the Health-RI Service Catalogue (https://www.health-ri.nl/services). Health-RI services are offered on a Fair Use basis and with a Freemium Pricing Strategy. The relevant policies are made available through the Health-RI website
Mechanisms for finalisation and devolution?	Yes, access provided only for the duration and with the scope of the project (https://www.health-ri.nl/terms-use-health-ri-services)

Legal references

[1] Art.33 Forthcoming European Health Data Space for Secondary Use[6]

[2] Art. 34 (e) Forthcoming European Health Data Space for Secondary Use[7]

[3] Art. 5 GDPR - General Data Protection Regulation[8]

[4] Art. 44.3 Forthcoming European Health Data Space for Secondary Use[9]

[5] Art. 44.1 European Health Data Space for Secondary Use[10]

[6] Art. 50 Forthcoming European Health Data Space for Secondary Use[11]

[7] Whereas 54 Forthcoming European Health Data Space for Secondary Use[12]

[6] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[7] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

[9] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[10] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[11] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[12] https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF