Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations

2021, 31st May

Enrique Bernal-Delgado on behalf of the WP10 team

Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu|Twitter: @JA_InfAct

This project is co-funded by the Health Programme of the European Union

Table of Contents

Executive summary

Information for Action! is a Joint Action (JA-InfAct) on Health Information promoted by the EU Member States and funded by the European Commission within the Third EU Health Programme (2014-2020) to create and develop solid sustainable infrastructure for EU health information. The main objective of this JA-InfAct is to build an EU health information system infrastructure and strengthen its core elements by three main actions: a) establishing a sustainable research infrastructure to support population health and health system performance assessment; b) strengthening the European health information and knowledge bases, as well as health information research capacities to reduce health information inequalities; and c) supporting health information interoperability and innovative health information tools and data sources.

Methodology: Following a federated analysis approach, JA-InfAct developed an *ad hoc* federated infrastructure based on distributing a well-defined process-mining analysis methodology to be deployed at each participating partner's systems to reproduce the analysis and pooling the aggregated results from the analyses. To overcome the legal interoperability issues on international data sharing, data linkage and management, partners (EU regions) participating in the cases study worked coordinately to query their real-world healthcare data sources complying with a CDM, executed the process mining analysis pipeline on their premises, and shared the analysis results, enabling international comparison and the identification of best practices on stroke care.

Results: The ad hoc federated analysis infrastructure was designed and built upon open source technologies providing partners with the capacity to exploit their data and generate stroke care pathway analysis dashboards. These dashboards can be shared among the participating partners or to a coordination hub without legal issues, enabling comparative evaluation of the caregiving activities for acute stroke across regions.

Nonetheless, the approach is not free from a number of challenges that have been solved, and new challenges that should be addressed in the eventual case of scaling up. For that eventual case, 12 recommendations on the different layers of interoperability have been provided.

Conclusion: The proposed federated analysis approach, when successfully deployed as a federated analysis infrastructure, such as the one developed within the JA-InfAct, can concisely tackle all levels of interoperability requirements from organisational to technical interoperability, supported by the close collaboration of the partners participating in the study. Any proposal for extension should require further thinking on how to deal with new challenges on interoperability.

Key points

The JA-InfAct federated analysis infrastructure is an empirical demonstration of how to put in practice a data-centric interoperable approach to analyse extremely sensitive personal data, such as health data. This kind of infrastructure, namely a data-centric computing infrastructure, is expected to be the most predominant in scenarios with highly restricted access to personal data.

The JA-InfAct experience has demonstrated that, pursuing a successful deployment of such an infrastructure, there is a significant effort to implement interoperability, primarily when it comes to organisational interoperability.

Should the final aim be scaling this infrastructure up, there is also a need to rethink some of the interoperability issues and the solutions provided in this pilot, as new challenges arise with the expansion, and new solutions are subsequently required. Twelve recommendations have been provided in this respect.

Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations

Authors: Enrique Bernal-Delgado,[1] Francisco Estupiñán-Romero,[1] Juan González-García,[2] Javier González-Galindo,[1] Carlos Tellería-Orriols,[2] Micaela Comendeiro-Maaløe,[1] Natalia Martínez-Lizaga,[1] Jane Lyons,[3] Ronan Lyons,[3] Damir Ivanković[4] and Jakov Vuković[4] on behalf of the InfAct Joint Action consortium

1. Data Sciences for Health Services and Policy Research, Institute for Health Sciences in Aragon (IACS), Spain
2. Biocomputing Unit, Institute for Health Sciences in Aragon (IACS), Spain
3. Swansea University, Wales, UK
4. Institute of Public Health, Zagreb, Croatia

Acknowledgments There have been special contributions in the nodes participating in the stroke care use case: Luigi Palmieri (Instituto Superiore di Sanitá, Rome, Italy), Andrea Faragalli (Università Politecnica delle Marche, Italy), Jelena Dimnjaković, Domina Vusio and Marko Brkić (Institute of Public Health, Zagreb, Croatia), Janis Misinš and Irisa Zile (Centre for Disease Prevention and Control, Latvia) and Zeynep Or (IRDRES, France).

I. Introduction

"Information for Action" is a Joint Action (JA-InfAct) on Health Information promoted by the EU Member States and funded by the European Commission within the Third EU Health Programme (2014-2020), including 40 partners in 28 EU and associated countries. The main aim of the JA-InfAct is to build an EU health information system infrastructure and strengthen its core elements by 3 actions: a) establishing a sustainable research infrastructure to support population health and health system performance assessment, b) strengthening the European health information and knowledge bases, as well as health information research capacities to reduce health information inequalities, and c) supporting health information interoperability and innovative health information tools and data sources. One of the underlying tasks has been setting up the pillars for the design, preparation and implementation of a federated research infrastructure that leverages the use of health data to carry out policy-oriented research.

Paramount in the development of a federated research infrastructure (FRI), where data is leveraged from multiple and heterogeneous data sources hosted in multiple sites with different governance models, is interoperability. According to [1], interoperability is defined as "*the ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems*". In the specific case of JA InfAct, interoperability refers to the capacity to capture coherent data from the different partners, being able to reproduce the same analyses and being capable of sharing the results of these analyses.

This definition of interoperability entails different levels of analysis framed in the recommendations report by the European Interoperability Framework (EIF) [1]. The EIF introduces a concise yet clear interoperability model, shown in Figure 1, that classifies the different interoperability elements in four layers that need to be addressed for a successful interoperable (public) service. So, the design, development and implementation of the JA InfAct federated research infrastructure have built upon these four layers.

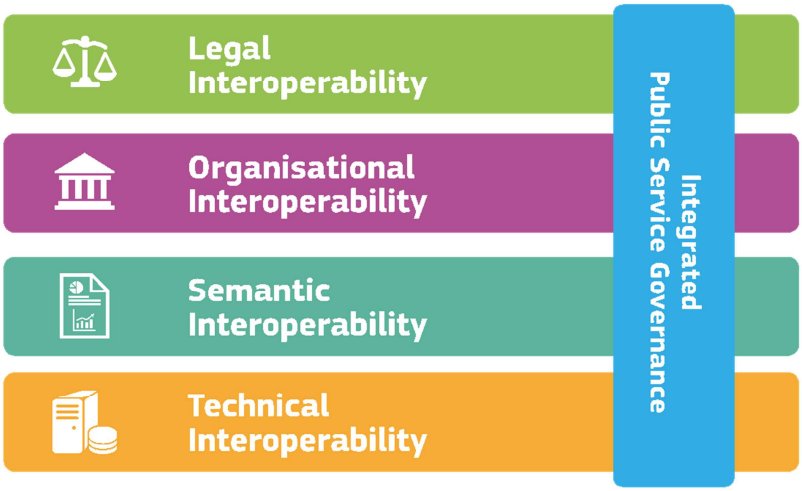


Figure 1. European Interoperability Framework interoperability model.

II. Aim

This report describes the process and challenges, in the context of JA InfAct, of coping with the different layers of interoperability when trying to answer population health research queries in the context of a federated infrastructure. Likewise, the report provides recommendations for the eventual real-life implementation of such a federated infrastructure.

III. Approach

InfAct has addressed the different interoperability challenges to build an FRI, following an approach based on case studies. In Table 1, there is a synthesis of the three use cases deployed through the Joint Action. The selection had mainly to do with different levels of maturity -- from the simplest definition of a study and how it was materialised into the required documentation (i.e., case study on dementia), the definition of a population-based indicator that materialises in a SQL code that is distributed to elaborate in-house indicators (i.e., case study on resilient populations), to a full distribution exercise where a data model is made common, an analytical pipeline developed to be interoperable is distributed among a number of nodes, which run the analyses and produce the expected research outputs that are sent elsewhere (i.e., case study on stroke care pathways). As will be detailed, the stroke care study has provided in-depth insight on how to make interoperability a reality in the context of JA-InfAct federated infrastructure. For the sake of clarity, we will be describing all the interoperability elements in the fully completed stroke case study thoroughly, providing details of the others in an appendix.

In short, the JA-InfAct federated analysis infrastructure is an *ad hoc* infrastructure solution proposed for cross-border analysis. The core element of the analysis infrastructure is the process-mining methodology, where real-world datasets are combined and then processed to generate the care pathway process models within the premises of each participating partner.

The analysis methodology is encapsulated in a software distribution solution that acts as the central element of the federated analysis architecture, leveraging the data capture and execution of the analyses across different partners and the exchange of the results. All the methodologies and solutions are designed and implemented following a *privacy-by-design* approach to fulfil the interoperability challenge: that is, how to work with the different information systems, data sets and software solutions that each of the participating partners has, with complete coherence of the analysis results.

This section describes these three elements, using the care pathways of acute ischemic stroke as a case study. Subsection 2.1 summarises the process-mining analysis, further commented in [2]. Subsection 2.2 details the technical aspects of the federated analysis infrastructure, considering the software and system orchestration elements to reach the desired solutions. Subsection 2.3 focuses on all the interoperability elements that have been considered within the first two elements (the analysis methodology and the analysis infrastructure) and specific organisational agreements between partners for successful deployment of such an infrastructure.

Table 1 JA-InfAct use cases.

Use case	Purpose	Data sources	Common data model (main entities)	Distribution	Hubs
Dementia care	Identification of 1-year follow up contacts and associated costs	Insurance data PC EHR Hospital stays Prescriptions ER data RHB contacts Billing data	Individual patient care provider contact Time stamps	Data model specification (v0.1)	Aragon (ES) France (FR)
Desirable health services utilisation	Elaboration of a population-based health indicator based on the users with the lowest use of health services	Insurance data PC EHR Prescriptions Hospital stays	Individual insuree's residence	Protocol, data model specification and SQL script for data transformation (v1.0)	Wales NHS (UK) Aragon (ES)
Stroke care pathway	Discovery of the actual care pathway for Acute Stroke patients	Insurance data ER data Hospital data	Individual patient care provider Contacts Time stamps Event	Complete solution: Docker with open source Log builder and Process Mining FAIR publication (v14.0)	Aragon (ES) Marché (IT) HU Zagreb (HR) HU Riga (LV)

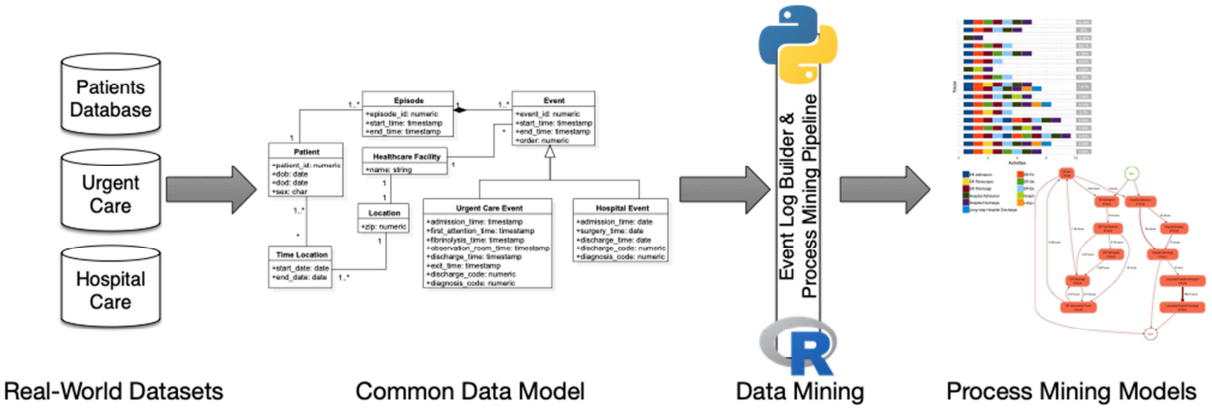
EHR: Electronic Health Recorder; ER: Emergency Records; ES: Spain (*España*); FR: France; HR: Hungary; IT: Italy; LV: Latvia; RHB: Rehabilitation; UK: the United Kingdom.

3.1 Description of the case study: process-mining based analysis methodology

The process-mining based analysis methodology used in JA-InfAct has been previously introduced for a similar use case at a regional level in work by González-García et al [2]. It is based on analysing real-world datasets related to stroke care using process mining techniques [3][4] and, specifically, process discovery. The analysis aims to detect how theoretical care pathways or clinical guidelines, such as the acute ischemic stroke care pathway defined in [5], are implemented in real life.

The methodology is illustrated in Figure 2 and comprises four main elements: 1) capture the real-world data to be further processed, considering that the start of the acute stroke care pathway should be captured in the urgent care unit (i.e., accident and emergency care services) and that hospital care information systems and the patient information database should contain specific details on patient characteristics; 2) transform the data from their specific information systems into a defined CDM, which contains the actual semantics of the contents in the form of different entities (i.e., patients, visits, procedures, etc.), the variables that define entities (i.e., age or sex for patients, visit date and hospital, or procedure date and code), their relationships (i.e., when a patient went to a hospital where he or she received a procedure) along with the encoding systems prevalent in the different nodes (i.e., the International Classification of Disease Version 9 [ICD-9]or ICD-10); 3) process the data stored in the CDM to generate an event log (where each register in the data set represents an activity and its attributes), using the *Event Log Builder* tool that sends input to the final *Process Mining Pipeline* tool,

which generates the empirical process models; and 4) compare and contrast the process models obtained to verify the actual care pathways in the different countries. The outputs of the methodology



are the process models that can be depicted, for example, as process maps that present the real-life transitions between events, the number of patients along each trajectory and throughput times (see the Results section).

Figure 2. Methodology and analytical pipeline supporting the case study on stroke.

3.2 Description of the federated infrastructure

The JA-InfAct federated analysis infrastructure has been designed to distribute the analytical pipelines using the data of the different partners in this case study. Furthermore, the objective is to replicate the process mining methodology without requiring the partners to move any data to other partners or to the coordination hub.

The coordination hub, in this context, is responsible for developing the analysis scripts, supporting each participating partner in script deployment and producing insights from comparing the partners. The term *federated infrastructure* is used because participating partners can act independently without requiring the rest of the partners to perform the analyses.

Figure 3 presents the schema of the distribution workflow. The partners involved in the infrastructure are generally considered *Data Hubs* in the figure. They are responsible for transforming and loading their data sets in the CDM format previously defined and agreed on. As stated initially, this is an *ad hoc* infrastructure, which means that the code distribution is actually handled in two steps: first, the coordination hub encapsulates the process mining analysis scripts into a portable execution environment and, second, each Data Hub gets this portable execution environment, deploys it and runs it in their premises, indicating where the data is located.

The analysis process starts (Point 1 in Fig. 3) when the coordination hub distributes the process mining analysis scripts among the Data Hubs. Next, in Point 2 of the figure, partners fetch the scripts into their systems, indicate the input data placement to the analysis code and run the analyses. As a result of this point, each partner obtains their own results (i.e., stroke care pathways). Then, as shown in Point 3 of the figure, partners send back their results to the coordination hub. This feedback in the

form of local results is sent (in this case manually) by compressing the output dashboard and mailing it to a specific address. Finally, once the coordination hub has gathered information from all partners, it compares them to present a final analysis (Point 4 in Fig. 3).

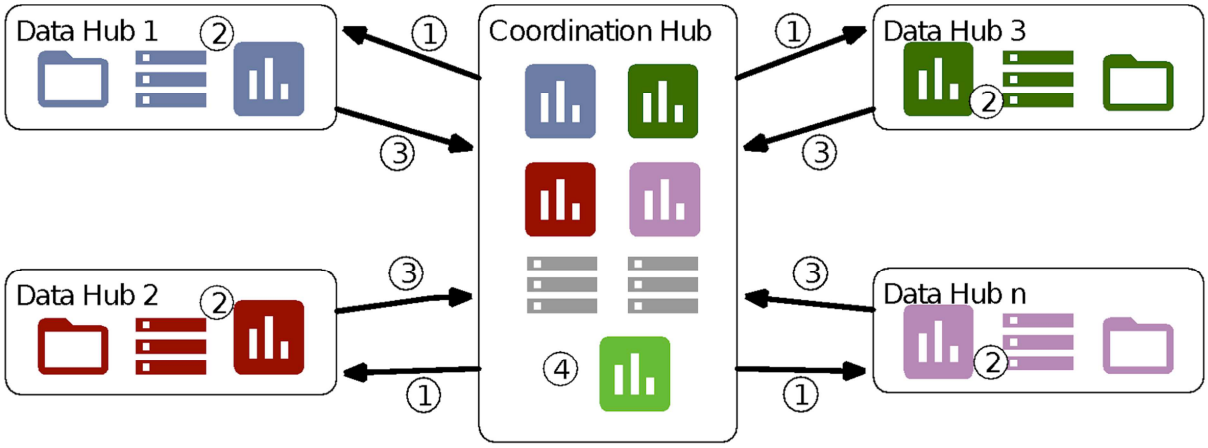


Figure 3. JA-InfAct federated analysis infrastructure.

Note that this architecture follows the *privacy-by-design* principle, in that all data is governed under the legal provisions of the hosting institutions and no individual data is moved outside partner premises. This deals with various relevant legal interoperability issues raising legal barriers to cross-border data sharing. Results and data that partners feed back to the coordination hub are always outputs, in this case, aggregated measures in the form of the process models, and runtime application error logs. In addition, following the *secure-by-design* principle and as a way to build trust among federated infrastructure participants, it is important to note that all the analysis and deployment solutions presented are open source. This means that all the source code of the scripts for data management and analysis are auditable by the participants. This ensures the credibility of what is going to be executed, enhances the reliability of the results and helps to increase the quality of the solution through partner contribution.

3.3 Interoperability layers

What has been described in the two previous subsections relies on guaranteeing interoperability among the different components of the solution, as well as the relationship within the partners involved in terms of trust, governance and legal compliance. In the following subsections, we describe how the different layers of the EIF have been achieved in the context of the JA-InfAct federated research infrastructure.

3.3.1 Legal interoperability. The General Data Protection Regulation (GDPR).

The top layer in the EIF interoperability model is legal interoperability. Given the compliance with GDPR and overarching ethical principles, legal interoperability is about "*ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together*."

As the objective of this federated infrastructure is to analyse care pathways using patient data, the legal frameworks to be considered are those regulating the use of personal health data for research purposes. European legislators have worked intensely on this matter so as to homogenise the use of this kind of data across the EU Member States. The result is the General Data Protection Regulation (GDPR) [6], the EU law that establishes the conditions for the legal interoperability of the JA-InfAct federated infrastructure.

3.3.2 Organisational Interoperability. The JA-InfAct proposal and grant agreement.

The second layer to be discussed is the organisational interoperability layer. Organisational interoperability refers "*to the way in which public administrations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals.*"

The JA-InfAct federated analysis infrastructure was designed as client-server infrastructure, with a coordination hub orchestrating the whole process (consolidation of the data model and development of the analytical pipeline, including technological solutions) and a node counterpart where a contact person performs the following duties: 1) detect the staff with the knowledge to perform the required tasks; 2) attend work meetings to ensure the proper coordination of the work; 3) commit the necessary resources to develop the work; 4) provide the required feedback to improve and solve any possible issues that appear during the development; and 5) carry out the analyses and feedback and interpret the results considering the local context.

3.3.3 Semantic interoperability. The common data model and data codifications.

Semantic interoperability ensures that "*the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties*". So, in practice, it is necessary to guarantee that, when different partners perform care pathway analyses, the data and the results refer to the very same caregiving processes.

The cornerstone of the semantic interoperability layer is the common data model (CDM), in which the data entities and their relationships are defined and which serves as the common storage for further analyses. The CDM design was iteratively refined to express the actual caregiving settings in the different partners. For example, initial versions of fibrinolysis treatment and the thrombectomy procedure (crucial for a rapid response in ischaemic strokes, were not considered to be part of urgent care events, while in further refinements these activities were included in this setting. At the moment of writing this paper, the 14th revision of the CDM has been finished.

It is important to note that, as a fundamental component of the CDM apart from the entities and relationships, how the information is codified within the data model variables is also defined. Consequently, the encoding systems or standards used in the different nodes of the federation are established.

3.3.4 Technical interoperability. The Event Log Build and the Process Mining Pipeline

The technical interoperability layer covers *"the applications and infrastructures linking systems and services. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols*"

The most important element in the technical interoperability layer of the JA-InfAct federated analysis infrastructure is the "*deployment package system.*" That is, this system establishes how each source code of the analysis scripts is encapsulated so as to be easily transmitted from the Coordinator Hub to the partners and easily executed by the partners to perform the analysis on their premises and so the partners can easily transmit the results back to the coordination hub.

As a final point regarding technical interoperability, it is important to highlight that the analysis code included in the Docker image deployed relies on having the input data of each partner in the CDM format. It is the responsibility of each partner to create the Extraction, Transformation and Load (ETL) processes that capture the required data from their health information systems and complete the CDM according to the definitions agreed.

IV. Results

The JA-InfAct federated infrastructure, as designed, has been able to yield the intended output in the four nodes comprising the case study on stroke (each node has been able to produce a dashboard depicting the real-life care pathways in the four countries). In this section, we explain the steps shown in Figure 1: specification of the CDM, distribution of the analytical pipeline, implementation of the pipeline in the different nodes and collection of the outputs.

4.1 Specification of the data model

The research query sent out to the federation materialised in: 1) a definition of the cohort of patients as those who, in the period of study, had been admitted with symptoms of stroke in an emergency ward; 2) because of the different theoretical pathways, a classification of patients as confirmed ischaemic stroke or haemorrhagic stroke patients; and, 3) the definition of the different activities (namely, events) that patients would follow in their journey from admission to discharge (Table 2).

As a second step, coding experts and neurologists among different partners were consulted to select the codes that conceptually better fitted the cohort definition. In our case, and for the nodes included in the exercise, we just needed to build interoperability between ICD-9 [7] and ICD-10 [8] codes.

Table 2 Types of activities considered in the process mining analysis

Activity Name	Activity Description
ER Admission	Administrative admission to Emergency Room Department
ER First Attention	First contact with an MD in the ER Department
ER CT	Computed Tomography Scan imaging at ER Department
ER Fibrinolysis	Fibrinolysis infusion at ER Department
ER Thrombectomy	Thrombectomy at ER Department
ER Observation Room	Observation room stay at ER Department
ER Discharge	ER Department administrative discharge
ER Exit	ER Department physical exit
Hospital Admission	Hospital administrative admission
Hospital Fibrinolysis	Fibrinolysis infusion during hospitalisation
Hospital Thrombectomy	Thrombectomy during hospitalisation
Hospital Discharge	Hospital administrative discharge
Long-stay Hospital Admission	Long-stay (recovery) hospital administrative admission
Long-stay Hospital Discharge	Long-stay (recovery) hospital administrative discharge

Once the cohort of patients was defined and the potential activities were identified and discussed, the data model was iteratively refined to express the actual caregiving settings in the different partners, so as to produce a CDM. For example, initial versions of the fibrinolysis treatment and the thrombectomy procedure (crucial for a rapid response in ischaemic strokes) were not considered to as part of urgent care events, while in further refinements these activities were included in this setting. At the moment of writing this paper, the 14th revision of the CDM has been finished.

As a final result, the CDM that supported the development of this case study comprised five main entities: patient, care provider, contact place, events (activities) and timestamps. The logic data model is shown in Figure 4.

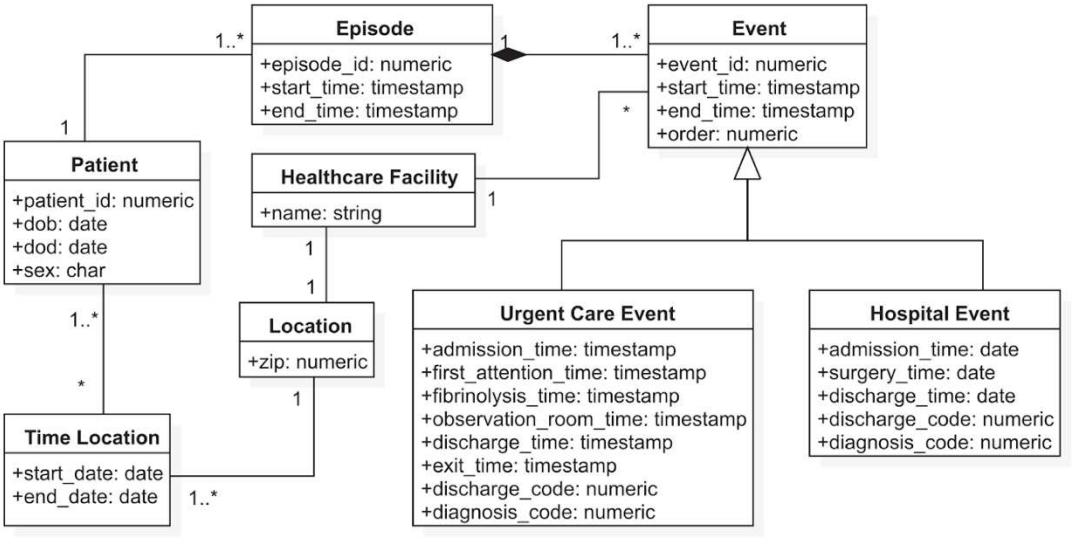


Figure 4. Stroke care logic data model.

The CDM for this case study can be found at http://doi.org/10.5281/zenodo.4879504.

4.2 Analytical pipeline

Once the data model was common to all the nodes in the federation, the orchestration orcommand node developed and distributed the analytical pipeline. In this case, two main pieces were distributed: a) a log builder script on Python, and b) the process mining script on R.

The coordination hub packaged both pieces using Docker as technology [9]. Docker is a compendium of existing Linux-based technologies that, in simple terms, is able to create isolated execution environments where software developers can guarantee that: 1) they will be exactly the same wherever they are executed; 2) code dependencies are easily managed; and 3) the deployment of the execution environments in different locations (in the current case, the different partner premises) is transparently managed. The first point is ensured by using operating system containerisation (that is, creating a virtual operating system), while the second point is guaranteed by having (nearly) infinite set packages for existing code libraries. The third point is handled by having an environment

exchange hub solution, publicly under the Docker Hub or privately deployed by software developers on their premises, where the execution environments are uploaded by their authors and easily downloaded by the users. For the purposes of this case study, we adapted the Docker solution used in another project, available at https://hub.docker.com/repository/docker/iacsbiocomputing/ictusnet_analysis.

4.3 Implementation of the pipeline and production of research outputs

The main tangible outcome of the JA-InfAct federated infrastructure is a number of interactive dashboards generated on each partner premises consisting of the actual care pathways followed by patients with a suspicious stroke in Marché (IT), Riga (LV), Zagreb (HR) and Aragon (ES).

The interactive dashboards contain the sequence of activities followed by the patients in the cohort at each of the sites (namely, process traces in Figure 5). Next, they contain the process map with the number of patients moving throughout the care pathway (Figure 6) and the throughput times among activities (Figure 7). Finally, they also contain the precedence matrix, a presentation of the information present in the process map with frequency information (frequency of transitions between caregiving activities analyses), but in a matrix form (Figure 8). Once the information has been gathered in the coordination hub, the work executed in the interoperability elements guarantees that the results can be compared.

Once the different nodes have been able to implement the analytical pipeline and produce the aforementioned outputs, dashboards are sent back to the Coordination Hub for further analyses. The different dashboards, with outputs for all patients, ischaemic strokes and haemorrhagic strokes can be consulted at http://doi.org/10.5281/zenodo.4878081.

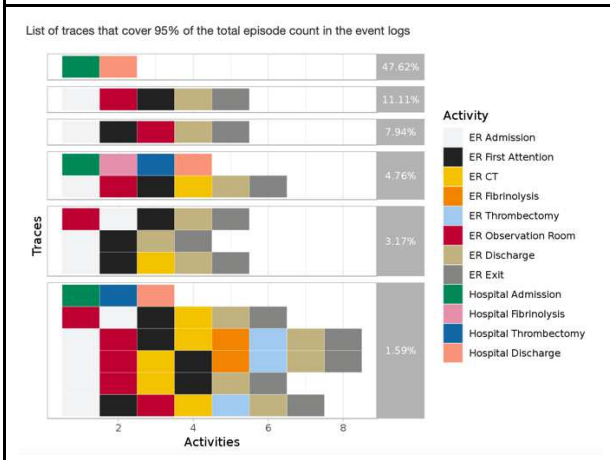
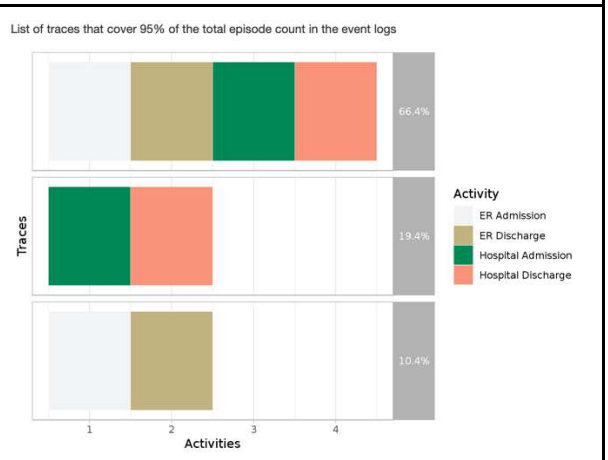
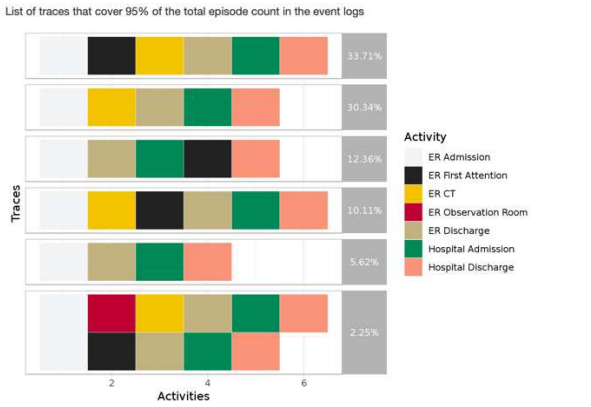
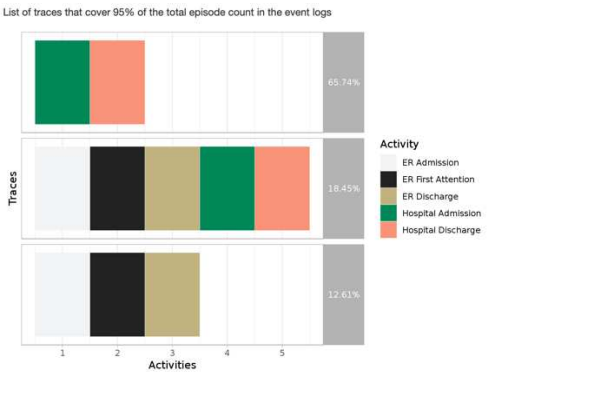
Aragon (ES)	Marché (IT)
	
Riga (LV)	**Zagreb (HR)**
	

Figure 5. Process traces.

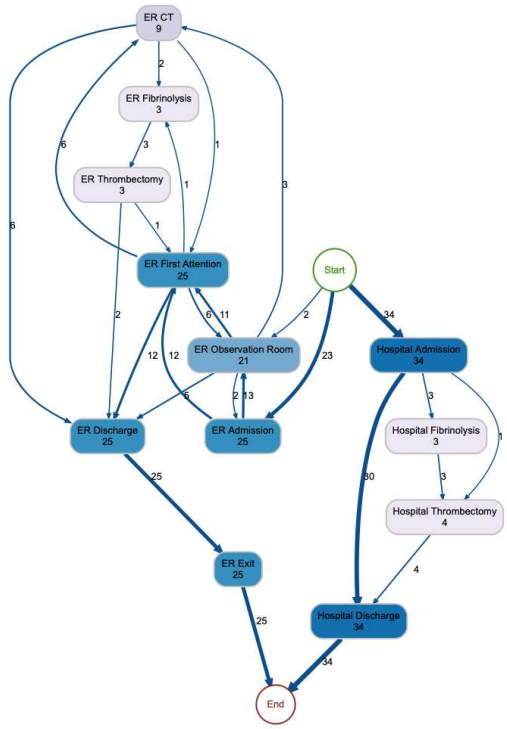
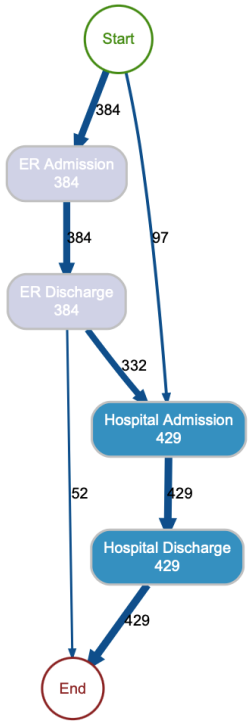
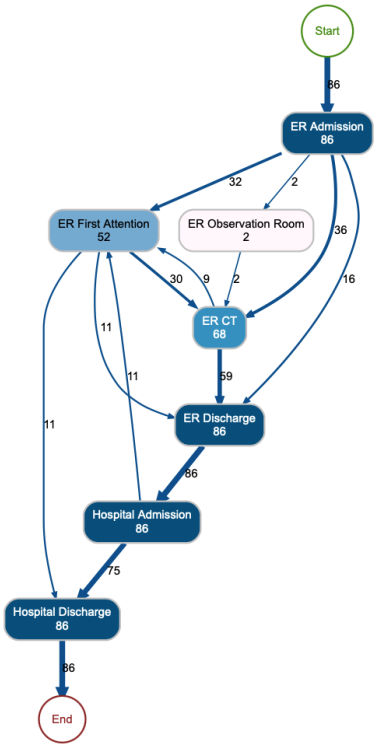
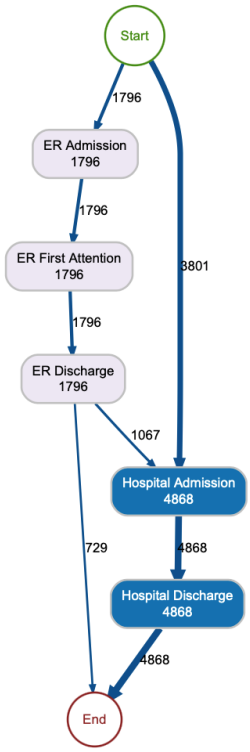
Aragon (ES)	Marché (IT)
	
Riga (LV)	**Zagreb (HR)**
	

Figure 6. Number of patients across the pathway.

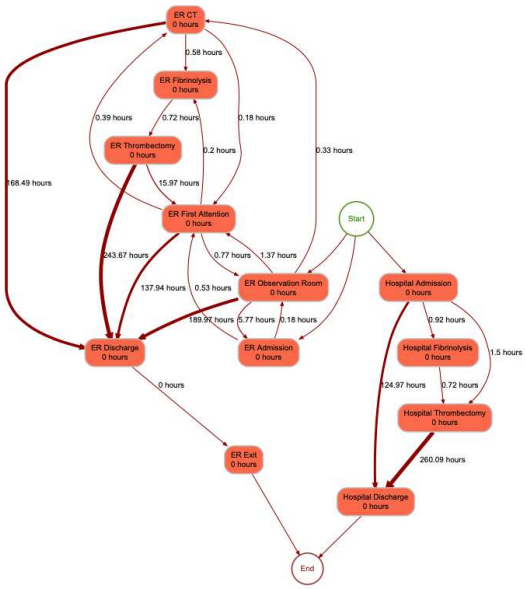
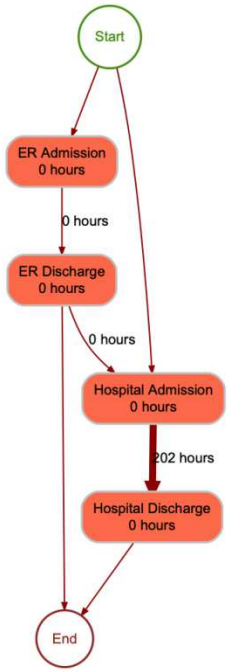
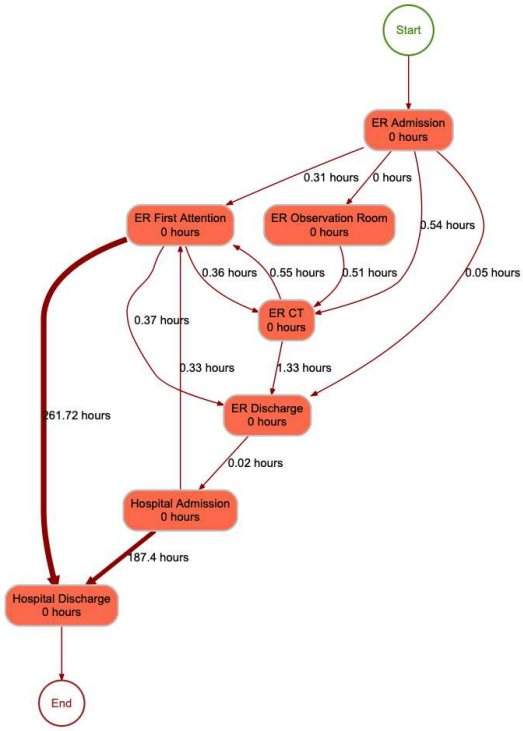
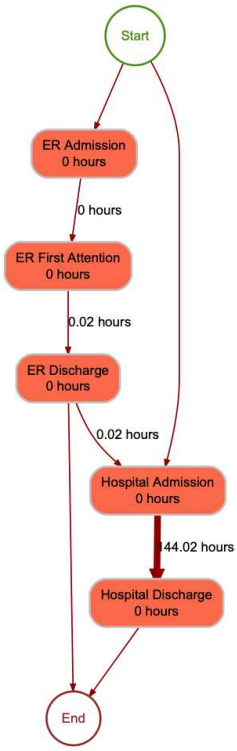
Aragon (ES)	Marché (IT)
	
Riga (LV)	Zagreb (HR)
	

Figure 7. Throughput times.

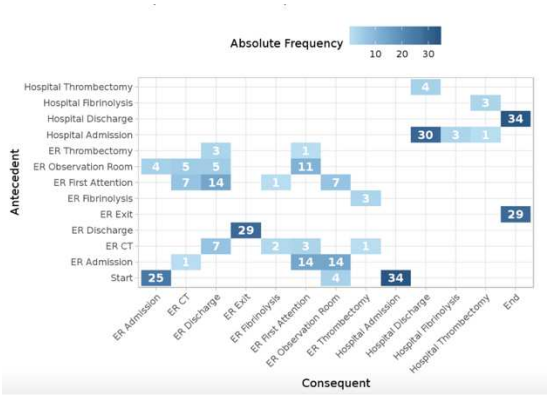
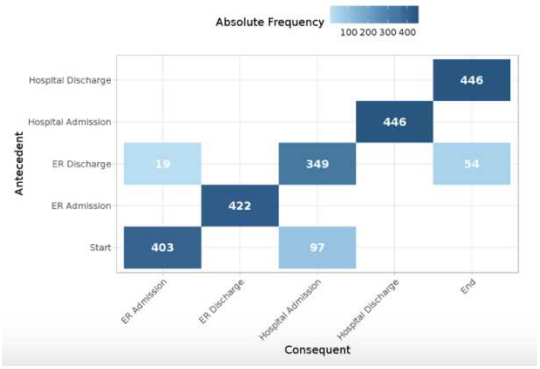
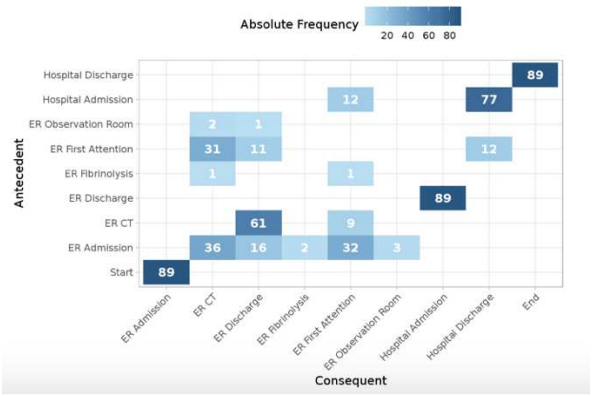
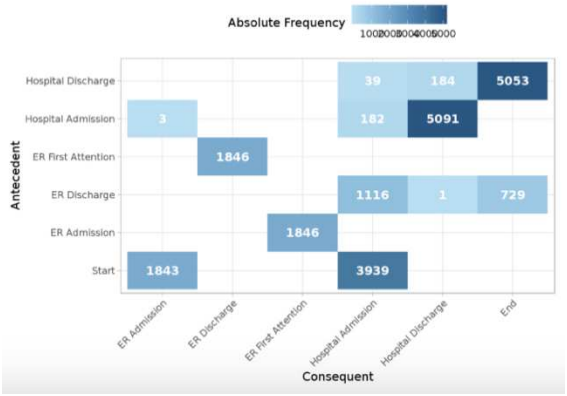
Aragon (ES)	Marché (IT)
	
Riga (LV)	**Zagreb (HR)**
	

Figure 8. Matrix of precedence.

V. Implications and challenges

Having completed the exercise, we have demonstrated the feasibility of the JA-InfAct approach to federated analyses. All the layers of interoperability have been successfully considered in this case study. The exercise has provided meaningful insight into the difficulties involved in the implementation of such a federated infrastructure and has shed light on the ways forward. In the following section, we specifically address the challenges faced and provide recommendations on how to tackle them in the eventual development of a federated research infrastructure.

5.1 Challenges in legal interoperability

In the JA-InfAct federated analysis infrastructure, GDPR compliance has been materialised avoiding the use of personal data during the analysis, minimising the data requirements to complete the case study and providing reliable procedures on how to manage the data at the participating partner's premises under their governance and security assurance procedures. All these concerns are captured in the CDM and how it is completed and processed. The CDM and the data processing designed in [2] had several characteristics to ensure GDPR compliance, especially those related to Art. 5 (related to personal data processing) and Art. 89 (regulating the research uses of personal data), as follows: 1) the data has been used for only the agreed analyses between partners ('*purpose limitation'*); 2) the data gathered has been limited to only those variables required for the analyses and the analysis time-frame ('*data minimisation*'); 3) the personal data has been stored using pseudonyms to avoid possible patient re-identification ('*confidentiality'*), and 4) patient data are stored on only the premises of the partners which have the mandate/responsibility of curating such original datasets ('*storage limitation*').

5.2 Challenges in organisational interoperability

Organisational interoperability is rooted in the willingness of the Data Hubs to participate and cooperatively respond to a research question that is relevant to the participants. It is labour-intensive and requires almost continuous communication between all the partners and the Coordination Hub. It is necessary to build a trust relationship based on the adjustment of use case specifications considering all inputs from participating partners, reaching consensus on each step of the process, and allocating tasks to staff with relevant skills in each organisation through the different steps in the process. Lastly, there has to be complete transparency in all developments aimed at producing a fully reproducible implementation of the analytical pipeline.

The JA-InfAct federated infrastructure has been designed as a client-server infrastructure with a coordination hub that orchestrates or commands and supervises the whole process in a strictly controlled modest case study. Key elements in terms of organizational interoperability have been to achieve a common understanding of the procedures and the roles of each of the parties, including data access; and to face the challenges involved in the technical deployment of the infrastructure.

As for the first element, it has been important to take the time to clarify the roles of the different actors in the federation. It has been necessary to extensively explain the coordination hub, as the one orchestrating all the procedures: developing the data model, implementing the analytical pipeline,

adapting the technological solutions to the computational environment of each node, acting as a *help desk* with the contact persons in each node, and supervising the process so as to synchronise the work in the different nodes. Most importantly, it has been necessary to explain the roles and requirements for the nodes. They are the ones that were basically in charge of providing information on the actual data access and the ability to comply with the CDM (providing insight to adapt it to local circumstances) and effectively access the required data. Above all, the nodes have to set up a team in house that is able to comply with the requirements of the case study, particularly the deployment of the technological solutions.

As for data access challenges, the case study was designed to get the intended outputs with a very simple data model and rather limited data requirements. Consequently, in this very controlled context, questions such as linkability of data sources, insufficient coverage or lack of relevance of the data sources, and more in-depth data quality elements at variable level such as incompleteness, missingness or systematic errors have not been deemed to be significant challenges to deal with in this exercise.

So, the main challenge in the deployment of the federation was expected to be (and it has been) the need for technical capacities in some of the infrastructure nodes. This is particularly true for needing individuals with IT profiles that can easily implement the technological solutions and software developed by the coordination hub. The coordination hub has taken on an extra effort in supervision and capacity building that would not be sustainable in the eventual scaling up of the infrastructure.

5.3 Challenges in semantic interoperability

As in any cross-national comparative research, semantic (and syntactic) interoperability is the main challenge. Each of the entities composing the data model (patient, contact, event, time) are subject to threats to semantic (and syntactic) interoperability. As an example, the definition of the cohort (what is stroke); how specific or sensitive the definition of ischemic or haemorrhagic stroke is, the definition of an episode, the identification of the care activities and where these activities are provided and if these activities can be separated out across care providers, the uneven granularity of the timestamps, or the definition of exiting the process.

The effort made to achieve the CDM has implied understanding the care processes in the different nodes of the federation, agreeing on common concepts (and then, definitions) for the different attributes within entities, building the appropriate cross-walks if there were different standards or encoding systems, transforming the variables when needed to a common format or, in the worst scenario, reaching a minimum common denominator.

In this exercise itself, the main threats to semantic (and syntactic) interoperability that the coordination hub had to solve have been:

- Reaching a consensus on the specification of the relevant activities to map in terms of acknowledging the hyper-acute care process in stroke (i.e., relevant therapies such as fibrinolysis or mechanical thrombectomy, but also CT imaging, etc.)
- Reaching a consensus on the classification of a stroke as ischaemic, transient ischaemic or haemorrhagic using both ICD-9th and ICD-10th.
- Defining standardised dictionaries for certain concepts such as ‘discharge_code’ based on mapping all existing ‘discharge_codes’ in each partner with similar descriptors.

- Setting a common default timestamp granularity and establishing the rules to comply while being consistent with the care process even when such level of granularity was not available at each site (i.e., requiring date-time granularity with 00:00:00 time when only the date was available and checking the consistency of the timestamps on an expected sequence of activities to assess irregularities).
- Establishing a normalised file format (comma separated value [CSV] file, pipe separated, without quotation marks) and encoding (Unicode Transformation Format [UTF]-8) with pre-set headers and variable names fitting the CDM specifications.

5.4 Challenges in technological interoperability

As stated previously, the Docker-based deployment relies on the availability of Linux servers among partners. They are needed to guarantee full technical interoperability, i.e., the analysis codes should be able to run independently of the systems available in the partners' sites.

However, some partners do not have Linux operating systems. The coordination hub had to implement a set of Virtual Machine images containing a Linux operating system; the scripts to fetch the Docker images and run the images were also created, in Open Virtualisation Format version 2 [10] and Virtual Hard Drive version 2 format [11]. These Virtual Machines can be deployed in virtually all commercial systems (Microsoft Windows, Apple macOS, multiple *NIX variants, etc.) and have demonstrated their utility during project development. For the purposes of this case study, we used an adaptation of the Virtual Machine used in [12].

VI. Upcoming challenges and recommendations

6.1 In legal interoperability

In the real-life implementation of a federated approach, where there would be many more nodes with different responsibilities for data curation and management, many more data sources could be used and data requirements could be larger in a context where research questions and data queries grow exponentially. In this case, assuring the compliance with GDPR principles −in particular minimisation and confidentiality− gains relevance and imposes new actions. In this context, Data Protection Officers (DPOs) will play a major role.

Recommendation 1: In the real-life expansion of JA-InfAct, data access will require documentation of how GDPR principles will be assured. This will mainly be guaranteed through: a) a protocol of the study behind the query (including the purpose and methodology) and a data management plan including the data schema (entities, variables, operational description with categories and values, and encoding systems), and b) what the measures to assure confidentiality and minimisation are, who the actors will be, what they will be paid for data management and for how long.

Recommendation 2: In the context of the nodes, the DPOs will need to understand how data accessing and data management procedures will work in the context of a federated approach. Specific training programs for DPOs could be recommendable. Conversely, the continuous exchange with DPOs will make each node aware of the local and specific requirements and anticipate the data accessing needs.

Recommendation 3: In a scaled context, there will be a need for technological solutions that ensure privacy and safety by design. There will be important authentication and authorisation features to limit data access to only authorised users and to provide information for forensic analyses in the follow-up of a given user.

6.2 In organizational interoperability

In the JA-InfAct case study, the number of actors interacting has been confined to a few: in the coordination hub, a technological and a domain expert; in the different nodes, one or two contact persons with mixed profiles. In this strictly controlled case study, bilateral interaction between the coordination node and the four participating nodes, close monitoring of the process, and even remote online intervention could be used to solve queries on the deployment of the technological solutions.

As mentioned when describing the federated research infrastructure, a good number of tasks are developed in-house by each of nodes within the federation. Examples are discussing the research question, agreeing on a CDM, accessing and collecting the data in the way required by such a data model, deploying the technologies developed elsewhere in their technological infrastructures, running the scripts, and interpreting the error logs and the outputs.

The need for organisational interoperability will sky rocket in special circumstances: in a federation with many more nodes, or in a hybrid federation with one node serving as an orchestrator of other nodes, or in a peer-to-peer federation where any node can orchestrate or any node can interrogate the federation.

Recommendation 4: In the context of a scaled up infrastructure, nodes in the federation will require individuals to cover a number of profiles: domain experts (depending on the research question), data scientist, data manager and SysAdmin engineer. The coordination hub requires, in addition, an IT profile expert in distributed computing.

Recommendation 5: Orchestrating the entire distribution in more complex federations will require a stepwise approach (see details in Appendix 2) that smooths the exchange between the coordination hub and the nodes, while deploying an analytical pipeline that is transparent and reproducible at any step.

Recommendation 6: In the institutions composing the federation, rating data curation institutions according to their procedures to get data up-to-date and high quality; agreeing on a common data quality framework (see, for example, [13]), cataloguing their data sources in a way that is standard (e.g., DCAT [14]); providing information on interoperability standards and reusability; and publishing clear procedures to access to their data.

6.3 In semantic interoperability

Data requirements within JA-InfAct case studies have been intentionally limited and, consequently, the number of data sources and the type of data have been restricted. Achieving a CDM has therefore been rather uncomplicated. An extended version of JA-InfAct federated infrastructure that is expected to interact with unlimited research questions will require considering multiple data sources and many more types of data. Some of them may come from routine collections; for example, administrative or claim data as we have used in the stroke case study, disease-specific registries, population-based registries, socioeconomic repertoires, electronic and medical health records, data from lab tests, data from imaging tests, etc. Some of these data may come from *ad hoc* data collections; for example, samples of human genes, biosamples, data from wearables, samples of texts, data from social media.

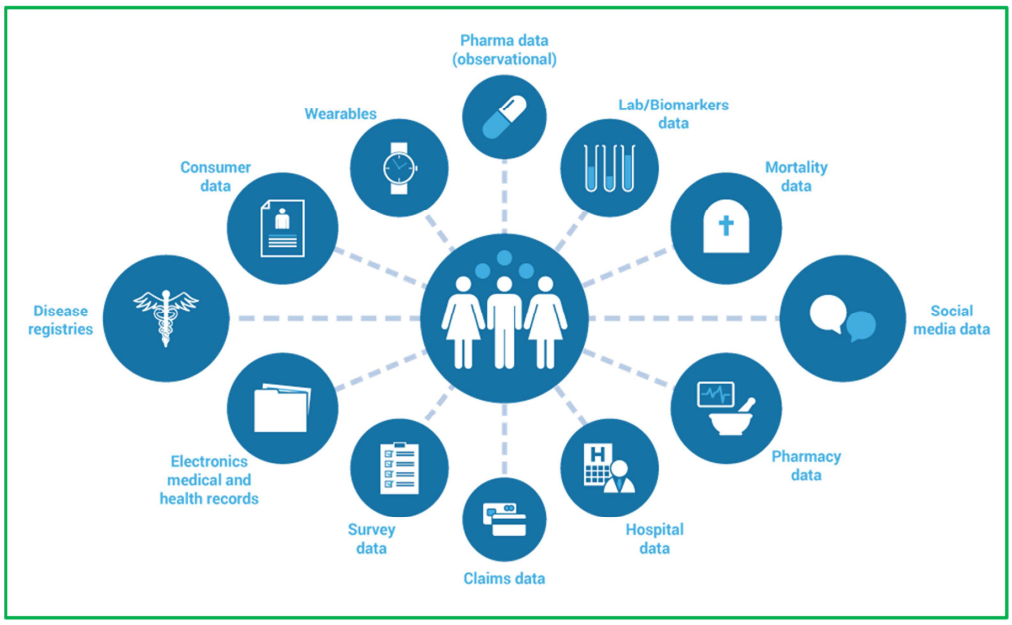


Figure 9. Data sources in the scope of a population health research infrastructure.

In addition to the variety of data sources, there is also the challenge of the heterogeneity of data in their very nature (at one end, administrative data; at the other end, natural language) but also heterogeneous in the encoding systems. Consequently, if the data model for the stroke use case is expanded to incorporate new data including new countries, the number of encoding systems would have to be more. For example, if Norway, the UK, Austria and Slovenia had joined, we would have needed to map in NOMESCO [15], OPCS [16], Leistsungkatalog [17] and ACHI [18], respectively. Or if data from lab tests had been needed, a new standard such as LOINC [19] would have been mapped out.

Recommendation 7. When it comes to the expansion of the InfAct federated infrastructure, it would be recommendable to map out and catalogue the most prevalent semantic interoperability standards. In that sense, future initiatives should link to standards developers and curators. An example would be SNOMED [20], the ontology of reference terms for medical conditions.

Recommendation 8. The future JA-InfAct federated infrastructure should link with the existing research infrastructures on health data. On the one hand, to learn how they have catalogued the standards of semantic (and syntactic) interoperability. On the other hand, to provide access to their standards to the population health research community that could be interested in data models including that variety of data sources. As examples, standards on biosamples [21] or molecular biology "omics" [22].

Recommendation 9. In any eventual future JA-InfAct research infrastructure, the vast majority of studies will be observational. A major multiparty initiative pursuing a CDM for observational research is OMOP [23]. A close follow-up of this initiative is recommendable, even proactively advocating improvements to get the specificities of population health research well represented in the OMOP CDM.

6.4 In technological interoperability

The aforementioned technological elements in the JA-InfAct stroke case study had a very modest scope. The study included distributing an analytical pipeline programmed using open source code (R and Python) within a Docker container (and Virtual Machines when needed), and producing output collated by the coordination hub with no further meta-analysis, although setting up the foundations for the distribution of more complex pipelines. The eventual expansion of a federation such as the one tested in InfAct would require a technological upgrade considering three elements: reducing human interaction in the steps proposed in 4.2.2, considering the possibility of heavier computational processes, and designing the architecture to allow full distribution of complex methodologies.

The JA-InfAct federated analysis infrastructure can be considered a step towards more sophisticated solutions. It is a reliable solution for a problem-specific scenario, but the foundations may be easily extended to include more analysis pipelines. For example, a generalised version of the infrastructure can support fully distributed statistical algorithms [24][25] and, in the final term, state-of-the-art federated learning algorithms [26][27][28], the current cutting-edge analysis approach when leading with huge data sets distributed across multiple locations, without having the possibility of merging them. In addition, the current client-server architecture, which relies on a coordination hub that agglutinates a high level of responsibility, can be moved to a peer-to-peer architecture, where all partners/Data Hubs can act as peers, having the capacity to coordinate analyses through the infrastructure.

Recommendation 10. When it comes to reducing human interaction, a way forward will be developing and implementing a user interface between the coordination hub and the different nodes that automates the activities included in the stepwise process presented in Appendix 2.

Recommendation 11. One of the tasks of the coordination hub in an eventual expansion of the JA-InfAct federated infrastructure should be the assessment of the computational needs of the different research queries. Instead of having and maintaining high capacities in-house, the way forward for the infrastructure will be linking with European providers of these services. At this time, noteworthy providers are EGI (computational capacities https://www.egi.eu) and EUDAT (storage capacities https://www.eudat.eu/).

Recommendation 12. In federated infrastructures, problems with the distribution of analyses become paramount when research questions and research methodologies become more demanding. To deal with these new and growing requirements, a future federated infrastructure, learning from distributed machine learning analyses (i.e., methods such as Bagging, Boosting and Stacking), should foresee, design and implement the architecture and analytical pipelines that support model assembly.

To conclude, it is important to note that all the know-how gathered during the development of the JA-InfAct federated analysis infrastructure and some of the recommendations provided are currently being implemented in population health information research infrastructure (PHIRI) [https://www.phiri.eu], a practical roll out of the distributed infrastructure on population health research (DIPoH), a current candidate for incorporation into the European Strategy Forum on Research Infrastructures (ESFRI) roadmap. In addition, all this insight is playing a fundamental part of the European Health Research and Innovation Cloud, a cloud for health data exchange between

European health research infrastructures and health services, to be designed under the HealthyCloud project [https://healthycloud.eu]. Finally, this knowledge is currently helping to give shape to the future European Health Data Space, the project to regulate the secondary use of health data across Europe, under the framework of the Joint Action Towards European Health Data Space (TEHDaS) project [https://tehdas.eu].

References

[1] Directorate-General for Informatics (European Commission), "New European Interoperability Framework," Brussels, Nov. 2017. doi: 10.2799/78681.

[2] J. Gonzalez-Garcia, C. Telleria-Orriols, F. Estupinan-Romero, and E. Bernal-Delgado, "Construction of Empirical Care Pathways Process Models From Multiple Real-World Datasets," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 9, pp. 2671–2680, Sep. 2020, doi: 10.1109/JBHI.2020.2971146.

[3] W. M. P. van der Aalst, *Process Mining: Data Science in Action*, Second Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[4] R. S. Mans, W. M. P. van der Aalst, and R. J. B. Vanwersch, *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*. Cham: Springer International Publishing, 2015.

[5] C. R. Gomez, M. D. Malkoff, C. M. Sauer, R. Tulyapronchote, C. M. Burch, and G. A. Banet, "Code Stroke. An attempt to shorten in hospital therapeutic delays," *Stroke*, vol. 25, no. 10, pp. 1920–1923, 1994, doi: 10.1161/01.STR.25.10.1920.

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

[7] World Health Organization, *International classification of diseases: [9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization, 1978.

[8] World Health Organization, *International statistical classification of diseases and related health problems. 10th Revision*, 5th ed. Geneva, 2015.

[9] C. Boettiger, "An introduction to Docker for reproducible research," in *Operating Systems Review (ACM)*, Jan. 2015, vol. 49, no. 1, pp. 71–79, doi: 10.1145/2723872.2723882.

[10] Distributed Management Task Force Inc., *Open Virtualization Format Specification (DSP0243)*. 2015.

[11] Microsoft Corporation, *Virtual Hard Disk v2 (VHDX) File Format, 5th Revision*. 2021.

[12] González-García, Juan, Estupiñán-Romero, Francisco, Tellería-Orriols, Carlos, & Bernal-Delgado, Enrique. (2020, July 20). Stroke Process Mining Analysis Virtual Machine Image (OVFv2 / VHD formats) (Version Data Model Slides v13 (Effective Data Model v11). Run Slides v7.). Zenodo. http://doi.org/10.5281/zenodo.3952495

[13] Health Data Research UK Data Quality and Standards Strategy Green Paper for Consultation September 2019; available at: https://ukhealthdata.org/wp-content/uploads/2019/09/HDRUK_Data_Quality_Standards_Green_Paper_Sept2019circulate.pdf (last access, 2021, 31st May).

[14] DCAT Data Application Profile for Data Portal in Europe, European Commission 2021, available at https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en (last access, 2021, 31st May).

[15] N.-N. Nordic Council of Ministers, NOMESCO Classification of Surgical Procedures (NCSP), Version 1.14. Copenhagen, 2009.

[16] National Clinical Coding Standards OPCS. Terminology and classifications delivery service. NHS Digital, 2021. Available at: https://classbrowser.nhs.uk/ref_books/OPCS-4.9_NCCS-2021.pdf (last access, 2021, 31st May).

[17] Leistsungkatalog, Ministry of Health Austria, 2021. Available at https://www.sozialministerium.at/public.html (last access, 2021, 31st May).

[18] ICD-10-AM/ACHI/ACS, IHAP 2021. Available at https://www.ihpa.gov.au/publications/icd-10-amachiacs-ninth-edition (last access, 2021, 31st May).

[19] Common LOINC Laboratory Observation Codes, LOINC, 2021. Available at https://loinc.org/usage/obs/ (last access, 2021, 31st May).

[20] K. A. Spackman, K. E. Campbell, and R. A. Côté, "SNOMED RT: A Reference Terminology for Health Care," J. Am. Med. Informatics Assoc., vol. 4, no. SUPPL., pp. 640–644, 1997, Accessed: Apr. 27, 2021. [Online]. Available: /pmc/articles/PMC2233423/?report=abstract.

[21] BBMRI-ERIC standardization services Available at https://www.bbmri-eric.eu/services/standardisation/ (last access, 2021, 31st May).

[22] ELIXIR-ERIC interoperability and standards. Available at https://elixir-europe.org/services/tag/interoperability-and-standards (last access, 2021, 31st May).

[23] OHDSI, Observational data sciences and informatics, 2021. Available at https://ohdsi.github.io/TheBookOfOhdsi/ (last access, 2021, 31st May).

[24] P. Shi, P. Wang, and H. Zhang, "Distributed Logistic Regression for Separated Massive Data," in *Communications in Computer and Information Science*, Sep. 2019, vol. 1120 CCIS, pp. 285–296, doi: 10.1007/978-981-15-1899-7_20.

[25] C.-L. Lu *et al.*, "WebDISCO: a web service for distributed cox model learning without patient-level data sharing," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 6, pp. 1212–1219, Nov. 2015, doi: 10.1093/jamia/ocv083.

[26] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, Jan. 2019, doi: 10.1145/3298981.

[27] K. Bonawitz *et al.*, "Towards Federated Learning at Scale: System Design," 2019, [Online]. Available: https://arxiv.org/abs/1902.01046.

[28] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.

Appendices

Appendix 1

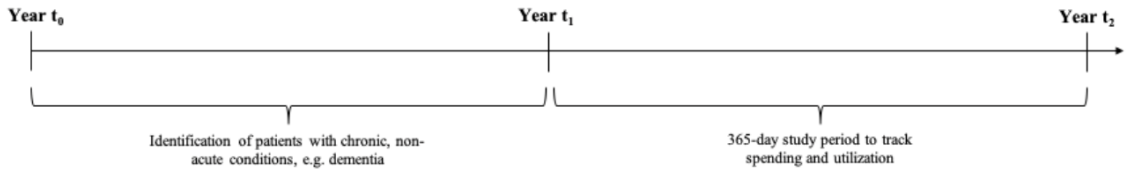
Three population health case studies were piloted, all building on the secondary use of routine health data at real-world settings requiring different stages of deployment of interoperability solutions to enable international comparison at some level. In this appendix, some details are provided on the other two case studies, whose scope was: 1) building a preliminary version of a data model for the study of dementia care; and 2) building a matured version with a script for distribution of the data requirement of a population-based indicator of desirable care. In the following section, the documentation for these two case studies is provided.

1. Dementia care:

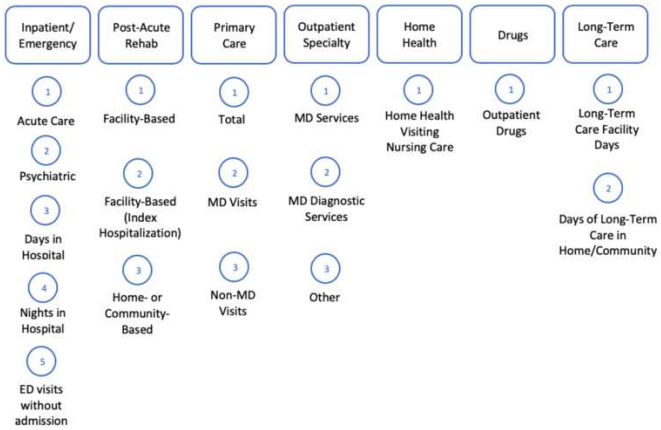
Purpose: Setting up a version 0.1 of the data requirements for the study of utilisation and associated costs in patients with dementia in two countries, (Aragon) Spain and France.

Definition of dementia case: individual older than 65 with a diagnosis of dementia, irrespective of the underlying reason, that has contacted any point of care in the health system in the previous year.

The operational definition of case relies on the following ICD-10 codes: F00: Dementia in Alzheimer disease; F01: Vascular dementia; F02: Dementia in other diseases classified elsewhere; F03: Unspecified dementia; G30: Alzheimer disease



Utilisation: in the natural year after the contact where the patient was identified, any contact in the following instances will count as an episode/visit/contact:



Spending: The following spending categories will be used to calculate the costs associated to the episodes/visits/contacts to the system

Category	Definition
Inpatient and Emergency Care	
Total Inpatient/Acute Hospital Spending	Sum of spending categories below.
General Acute Care Hospital Spending	Acute hospitalisations are defined as any hospitalisation that occurred in general hospitals for any condition. This includes physician fees, inpatient laboratory, imaging and drugs given. All admissions were counted even if the patients were discharged the same day.
Psychiatric	Spending related to hospitalisations occurring in psychiatric hospitals.
ED Spending that did not result in hospitalization	All spending related to the emergency department (ED) that did not result in hospital admissions. We include ED spending as part of inpatient spending since some countries may not be able to exclude this type of spending from their hospital costs.
Post-Acute Rehabilitative Care	
Facility-based post-acute rehabilitative care	All spending related to inpatient rehabilitative care or skilled nursing facilities or other spending related to rehabilitative care that requires a facility-based stay. (Please specify the source of other facility-based stays and whether you are able to distinguish between inpatient care and inpatient rehab spending.)
Facility-based post-acute rehabilitative care for Index Hospitalisation	Same as above but for Index Hospitalisation only.
Home-based post-acute rehabilitative care	All spending related to other rehabilitative care, such as outpatient rehabilitative care and home-based physical therapy.
Primary Care	
Total Primary Care Spending	Sum of spending categories below.
Primary Care Services provided by a medical doctor (MD)	Costs for any service provided by general practitioners, general internists or the equivalent in a primary care/ambulatory setting.
Non-MD Primary Care Services	Primary care services provided by nurses, nurse practitioners or other non-MD equivalent (physician assistants, nurse practitioner). (Exclude phone calls.)
Outpatient/Ambulatory Specialty Care	
Total Outpatient Specialty Care	Sum of spending categories below.
Outpatient MD Specialty Services	All visits to specialists who are MDs. These include MDs such as cardiologists, gastroenterologists, surgeons, etc. Do not include radiologists or pathologists in this category.
Outpatient Diagnostic MD Specialist Visits	All visits to radiologists or pathologists if they are actual patient encounters.
Outpatient Other Specialist Visits	All visits to specialists who are non-MDs. Please include the type of non-MDs that you capture in this data.
Durable medical equipment	Any costs that are unable to be specifically classified as specialty or primary care. This includes durable medical equipment, ancillary testing, etc.

Other costs not captured above (structural costs, other operating costs)	Please include costs that are not captured in the categories above. Please outline in detail what these costs pertain to in your country.
Home Health	
Home Health Visiting Nursing Care	Any care delivered at the residence of the patients by visiting nurses.
Drugs	
Outpatient Drugs	Any costs attributable to drugs prescribed to the patient in the outpatient setting are included in this section. Drugs administered as part of a hospital stay are not included in this category. Of note, drugs administered in the inpatient setting are included in a different category outlined above.
Long-Term Care	
Long-term care facilities	All spending related to long-term care facilities.
Long-term care at home/community	All spending related to long-term services provided at home or the community.

Acknowledgement: to the ICCONIC collaborative that has inspired this case study (https://icconic.org).

2. Desirable Health Service Use indicator calculation

Authors: Jane Lyons and Ronan Lyons
Version 1.0, 27th March 2020

Purpose: Setting up a version 1.0 of the data requirements for the study desirable utilization in two countries, (Aragon) Spain and Wales. Unlike the previous case, the data schema is extensively explained and the extraction process is standardised using an SQL script developed by one node and distributed to the other node.

Study protocol and data schema

Introduction

This document details the study design and methodology for creating a desirable Health Service Utilisation indicator for Wales as part of the Information for Action (INFACT) European project. This study used anonymised and encrypted demographic and healthcare data held in the Secure Anonymised Information Linkage (SAIL) databank (www.saildatabank.com) at Swansea University, Wales, UK. SAIL is acknowledged as one of the world's leading trusted research environments (TREs) and contains many different health and non-health individually-linked pseudonymised datasets on the population of Wales (1).

Project Aims

To create a desirable health service use indicator. This indicator was conceived and discussed at the meeting in Zagreb as potentially being an inspirational case study that utilised multiple sources of data from several countries.

Briefly, the underlying issue is that health services are under pressure everywhere due to a number of factors, including population aging, rising expectations, increasing technological opportunities for treating more diseases, and the huge growth of multi-morbidity (2).

There are few health indicators that, as well as providing an overview, can also be used to evaluate cross-national, national and local policy initiatives and interventions. Modern health informatics increasingly needs to link observation with intervention and evaluation, supporting targeting and testing of new ways of delivering care and improving health.

The proposed indicator would measure aspects of health and resilience that are available through routine data in an increasing number of settings. A first principle is that almost nobody starting a new year would wish to require hospital treatment as an inpatient, go to an emergency department or require treatment for infection, pain or mental health. We focused on developing a parsimonious indicator that measured the proportion of the population, subdivided by age group, gender and possibly socio-economic status (SES), that are likely to be free from these issues during a calendar year.

The indicator requires separate data analyses to produce five profiles that could be linked to create a composite where individual linkage is possible.

Data needed

1. Population register data (denominator).
2. Prescribing/dispensing datasets to create three profiles.
    a. Those not prescribed antibiotics in a year.
    b. Those not prescribed analgesics in a year.
    c. Those not prescribed mental health drugs (including sleeping tablets) in a year.

3. Hospital inpatient data
    a. Those not admitted during a year (including day cases).

4. Emergency Department data
    a. Those with no attendance in a year.

SAIL Data Sources

Within the SAIL system, we developed this concept within the Wales Multi-Morbidity Cohort, the protocol of which has since been published (3).

The following datasets have been utilised for the creation of the Welsh e-cohort:

- SAILW0911V.JL_MM_WDS_V7 – Welsh MM cohort - denominator data for cohort from the Welsh Demographic Service Dataset (WDSD), a centrally maintained list of people with free access to NHS provided healthcare.
- SAIL0911V.PEDW_SPELL_20191213 – hospital inpatient and outpatient datasets. Patient Episode Database for Wales (PEDW).
- SAIL0911V.EDDS_EDDS_20191213 – Emergency Department Data Set (EDDS)
- SAIL0911V.WLGP_GP_EVENT_ALF_CLEANSED_20200127 – General Practice dataset

The INFACT_DESIRABLE_HEALTH_INDICATOR.sql (please see attached document) contains the data management code for creating and calculating the desirable health use indicator per year (2010 – 2017).

Study participants design

This concept was designed around population data from the 2017 calendar year. Cohort entry includes all Welsh residents from the Welsh multi-morbidity e-cohort (WMC), alive and living in Wales on 31st January 2017. Study participants had to have been registered with SAIL providing practice up to and including 31st January 2017. The latter covers 80% of the population, whereas the other datasets cover 100%.

Demographic Variable source

- Age has been calculated using the WDSD recorded Week of birth and at the mid-year point (YYYY-07-01) per year.
- Sex: WDSD recorded.
- Welsh Index of Multiple Deprivation version 2011: Area deprivation level at Welsh multi-morbidity e-cohort start date (2000-01-01) and based on Welsh Lower Super Output Area (LSOA) version 2001. There are approximately 1900 such LSOAs in Wales, each with an assigned Welsh Index of Multiple Deprivation score. Usually these are ranked and divided into five equal groups with comparisons across the fifths to study socio-economic inequalities in health.

Data Management

The Welsh multi-morbidity e-cohort WDSD is the core dataset for identifying appropriate Welsh residents and overall study participants. This dataset is left-joined with the emergency department dataset, hospital admission dataset and GP dataset to identify anyone who has gone to an emergency department, been admitted to hospital or been prescribed the above drugs per year per person.

Inclusion and exclusion criteria for baseline population

- In the SAIL system, personal identifiers are replaced by Anonymised Linkage Fields (ALFs), which are derived from multiple encryptions of unique National Health Service (NHS) numbers.
- Welsh multi-morbidity cohort (WMC) ALFs only.
- Cohort end date >= 2018-01-01.
- GP coverage end date >= 2018-01-01.
- Age at cohort start date <= 110.

Inclusion and exclusion criteria for hospital admissions

- Spell number is not null. Spells are periods within hospital inpatient stays.
- Provider unit code is not null. Each hospital site has a unique provider code.
- Admission date is between 2010-01-01 and 2017-12-31.
- Good ALF status code: 1,4,39. These codes are for high quality linkage codes derived from identity matching using deterministic and probabilistic identity linkage used in the SAIL system (1).

Inclusion and exclusion criteria for Emergency Department visits

- Administration arrival date between 2010-01-01 and 2017-12-31.
- Good ALF status code: 1,4,39.

Inclusion and exclusion criteria for GP events

- GP practice code is not null.
- Event date between 2010-01-01 and 2017-12-31.
- Good ALF status code: 1,4,39.
- Version 1: This is based on the use of the Read 2 General Practice codes used in the UK and New Zealand electronic primary care systems.
    - Read codes V2 for infections: e%
    - Read codes V2 for musculoskeletal pain relief drugs: j%, di%, dj%'
    - Read codes V2 for mental health drugs: d1%,d2%,d3%,d4%,d5%,d6%,d7%,d8%,d9%,da%
- Version 2: This is based on mapping Read codes to ATC codes that are used in many systems:
    - ATC codes for infections: J%,A01A B%,A02B D%,A07A,D01%,D06%,D07C%, D09A A%,D10A F%,G01%,P%,R02A B%,S01%,S02%,S03%
    - ATC codes for musculoskeletal pain relief drugs: N02%,M01%,M02%,
    - ATC codes for mental health drugs: N05A%,N05B%,N05C%,N06A%,N06C%,

Table 1: ALF status codes and descriptions

Alf status codes	Description
1	NHS Number passes check digit test
4	Surname, First Name, Postcode, Date of Birth and Gender Code match exactly to WDS
35	Fuzzy Matching probability >= 0.5 & < 0.9
39	Fuzzy Matching probability >= 0.9
99	No match or Fuzzy Matching probability < 0.5

Calculating desirable Health Service Use indicator

The desirable health service use indicator has been calculated per person per year and identifies individuals who have not been a) admitted to hospital, b) had an ED attendance or c) been prescribed any of the above defined drugs per calendar year. A score of 0 identifies individuals who fit all these desirable criteria. A score of 1-3 identifies individuals who meet the a, b or c criteria. For example, a score of 2 could refer to someone being prescribed one of the defined drugs and also having gone to the ED in a year. The maximum score is 3, which identifies anyone who has been to the ED, been admitted to hospital and been prescribed at least one drug for either an infection, mental health issue or pain management.

SQL script is published at https://zenodo.org/record/4880073#.YLSZSC0RoUs.

Appendix 2

Stepwise approach to achieve organisational interoperability in a federated infrastructure

In the implementation of a research query, we strongly recommend achieving semantic interoperability to follow these 8 activities (some of them recursive). In the client-server approach, the coordination node has to orchestrate all the activities across nodes, aiming to reduce human interaction as much as possible.

Step 1. Any member of the user community could broadcast a research question to the federated research infrastructure.

Step 2. Once a research question (RQ) has been sent out to the federation of nodes, the coordination hub will start a process to define the key components of the RQ, so as to establish a detailed operational definition (i.e., unit of analysis, period of study, target population, inclusion and exclusion criteria, etc.).

Once the research question is clear, the originator of the query will draft the data schema of the use case defining in detail. The data schema will require the following: the entity, the attribute of interest (variable) with a label and a normative description, the encoding system of the variables, the unit and values in which the information is stored, the validation rules, whether the variable is required or optional and, lastly, the data source from which the variable has been extracted. In addition, the data model will specify the mapping of the variables to different encoding systems.

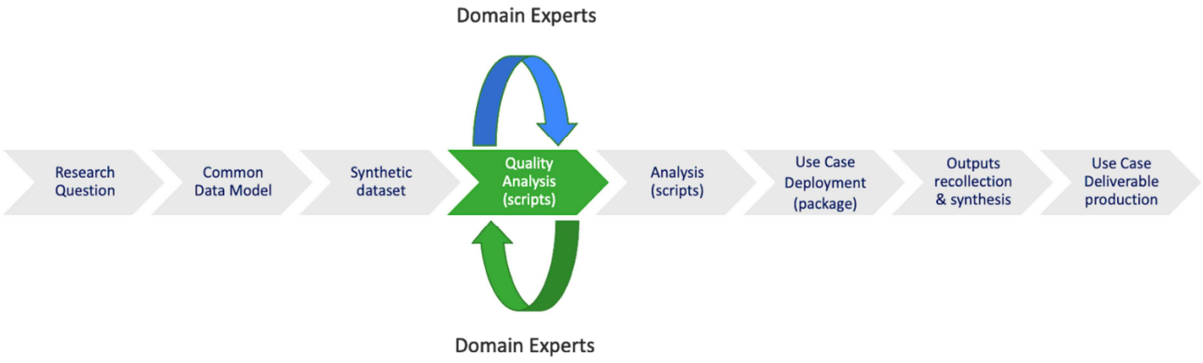
Data model entity	Variable						Data Quality Assessment
Associated entity in ERD	Label (var_label)	Name (var_concept)	Classification/Encoding	Units	Format	Description	References to validation rules
patient	**patient_id**	patient identificator	private key ciphering function	none	string	patient pseudonymized identificator	SHA256
patient	**age_nm**	age	none	years	integer	patient's age at the moment	3-digits; min 18; max 80
patient	**socecon_lvl_cd**	socioeconomic level	quintile	quintiles	integer	patient's socioeconomic level (quintile)	min 1; max 5
patient	**country_cd**	country (residence)	ISO3166	none	string	patient's country of residence	ISO3166-3
patient	**country_origin_cd**	country (origin)	ISO3166	none	string	patients' country of origin	ISO3166-3
procedure	**ttm_type_cd**	type of treatment	types of treatment referred below or a combination of them	none	integer	type of treatment received by the patient	values restricted to existing categories
procedure	**time_dx_to_surgery_nm**	[time til first surgery]	none	days	double	time from breast cancer diagnosis to first surgical procedure	no negative values allowed
procedure	**time_dx_to_radiotherapy_nm**	[time til first radiotherapy session]	none	days	double	time from breast cancer diagnosis to first radiotherapy session	no negative values allowed
procedure	**time_dx_to_chemotherapy_nm**	[time til first prescription/administration of a chemotherapy treatment]	none	days	double	time from breast cancer diagnosis to first prescription/administration of a chemotherapy	no negative values allowed

Data model description | Cohort (definition) | Surgery (procedure definition) | +

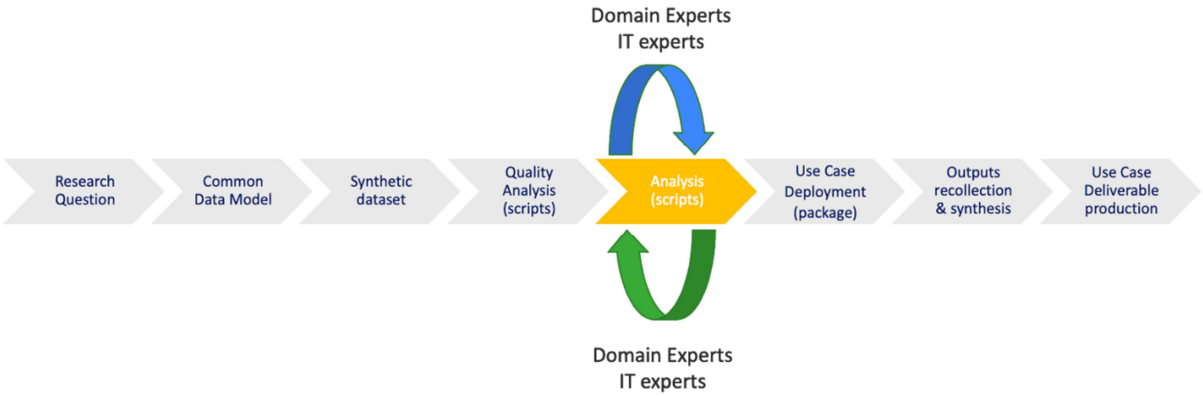
Step 3. As a continuation, a synthetic dataset based on these data requirements will be produced to document the data model and formalise the analytical pipeline (see the next steps) in a transparent, auditable way. The idea is to anticipate and programme all the critical elements in the pipelines and solve any potential problems before real data are used.

When it comes to the formalization of the data model there are multiple options; for example, the minimum metadata schema provided by the `dataspice` implementation (both in R and Python) complying with the Schema.org international standard (see the example below).

Note: Steps 2 and 3 are recursive and end once the data model is finally approved by the federation. The participation of domain experts and data scientists is essential at this point.



Step 4. Data quality assessment. Using the synthetic dataset and open source libraries, the coordination hub will prepare a quality analysis for those key variables, looking for incompleteness, missingness, outlier values, anomalous distribution, flaw associations, etc. There are open source libraries (e.g., '*dlookr*') that provide you with standard scripts for data quality evaluation.



Step 5. Configuration of a preliminary analytical script according to the objectives of the study. Iteration will be critical, which will help to update and improve the scripts. For that purpose, the exchange between domain experts, data scientists and IT is essential, particularly if the distribution implies model assembly.

Step 6. Once Steps 3 to 5 are finished, the coordination hub will package the whole pipeline, distributing the solution (for example, DOCKER) to the different nodes. They will then transform their data to the data schema, and will run the different parts: data quality assessment and analyses.

Step 7. In this step, the results of the analysis are collected; and if the research question requires meta-analysis, it will be sent back to the coordination hub, which will run the meta-analyses. Depending on

the type of meta-analysis, there could be the need for some recursive interaction between the coordination hub and the nodes.

Step 8. The entire pipeline is published in a repertoire as ZENODO. It is worth highlighting that all these pipelines have to be findable, accessible, interoperable and reusable in the OPENAIRE community (https://www.openaire.eu). At this very moment, the federation has to decide what level of access should be given to the pipeline (https://guidelines.openaire.eu/en/latest/literature/index_guidelines-lit_v3.html).

Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu |Twitter: @JA_InfAct