SYNCHROS

WORK-PACKAGE 2

Methodological issues in harmonising and integrating cohort data

D2.4

Inventory of infrastructures providing support for integration of cohort data

SYNergies for Cohorts in Health: integrating the Role of all Stakeholders

Grant Agreement No. 825884

Start Date: 01/01/2019

Duration: 36 months

DOCUMENT INFORMATION

Authors	**Maria Panagiotopoulou, Sergei Gorianin**
Contributors	Albert Sanchez-Niubo, Juan R González
Reviewer	Ellen Vorstenbosch
Responsible Partner	ECRIN / PSSJD
Dissemination Level	Public
Nature	Report
Keywords	Infrastructure, Data harmonisation and integration, Data Standards, GDPR
Due Date	31/12/2020
Actual Submission Date	23/12/2020
Version	1.0

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

ACRONYMS AND ABBREVIATIONS

ADaM : Analysis Data Model
ADaMIG : ADaM Implementation Guide
BRIDG : Biomedical Research Integrated Domain Group
CDER : Center for Drug Evaluation and Research
CDISC : Clinical Data Interchange Standards Consortium
CDM : Centralized Data Model
CRF : Case Report Form
CTR : Clinical Trials Regulation
CTU : Clinical Trial Unit
DAM : Domain Analysis Model
DMIM : Domain Message Information Model
DNA : Deoxyribonucleic acid
EDPB : European Data Protection Board
EEA : European Economic Area
EHR : Electronic Health Record
ETL : Extract, transform and load
FDA : US Food and Drug Administration
FDM : Federated Data Model
GA4GH : Global Alliance for Genomics and Health
GDPR : The General Data Protection Regulation 2016/679
HIMSS : Healthcare Information and Management Systems Society
HL7 : Health Level 7 International
HUPO-PSI : Human Proteome Organisation-Proteomics Standards Initiative
IHE : Integrating the Healthcare Enterprise
MedDRA : Medical Dictionary for Regulatory Activities
MoU : Memorandum of Understanding
MSI : Metabolomics Standards Initiative
OBO : Open Biological and Biomedical Ontologies
PMDA : Japan's Pharmaceutical and Medical Devices Agency
QRS : Questionnaires, Ratings and Scales
RCRIM : Regulated Clinical Research Information Model
RDA : Research Data Alliance
RNA : Ribonucleic acid
SEND : Standard for Exchange of Nonclinical Data
SENDIG : SEND Implementation Guide
SDTM : Study Data Tabulation Model
SDTMIG : SDTM Implementation Guide
WHO : The World Health Organisation
WHODD : WHO Drug Dictionary

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

TABLE OF CONTENTS

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

EXECUTIVE SUMMARY

The SYNCHROS repository, already presented in Deliverable 2.3 and available at https://repository.synchros.eu, was created to share key information on initiatives that harmonise and/or integrate cohort data. An important feature of all these initiatives is the type of infrastructure used to gather the information from the different cohorts and perform data analysis across them.

The aim of this deliverable is to provide an inventory of existing infrastructures from the revised SYNCHROS repository initiatives. For the purpose of this deliverable, only those initiatives where an infrastructure for data analysis was identified were considered. Both the data layout within the infrastructure and the software used for the data analysis are discussed in detail. In addition, potential challenges in data analysis within the infrastructures have been identified.

Therefore, the document contains the following information:

1. Infrastructure inventory of the SYNCHROS repository initiatives.

2. Types of data lay-out within the infrastructures.

3. Types of software mostly used in the different infrastructures.

4. Challenges in analysing data within infrastructures.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

1. INTRODUCTION

In the early 2000s there was much political discussion on European level concerning how to make scientific collaboration more effective. To this end, infrastructures that facilitate international collaboration are essential as researchers need to achieve statistical inference through comparing data extracted from different population cohorts, patient cohorts and clinical trials. The result of these discussions led to the European Strategy Forum on Research Infrastructures - ESFRI, a European cooperative body for infrastructure initiated by the European Commission (https://www.esfri.eu/).

In the same lines, Larsson conducted in 2017 a literature review that aimed to identify the most pressing issues for rendering biomedical research more efficient. Essentially, four different needs were highlighted by the scientific community[1]:

1. Cultural/procedural harmonisation: Emphasizes the need of securing harmonisation of "softer components" that is, employees and/or managers at the various organizations. It suggests the impediment is mainly attitudinal.

2. Data harmonisation: Emphasizes the need of updating the technical procedures and/or hardware to a uniform system that is used by all participating members.

3. **Infrastructure**: Emphasizes the need of an actual infrastructure, usually in the form of a physical infrastructure, but sometimes in more a conceptual sense.

4. Regulatory harmonisation: Emphasizes the need of securing harmonisation on higher, political level, usually via policy-making.

This report focuses mainly on the aspect "infrastructure" and provides an inventory of cohort initiatives with an identified infrastructure for the data analysis. This inventory constitutes a subset of the initiatives identified and included in the SYNCHROS repository and for each initiative we identify characteristics about the data layout and the types of software used for data management and analysis within every infrastructure.

With regard to the data layout, we can differentiate between two main types of infrastructures for the sharing of individual data from the cohorts within each initiative: 1) the individual data is centralised in one institution or server and 2) the individual cohort datasets reside in different institutions (federated), mostly on the server of origin. Both types of infrastructure are compared and their advantages and disadvantages are presented.

In addition, several challenges in analysing data across patient cohorts, clinical trials and population cohorts have been identified and discussed. The first challenge was the different data standards used across different types of studies. Data standards vary between population studies, and also differ from the standards used for health records, and for clinical trial data. Another challenge is the interoperability, which should be promoted in the future to avoid creation of non-interoperable data silos. However, in the short term, cohort integration requires harmonisation of content: core datasets should be defined and promoted, according to a consensus procedure similar to the one used to define core sets of outcome measures for clinical trials, or patient-

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

reported outcome measures. This ensures not only the relevance of these data for the patients, but also consistency of data collection across clinical studies. Finally, ethical challenges and data protection issues are presented. They are an important topic because national implementation of the GDPR has resulted in partly convergent data protection policies.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

2. INVENTORY OF INFRASTRUCTURES PROVIDING SUPPORT FOR INTEGRATION OF COHORT DATA

The inventory constitutes a subset of the initiatives identified and included in the SYNCHROS repository (https://repository.synchros.eu/). Only the initiatives where an infrastructure for the data analysis was identified are considered in Table 1.

It was a laborious task trying to obtain information on the infrastructure used by the initiatives and in some cases we could not obtain clear information regarding the use of some common computer environment, even if this was privately accessible among the researchers of the initiatives. We think that these initiatives tend to centralise the data in local supports and manage and analyse the data with common statistical software. Other initiatives were found to have some kind of infrastructure but kept it private and thus, no detailed information could be obtained. For all the rest, information was collected regarding the level of access and location of the data, on harmonisation, on whether they provide information on ethical and legal issues of the data, whether the data can be analysed within the infrastructure, and finally the type of software they use for data management and analysis.

Table 1 Inventory of infrastructures providing support for integration of cohort data

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
1	International HundredK+ Cohorts Consortium (IHCC)	Yes	Not yet	No	Yes	No	Not yet	No	Not yet	Information not found.	Federated	SAS
2	Canadian Partnership for Tomorrow's Health (CanPath)	Yes	By request	By request	By request	By request	Yes	Yes	Yes	Information not found, but implemented in the OBiBa software	Federated	OBiBa
3	Ageing Trajectories of Health: Longitudinal Opportunities and Synergies (ATHLOS)	Yes	By request	By request	By request	By request	Yes	Yes	No	Yes	Centralized	OBiBa (Opal/Mica); R/Rmarkdown

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
4	Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BioSHaRE-EU)	Yes	Some cohorts	No	Some cohorts	No	No	No	No	No	Federated	OBiBa (Opal/Mica); Datashield
5	Promoting mental well-being and healthy ageing in cities (MINDMAP)	Yes	Yes	No	Yes	No	Yes	Yes	No	Information not found, but implemented in the OBiBa software	Federated	OBiBa (Opal/Mica); R/Rmarkdown; Datashield
6	Interplay of Genes and Environment across Multiple Studies (IGEMS)	Yes	No	No	No	No	No	No	No		Centralized	NA
7	Swedish Cohort Consortium (Cohorts.se)	Not yet	NA	NA	Not yet	Not yet	Not yet	Not yet	NA	Not yet	Federated	OBIBA; DataShield
8	Cohort and Longitudinal Studies Enhancement Resources (CLOSER)	Yes	Yes	By request	Yes	By request	Yes	Yes	Yes	Not yet implemented: «CLOSER is currently working on projects across multiple research themes to produce new data resources including guides to cross-study data comparability and harmonised datasets.»	Centralized	Other
9	The Gateway to Global Aging Data (g2aging)	Yes	Yes	Most of them	Yes	Yes	Yes	Most of them	Yes	No	Centralized	STATA

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
10	Sino-Quebec Perinatal Initiative in Research and Information Technology (SPIRIT)	Yes	Yes	No	Yes	No	Not yet	Not yet	No	Information not found.	Federated	OBiBa (Opal/Mica); Datashield
11	National E-Infrastructure for Aging Research (NEAR)	Yes	Yes	By request	By request	By request	Yes	Yes	Yes	Information not found.	NA	OBiBa
12	International Childhood Cardiovascular Cohort Consortium (I3C)	Yes	No	No	No	No	No	No	Yes	Information not found.	Centralized	NA
13	Network on the Coordination and Harmonisation of European Occupational Cohorts project (OMEGA-NET)	Under development. They are constructing an inventory of European occupational, industrial, and population cohorts, including registry-based cohorts.									Federated	DataShield
14	euCanSHare	Yes	Yes	By request	Most of them	Most of them	Yes	Yes	Yes	Information not found.	NA	OBiBa
15	Healthy Life Trajectories Initiative (HeLTI)	Under development. HeLTI will use the infrastructure from ReACH. They state that «ReACH will act as a platform for the HeLTI initiative, allowing researchers to build upon it and utilize it as a template for harmonisation.»									NA	OBiBa
16	The Asia Cohort Consortium (ACC)	They use a private infrastructure.									Centralized	NA
17	BBMRI-NL-Biobank (BBMRI-NL)	Yes	Partially	No	Partially	No	No	No	No	No	Federated	Molgenis
18	Consortium on Health and Ageing: Network of cohorts in Europe and the United States (CHANCES)	Yes	Partially	No	Partially	No	Yes	No	No	No	Centralized	NA
19	Finnish Genome Project (FinnGen)	Yes	No	No	No	No	Yes	Yes	No	No	Federated	Github, R, Other

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
20	South African Population Research Infrastructure Network (SAPRIN)	Yes	Yes	By request	Yes	No	Yes	No	Yes	No	Centralized	NA
21	Maelstrom Research	Yes	Yes	No	Yes	No	Yes	No	No	No		
22	LifeCycle EU Child Cohort Network (LifeCycle)	Yes	No	No	Yes, but only harmonised metadata	No	Yes	No	Yes	Yes	Federated	Molgenis; Datashield
23	EUCAN CONNECT	Not yet	Not yet	NA	Not yet	NA	Not yet	NA	NA	Not yet	Federated	OBiBa; Molgenis; DataSHIELD
24	Research on European Children and adults born preterm (RECAP)	Yes	Yes	By request	Yes	Yes	Yes	By request	No	No	NA	OBiBa (Opal/Mica)
25	The Human Early-Life Exposome (HELIX)	Yes	Partially	No	Partially	No	Partially	No	No	No	Centralized	Regular statistical software
26	Child Cohort Research Strategy for Europe (CHICOS)	Yes	Yes	No	Yes	No	No	No	Yes	No	NA	NA
27	Research Advancement through Cohort Cataloguing and Harmonization (ReACH)	Yes	Yes	No	Yes	No	No	No	No	Information not found.	Federated	OBiBa; DataShield
28	interconnect	Yes	NA	NA	NA	NA	NA	NA	NA	Yes	NA	OBiBa (Opal/Mica); Datashield
29	National Cancer Institute Cohort Consortium (NCI)	Yes	By request	By request	By request	By request	Yes	No	Not found	Yes	Centralized	Other

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
30	Exposome Project for Health and Occupational Research (EPHOR)	Under development.									NA	NA
31	Dynamic longitudinal exposome trajectories in cardiovascular and metabolic non-communicable diseasesn (LONGITOOLS)	Under development.									NA	NA
32	International Childhood Cancer Cohort Consortium (I4C)	Private access									Centralized	NA
33	Biomarkers in Atopic Dermatitis and Psoriasis (BIOMAP)	Yes	Yes	Yes	Yes	Yes	NA	NA	Yes	Yes	Some centralized, some federated	RedCap; tranSmart; others
34	Common Infrastructure for National Cohorts in Europe, Canada and Africa (CINECA)	Yes	Yes	NA	Yes	NA	NA	NA	Yes	Yes	Federated	NA
35	The Canadian Network for Observational Drug Effect Studies (CNODES)	Private access									Federated	NA
36	Collaboration of Observational HIV Epidemiological Research Europe (COHERE)	Information not available or not found							Yes	No	Centralized	Other
37	Cohort Studies of Memory in an International Consortium (COSMIC)	Yes	No	No	No	No	No	No	No	No	Centralized	NA
38	The Breast Cancer Association Consortium (BCAC)	Private access									Federated	NA
39	35th Multicenter Airway Research Collaboration (MARC-35)	Private access									NA	R/Rmarkdown

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
40	National Cancer Institute Cohort Consortium (NCI)	Private access									Centralized	Other
41	European Sudden Cardiac Arrest network: towards Prevention, Education and NEw Treatment (ESCAPE-NET)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Centralized	NA
42	Malignant Germ Cell International Consortium (MaGIC)	Yes	No	No	Yes	Partially	Yes	Private	No	Yes	Centralized	R; Other
43	Global ECT-MRI Research Collaboration (GEMRIC)	To determine by contacting the initiative.									Centralized	NA
44	Reconciliation of Cohort data in Infectious Diseases (ReCoDID)	Currently being developed.									Federated	NA
45	Genomics Evidence Neoplasia Information Exchange (GENIE)	Registration is needed to obtain this information.									Centralized	NA
46	RESPOND: International Cohort Consortium of Infectious Disease	Information not available or not found.									Centralized	NA
47	European Health Data and Evidence Network (EHDEN)	Currently being developed.									Federated	NA
48	HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) towards improved stratification, treatment and health policy making disease (HarmonicSS)	Currently being developed.									Centralized	NA

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

		Level of access to data			Level of data discovery		Are harmonised data accessible?					
No	Initiative	Cohort general information	Cohort metadata	Cohort individual data	By only metadata	Individual data statistics?	Only metadata	Also individual data	Is information on ethical and legal issues available within the infrastructure?	Can Integrated data analysis be performed within the infrastructure?	Infrastructure	Software
49	HARMONY: European Public-Private Partnership for Big Data in Hematology	Information not available or not found										NA
50	Medical Informatics in Research and Care in University Medicine (MIRACUM)	Information not available or not found									Centralized	NA
51	Sentinel Initiative	Information not available or not found									Centralized	NA
52	Sildenafil TheRapy in dismal prognosis early onset fetal growth restriction (STRIDER)	Yes	Partially	No	No	No	No	No	NA	No	Centralized	STATA
53	ClinicalTrials.gov (NIH)	Yes	Partially	No	Partially	Some results	No	No	Yes	No		
54	DukeHealth	Yes	No	No	No	No	No	No	No	No		

Population cohort initiatives

Patient cohort initiatives

Clinical trial initiatives

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

3. DATA LAY-OUT WITHIN THE INFRASTRUCTURE INVENTORY

Out of the 54 initiatives with an infrastructure identified within this inventory, 17 (31%) use federated data analysis, 23 (43%) centralized and 1 centralized in some cases and federated in others (2%). For 13 initiatives (24%) this information was not accessible publicly. Such information could potentially be obtained after registration or contacting the initiative or if collaboration is established.

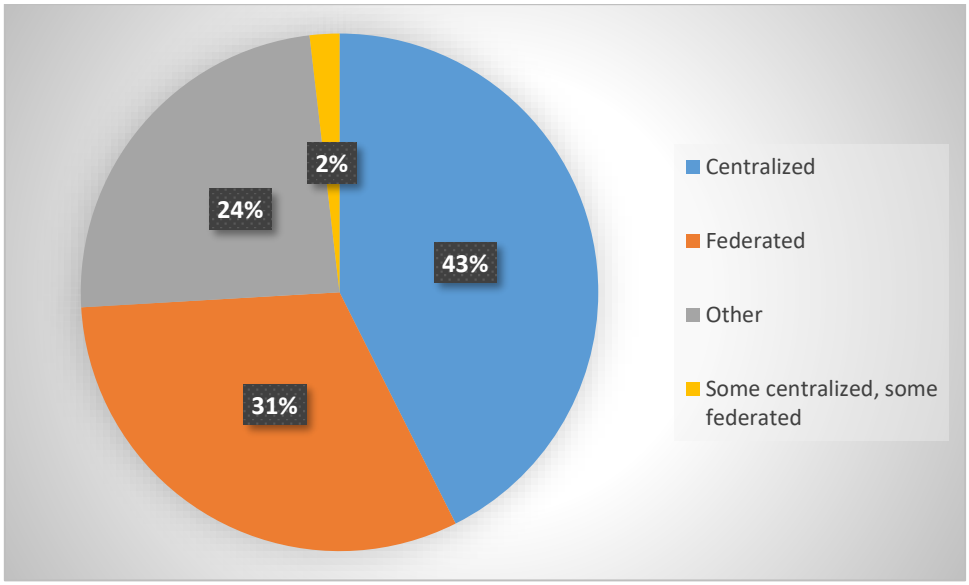

Figure 1 Data layout within the infrastructures of the inventory

In general, the most common type of data analysis is the centralized analysis according to which the data are gathered in a single common database, i.e., a centralized database. With regards to cohort studies and clinical trials, this type of analysis presents several data sharing issues, especially in the case of prospective studies where the patient data are updated. In addition, centralized databases are prone to data breaches.

On the other hand, federated analysis does not involve any patient data breaches as the data never get out of the clinical/research centre (e.g., hospitals). According to this concept, the initial data model is distributed from an authorized reference centre (i.e., an external database) to each clinical/research centre's local database for training and testing purposes (e.g., assume a Bayesian network model that is distributed and executed separately on each clinical/research centre). In fact, the data model is executed on each clinical/research centre (i.e., in a parallel way), and the individual results are returned to the reference centre where they are finally combined and distributed to all involved clinical/research centers[2].

The main features of the two data models are presented below.

3.1. The Centralized Data Model

In a Centralized Data Model (CDM), the data provider consolidates the infrastructure data in one repository[3].

Using a centralized data system may resolve data duplications, inconsistent master data, and improve data quality. However, implementing a centralized data system may require users to overcome challenges such as geographical locations of the applications, cost of the implementation, and compliance with different national (or even regional) rules and regulations [3,4].

In a centralized data system, all participating source systems copy their data to a single, centrally-located data repository where they are organized, integrated, and stored using a common data standard (for instance, clinical CDM)[7]. As depicted in Figure 2, data in a centralized data system are periodically matched, integrated, and loaded into a central repository. Users query the system and can access the data to which they have been authorized to view and use.

The most interoperable data architecture, the CDM is also the most expensive to establish and maintain because it requires a large upfront investment in technology in the form of servers, which need to be monitored and stored in a secure, separate location.

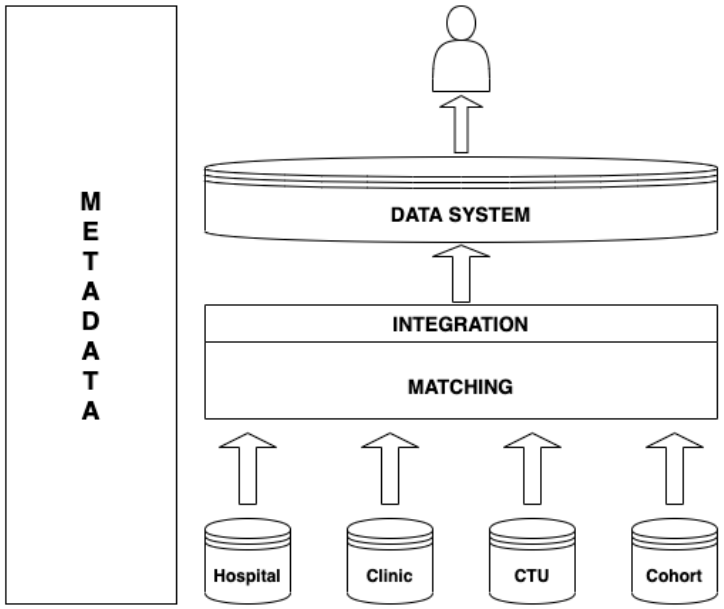

Figure 2 Basic structure of a centralized data system

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

3.2. The Federated Data Model

The Federated Data Model (FDM) allows an organization to extend data and infrastructure services to inquire data from multiple sources[3,7].

The goal of a federated data model is to make infrastructure data available to all departments and partners of a network, initiative or organization. Yet, implementing a federated data model comes with many challenges such as synchronization of data between transactional and master data, network connectivity between the sources and master data management hub, performance, maintenance, and identifying roles and responsibilities.

In a federated data system, individual source systems maintain control over their own data, but agree to share some or all of this information to other participating systems upon request. System users submit queries via a shared intermediary interface that then searches the independent source systems. In a federated system, as depicted in Figure 3, data are queried from source systems and records are matched to fulfil a data requestor's information needs. The linked data are not stored by the system, but rather, are removed once cached and delivered. The individual sources of data maintain control of their data, storing and securing them, and providing them to the system only upon request.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

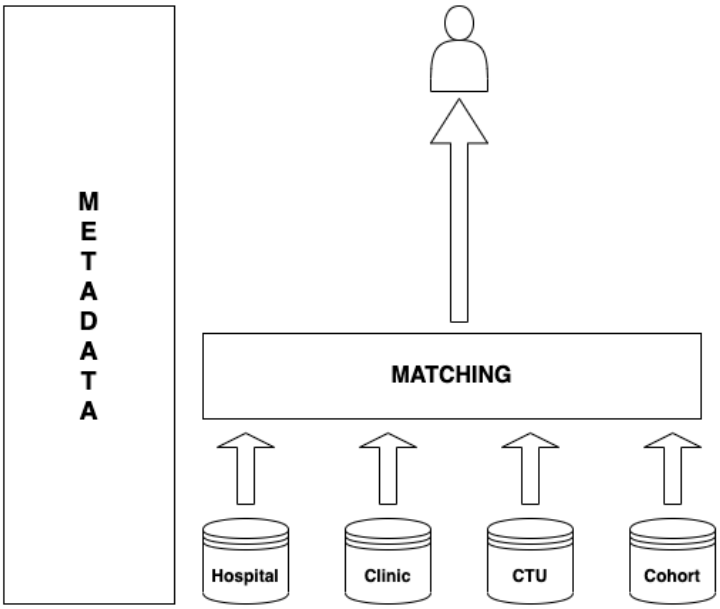

Figure 3 Basic structure of a federated data system

3.3. Comparison

Determining which architectural model is suitable for an infrastructure depends on several factors; including use of the infrastructure data, number of applications (domains) that will use the master data, development and availability costs, delivery schedule, performance, efficiency, limitations, risk, training, operations, compliances, deployment, security, accessibility, dependability, data quality, stability, maintainability, reliability, availability, flexibility, scalability, and predictability[6].

Table 2 Comparison between the centralized and the federated models

Criteria	Centralized	Federated
Data ownership	Data ownership is with the source agency with shared data stewardship with the centralized data warehouse agency/entity. Responsibility for this data stewardship should be spelled out	Data ownership is with the source agency with no need for shared data stewardship.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Criteria	Centralized	Federated
	in Memoranda of Understanding (MoU).	
Staff resources	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will need to be given to the extract, transform and load (ETL) processes to reflect changes in source data systems and data element modifications. Staff will also be needed to support the centralized database system.	Staff resources are required of each source system to oversee and maintain required data access. In addition, support will need to be given to the extract, transform and load (ETL) processes to reflect changes in source data systems and data element modifications. Staff resources are required from each participating agency to review and approve data requests.
Technical requirements	Each source system will need to be willing to allow access or provide the data to be included in the centralized data system. An infrastructure to support the centralized system along with ETL tools, conduct matching processes and storing the results. There will also be a need to deliver the matched resulting dataset (e.g., via portal or business intelligence (BI) solution).	Each source system will need the required hardware and network bandwidth to facilitate and process external queries (ETL tools), conduct matching processes and return the resulting dataset. There will also be a need to deliver the matched resulting dataset, i.e. portal or business intelligence (BI) solution.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Criteria	Centralized	Federated
System Performance	Data extraction is generally fast since all data matches have occurred in the transformation and load steps. Match once, use many times. Scheduled extracts can occur on source systems during off- peak hours to minimize impact on sources. Centralized data system architecture can be designed specifically for this purpose, thus increasing response times. Established technology and procedures; proven technology.	Subject to longer delays in data delivery due to load on source systems, etc. Agency specific performance issues can affect the performance of the entire system. Also the possibility of limited or narrow windows of processing time due to other/competing priorities. Relatively new technology; accounts for less than 10 percent of all data warehouse projects; not a proven technology.
Privacy/Security	Primary responsibility is with the centralized data system agency/entity as the data steward, but is dictated by source system agencies via memoranda of understanding. Security is handled through access rules for users. May make it easier to account for data integrity. Stakes may be higher in the event of a breach since all data are stored in one location (though typically records are deidentified as part of the load process).	Primary responsibility is with the source system agencies. Secure process needed for handling of data queries. Data are diffused, allowing for tailored protection based on sensitivity of each source system's data, and reducing the amount of data that could be accessed through a breach.
Data updates/corrections	Establish processes for ETL either when data are changed (if required to have near real- time data in centralized data system) or at a specific periodicity to capture changes, corrections, or updates.	Data reside within each agency. Each agency is responsible for communicating and possibly updating the data extract processes to reflect changes, corrections or updates.
Data availability	Based on when data are available in the source and made available for extract. Access to data is determined by source agency via MoU.	Based on when data are available in the source and made available for extract. Access to data is determined by the source agency.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Criteria	Centralized	Federated
Data quality	Process for data cleansing apply to all data as agreed upon by the source system agencies; consistency of data cleansing processes and data quality checks. May provide more reliable data since the compiled data from various systems are validated as part of the load process.	Dependent on processes implemented at each agency.
Implementation	Longer implementation period due to the need to build the centralized data system database/warehouse. But equal time is also needed to determine requirements and processes for ETL and data provision.	Generally requires less time; although equal time is needed to determine requirements and processes for ETL and data provision.
Scalability	Potentially supplementing or expanding centralized data system architecture to accommodate additional agency source system data. Writing ETL processes and matching/integration rules.	The addition of any required hardware and other resources (as mentioned above) required for data queries/matches across the system. Writing ETL processes and matching/integration rules.
Production of standard reports	Can be an automated process; less expensive and timelier to accomplish.	Dependent on an agency accepting this as a responsibility.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Criteria	Centralized	Federated
Sustainability	Possible approaches include a state appropriation to the centralized data system agency/entity for the development and ongoing support and maintenance of the centralized system. This would have no fiscal impact on the participating agencies. Another approach would be for each participating agency to pay for a proportional part of the needed funds for the support of the centralized system, in a cost recovery model. This could be a deterrent for agencies to participate.	Possible approaches are for each participating agency to make their contribution for the corporate support of the processes needed for the federated system. This may be a deterrent for agencies to participate. Another approach would be specific appropriation that is allocated to each participating agency, based on a funding formula.
Usability	Longitudinal data all in one place. Facilitative of data mining.	Multiple years of data must be queried from partner agencies, which requires assurance of comparability. If additional years of data are needed for a given cohort, the entire data set will need to be rebuilt.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

3.4. Advantages and Disadvantages

Table 3 Advantages and disadvantages of the centralized and federated models

	Centralized	Federated
Advantages	+ Proven technology + Better performance + Better for data mining + Easier to account for data integrity/security + Central data policy + Easier to ensure data quality + Quicker data results	+ Shorter development time + Mitigates turf battles/get around trust issues + Diffuses data and allows for tailored protection of data based on sensitivity + More easily scalable
Disadvantages	- Higher costs for infrastructure development and training - Data only as current as most recent load - Higher risk in event of breach due to amount of data contained in single repository	- Requires development and maintenance of multiple data sharing policies - Data linked every time a dataset is generated - Unproven technology (for example, response time not yet tested) - Investment and support of intermediary interface by each of the participating agencies - Limited data integration

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

4. OVERVIEW OF SOFTWARE USED WITHIN THE INFRASTRUCTURES OF THE INVENTORY

Obtaining information regarding the software used in each of the initiatives listed in this inventory for harmonisation and integration of data was all but trivial. 39% of the initiatives did not publicly include information on the software they use. 20% use OBiBa (Opal/Mica), 14% DataSHIELD, 6% R/R markdown and 5% Molgenis. 16% used "Other" tools for analysing their data such as REDCap and Excel. 1 initiative listed in this inventory reported using SAS and 2 Stata.

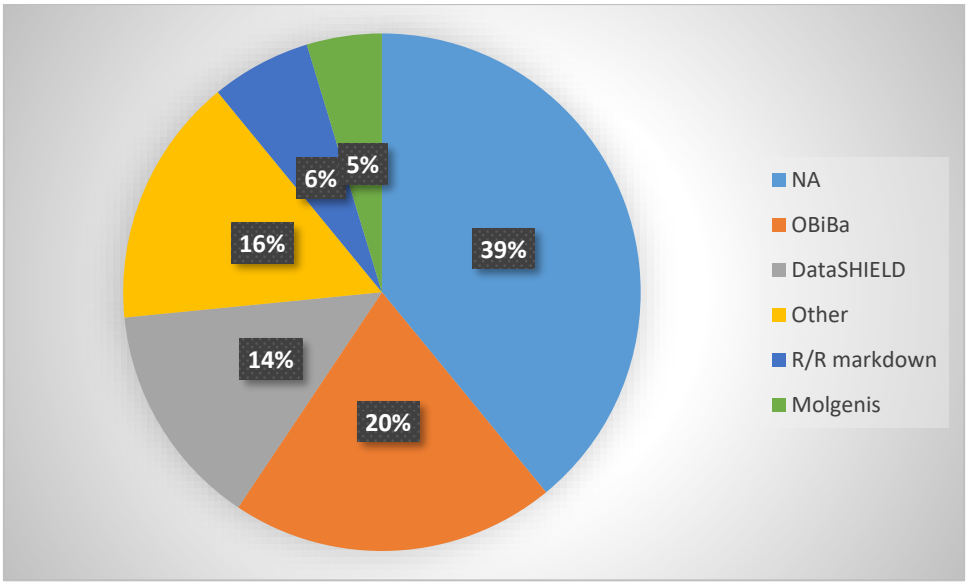

Figure 4 Software used in the infrastructures of the inventory

The following is a description of the main softwares collected in the inventory.

4.1. OBiBa (Opal/Mica)

OBiBa (https://www.obiba.org/) is an international project committed to build open source software for epidemiological studies. As part of the Maelstrom Research program, OBiBa software suite is developed in close partnership with large-scale studies and supports the entire data management lifecycle including data collection, integration, harmonisation, sharing, and analysis.

Opal is OBiBa's core data warehouse. This application provides all the necessary tools to import, transform and describe data. Subject's identifiers can also be managed at data import and export time.

Opal is integrated with R and complex statistical analysis and reports can be performed. The implementation of the DataSHIELD process allows advanced statistical data analysis across multiple studies without sharing and disclosing any individual-level data. Being integrated with

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Onyx and Mica, studies using Opal can securely import data collected with Onyx. They can also create web data portals with Mica that query Opal databases to obtain real-time aggregated reports on subject's data. Secured REST web services are also available allowing to automate server management (Python command line tools) or to access to data (from R or any tools that are web-capable). Features:

- Store data on an unlimited number of variables,
- Support MongoDB , Mysql , MariaDB and PostgreSQL as database software backend,
- Customized variable dictionaries,
- Import data from CSV, SPSS, SAS, Stata files and from SQL databases,
- Export data to CSV, SPSS, SAS, Stata files and to SQL databases,
- Incremental data importation,
- Connect directly to multiple data source software such as SQL databases and LimeSurvey,
- Store data about any type of "entity", such as subject, sample, geographic area, etc.,
- Store data of any type (e.g., texts, numbers, geo-localisation, images, videos, etc.),
- Import and store genotype data as VCF files (Variant Call format),
- Advanced indexing functionality using ElasticSearch.

Mica is a software application used to create data web portals for large-scale epidemiological studies or multiple-study consortia. It helps studies to provide scientifically robust data visibility and web presence without significant information technology effort. It also provides a structured description of consortia, studies, annotated and searchable data dictionaries, and data access request management.

4.2. DataSHIELD

DataSHIELD (https://www.datashield.ac.uk/) is an infrastructure and series of R packages that enables the remote and non-disclosive analysis of sensitive research data. Users are not required to have prior knowledge of R.

DataSHIELD provides a technological solution that can circumvent some of the most basic challenges in facilitating the access of researchers and other health care professionals to individual level data.

DataSHIELD facilitates important research in settings where:

- a co-analysis of individual-level data from several studies is scientifically necessary but governance restrictions prevent the release or sharing of some of the required data, and/or render data access unacceptably slow
- equivalent governance concerns prevent or hinder access to a single data set
- a research group wishes to actively share the information held in its data with others but does not wish to cede control of the governance of those data and/or the intellectual property they represent by physically handing over the data themselves

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

- a data set which is to be remotely analysed – or included in a multi-study co-analysis – contains data objects (e.g. images) which are too large to be physically transferred to the site of analysis.

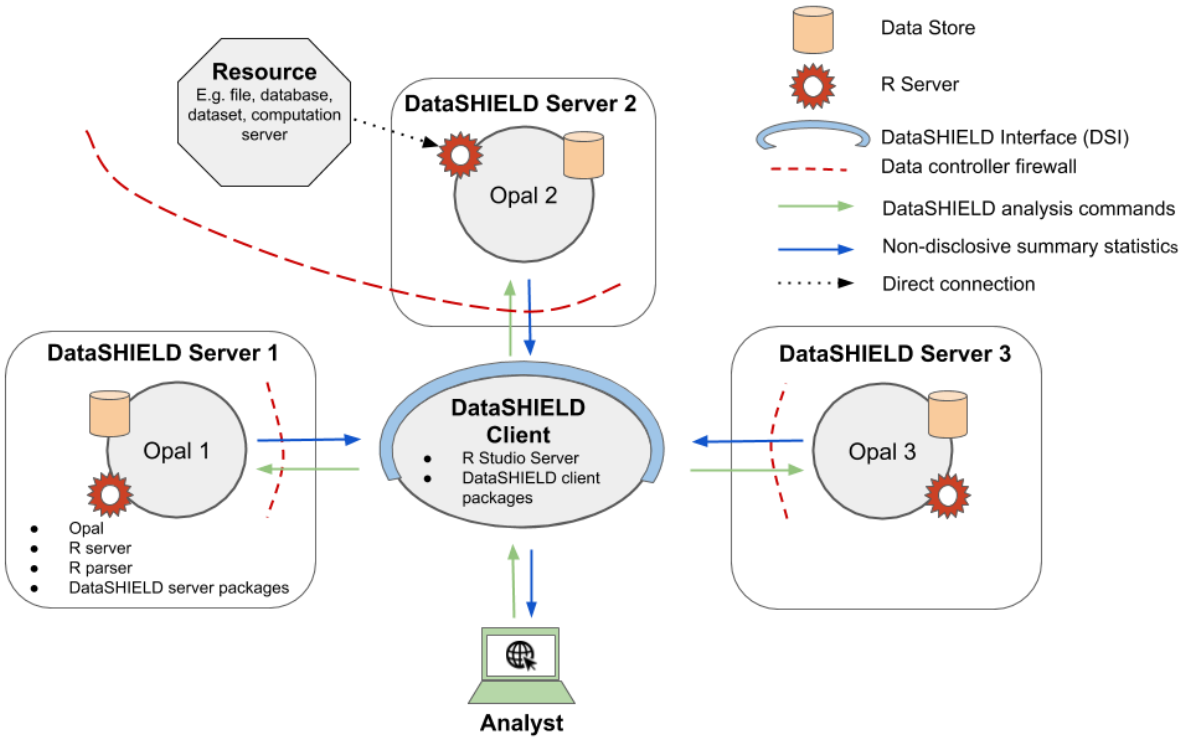

Figure 5 DataSHIELD Deployment Architecture
[from https://www.datashield.ac.uk/about/datashielddetailedoverview/]

4.3. Molgenis

Molgenis (https://www.molgenis.org/) is a data platform facilitating scientific collaborations for researchers and bioinformaticians.

Features:

- Structured data management - Uploaded user's data can be refined by using Molgenis advanced 'object-relational' data definition format and the online metadata editor.
- FAIR data sharing
- Secure access - integrated AAI system, which supports connection by using institutes' accounts (for example: SURFconext, BBMRI, ELIXIR) or Google two-factor authentication.
- Scripting and visualization -supporting of external JavaScript, R, HTML and API integration.
- Harmonisation and integration
- Task automation
- Questionnaires - to get data directly from the source.
- Customization

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

- App development platform - creation of independent applications and plugins within the Molgenis environment.
- High performance computing

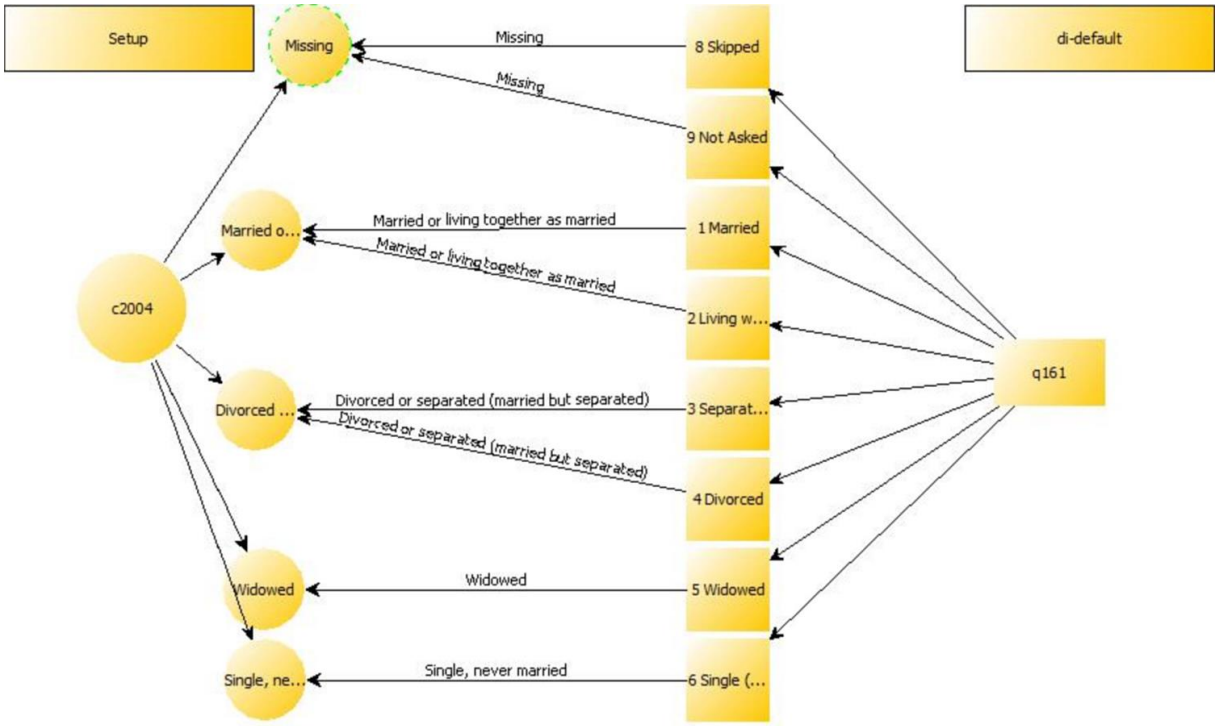

Figure 6 Marital Status harmonisation mapping

4.4. R/Rmarkdown

R (https://www.r-project.org/) is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

Besides, Markdown is a lightweight markup language that you can use to add formatting elements to plaintext text documents. R together with Markdown provides a very useful combination to generate high quality reports about data integrative analysis. They can be designed to produce automatically results and can be fully reproducible and shared among researchers.

5. CHALLENGES IN ANALYSING DATA WITHIN INFRASTRUCTURES

5.1. Data standards issues

One of the most important factors for the biomedical research to thrive is to standardise the data. Standards can be defined as an agreed compliant term or structure to represent a biological entity. Entities are all types of units of biological information. For example, we use T, G, A, C as a standard way to refer to the nucleotides that make the DNA. Lots of standard initiatives exist nowadays, sometimes redundant, often non driven by the end users' communities. The following table lists important standard initiatives[8]:

Table 4 Important international initiatives on data standards in life sciences

Name	Acronym	Aim	Website	Reference
The Open Biological and Biomedical Ontologies	OBO	Establish a set of principles for ontology development to create a suite of orthogonal interoperable reference ontologies in the biomedical domain	http://www.obofoundry.org/	DOI: 10.1038/nbt1346

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Clinical Data Interchange Standards Consortium	CDISC	Establish standards to support the acquisition, exchange, submission and archive of clinical research data and metadata	http://www.cdisc.org	DOI: 10.4103/2229-3485.111779
Health Level 7 International	HL7	ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.	http://www.hl7.org/index.cfm	
Human Proteome Organisation-Proteomics Standards Initiative	HUPO-PSI	Defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification	http://www.psidev.info	DOI: 10.1089/omi.2006.10.145
Global Alliance for Genomics and Health	GA4GH		https://www.ga4gh.org/	DOI: 10.1089/gtmb.2014.1555
Computational Modeling in Biology	COMBINE	Coordinate the development of the various community standards and formats for	http://co.mbine.org/	DOI: 10.3389/fbioe.2015.00019

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

		computational models		
Metabolomics Standards Initiative	MSI	Define community-agreed reporting standards, which provided a clear description of the biological system studied and all components of metabolomics studies	http://msi-workgroups.sourceforge.net	DOI: 10.1038/nbt0807-846b
Research Data Alliance	RDA	Builds the social and technical bridges that enable open sharing of data across multiple scientific disciplines	https://rd-alliance.org/	
Standards organization for open and FAIR neuroscience	INFC	Develop, evaluate, and endorse standards and best practices that embrace the principles of Open, FAIR, and Citable neuroscience	https://incf.org/	

It is out of the scope of this deliverable to review all the standards in life sciences. We will rather focus on highlighting some issues with the use of the current standards in interventional and observational studies.

5.1.1. CDISC standards in interventional studies

In 1997, CDISC was founded on the growing industry recognition of the need for standards. Its mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC "provides clarity by developing and advancing data standards of the highest quality to transform incompatible formats, inconsistent methodologies, and diverse perspectives into a powerful framework for generating clinical research data that is as accessible as it is illuminating".

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

CDISC has made huge progress in creating shareable, end-to-end data standards for clinical and non-clinical research. To date, the foundational standards focus on core principles of data standard definitions that include models, domains and specifications for data representation. Standards around protocols and data collection exist but are currently underused, which indicates that further work is necessary for them to be widely adapted by the scientific community.

Mature standards include[9]:

- The Study Data Tabulation Model (SDTM); the tabulated representation of the collected data and one of the first standards developed, has evolved to not only support clinical data but also non-clinical, medical devices, and pharmaco-genomics data. SDTM Implementation Guide (SDTMIG) guides the organization, structure, and format of standard clinical trial tabulation datasets.
- The Analysis Data Model (ADaM) and the ADaM Implementation Guide (ADaMIG), provide a standard structure of analysis data which is derived from SDTM data to support the creation of tables, listing, and figures as part of the study's clinical study report.
- Questionnaires, Ratings and Scales (QRS) supplements - Each QRS instrument is a series of questions, tasks or assessments used in clinical research to provide a qualitative or quantitative assessment of a clinical concept or task-based observation. The QRS team develops Controlled Terminology and SDTM (tabulation) supplements; the ADQRS Team develops ADaM (analysis) supplements.

Both US Food and Drug Administration (FDA) and Japan's Pharmaceutical and Medical Devices Agency (PMDA) require CDISC standards for the applications they receive.

The Center for Drug Evaluation and Research (CDER) has reported several data standards issues as it kept receiving numerous "SDTM-like" applications over the past several years in which sponsors failed to follow the SDTM Implementation Guide. In addition, some sponsors have wrongly believed that the submission of SDTM datasets obviates the need for the submission of analysis datasets, resulting in the delay in review due to the need to request these datasets.

CDER issued guidelines[10] stating that sponsors should refer to http://www.cdisc.org for the latest version of the implementation guides for SDTM, SEND, and ADaM (SDTMIG, SENDIG and ADaMIG), in addition to other documentation related to the study data standards. Other reported issues and the respective guidelines of CDER include:

- No Standard Terminology Exists

For variables for which no standard terminology exists, or if the available terminology is insufficient and needs to be extended, the sponsor may propose their own terminology.

- Consistent Drug Dictionary

It is strongly preferred that a consistent drug dictionary (for example, the WHO Drug Dictionary) terminology be used for coding of the concomitant medications. The generic preferred term for a drug should be used for the SDTM standardized medication name variable, CMDECOD.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

- MedDRA/Pathology Findings

When using MedDRA for adverse events and medical history terms, sponsors should exactly follow the spelling and case of the MedDRA terms. A common error that has been seen is a misspelling of a System Organ Class term or other MedDRA term. Sometimes trials are conducted at different times during the development cycle which results in the use of different versions of MedDRA from one study to the next. It is expected that the Adverse Event dataset for the Integrated Summary of Safety include MedDRA Preferred Terms from a single version of MedDRA. The reason for this request is that reviewers often want to analyze adverse events across trials, including the use of Standardised MedDRA Queries. If different dictionary versions are used for data included in the same analysis, there is the potential for confusion or incorrect results.

- Common Dictionaries

It is expected that common dictionaries are used across trials and throughout the submission for each of the following: adverse events, concomitant medications, procedures, indications, study drug names, and medical history. Implementation of such dictionaries should be careful to exactly follow the terminology conventions (e.g., spelling and case) specified by the dictionary or according to a single consistent sponsor specification if no pre-existing terminology exists. CDER reported receiving frequently data in which terminology conventions were not followed, for example, misspelling of MedDRA or WHO Drug terms or lack of conformance to upper/lower case or the use of hyphens.

5.1.2. CDISC standards in observational studies

When one tries to apply CDISC standards in observational studies there is a series of aspects to take into consideration. Observational studies differ from randomized controlled trials in significant ways regarding study goals and design, subject populations, clinical settings, regulatory/study oversight requirements, and data collection/data management practices. Many of these differences present challenges that are seen as barriers to the adoption of CDISC standards in observational research.

CDISC's "PhUSE Data Standards for Non-Interventional Studies" working group presented issues users face when attempting to implement CDISC standards in observational research[11].

A commonly-identified challenge related specifically to using CDISC standards in observational studies is the inability to meet SDTM conformance rules and subsequently failing validation checks. Conformance rules can apply at the dataset level, the variable level and at the controlled terminology level. Validation checks cover these rules and additional rules applicable to regulatory submissions. Table 5 below summarizes FDA validation rules around inclusion of datasets, and how these rules may present problems in observational research[11].

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Conformance rule	Flag Type	Challenge presented
Demographics (DM) dataset must be included in every submission.	Error	Inclusion of the dataset should not present a problem. However, some required/expected variables will not be available (See table 6 below)
Adverse Effects (AE) dataset should be included in every submission.	Warning	Depending on study type, these data may not be available
Lab Test Results (LB) dataset should be included in every submission.	Warning	Depending on study type, these data may not be available
Vital Signs (VS) dataset should be included in every submission.	Warning	Depending on study type, these data may not be available
Exposure dataset (EX) should be included in every submission.	Warning	Observational studies are not interventional studies. As such, exposure data will not be relevant.
Disposition dataset (DS) should be included in every submission.	Warning	Subjects won't likely meet formal milestones, nor will they have formal study completion/withdrawal dates.
Subject elements (SE) dataset should be included in every submission.	Warning	Trial arms and elements are not relevant to observational research. Therefore neither is subjects' progression through these.
Trial Arms (TA) should be included in every submission.	Warning	Observational studies do not have rigid study designs with planned arms.
Trial Elements dataset (TE) should be included in every submission.	Warning	Without trials arms there are no elements to describe.
Trial Summary (TS) dataset must be included in every submission.	Error	Observational studies are not trials, but investigators could possibly create study parameters to describe here. It would require new controlled terminology and could be burdersome if it were considered a "required" dataset in observational research.

Table 5 SDTM datasets required or expected by FDA and the challenges these requirements would present in observational research. Adapted from[11]

In addition to these dataset-level conformance rules, Table 6 describes several variable-level expectations and the challenges those may present. Given the nature of the data sources in observational studies, potentially any required/expected - or conditionally warranted/useful) SDTM variable could be missing. Potentially many more expected variables than the listed ones can present challenges for specific studies. Additionally, when it is expected that an entire dataset would most likely be wholly missing (as described in Table 5), required and expected variables from that dataset are not all shown here. It is not likely that investigators would be able to produce even partial datasets in these cases and would therefore not encounter validation errors from individual missing variables[11].

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Variable(s)	Domain	Core	Challenge presented
RFSTDTC/ RFENDTC	DM	Expected	Study reference periods will not always be relevant. Defining these dates can be challenging. Sometimes dates will be missing altogether.
RFXTSDTC/ RFXENDTC	DM	Expected	Observational research does not include regimented exposure to a protocol-defined drug. Phase IV studies/ Post-marketing surveillance could possibly provide these.
SITEID	DM	Required	Observational research includes observations from across healthcare and clinical settings. These will likely vary and not be available in the data anyway.
ARM / ARMCD ACTARM / ACTARMCD	DM	Required	There are no arms to describe in observational research.
VISITNUM	Multiple	Sometimes required	The concept of "visit" may not be relevant in observational research.
EPOCH	Multiple	Sometimes required	Use cases for observational research have not been explored. Existing controlled terminology is specific to clinical trials.

Table 6 Examples of SDTM variables required or expected by FDA and the challenges these requirements would present in observational research. Adapted from[11]

CDISC and PhUSE are currently working on achieving further interoperability.

Healthcare standards also differ from the ones used in clinical research and cohort studies with HL7 standards being the most widely used. For an overview of the healthcare standards landscape the reader can refer to [12].

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

5.2. Interoparability

Interoperability, as defined by the Healthcare Information and Management Systems Society (HIMSS), "is the ability of different information systems, devices or applications to connect, in a coordinated manner, within and across organizational boundaries to access, exchange and cooperatively use data amongst stakeholders, with the goal of optimizing the health of individuals and populations."

It is now well understood that semantic interoperability relies on the adoption of interoperability standards (reference information models/templates and terminologies) that support information sharing among systems[13]. Healthcare information (clinical facts, decisions, activities, workflows) need to be standardized in order to be interoperable and used by humans or machines in contexts different than the original collection purpose.

Nevertheless, the emergence of operational solutions for semantic interoperability is hampered by the inability of EHR applications to conform to interoperability standards[14–17]. These applications provide interfaces to health professionals in order to collect data in a way adapted to their use and incorporated with their daily practice but usually not conform to standards.

In order to collect healthcare information in an evolutionary manner taking into account local organizations and clinical characteristics, EHR applications are often based on clinical information models that are legacy systems, specific and locally implemented. Even when several care settings use the same commercial EHR application, there is very little sharing of common clinical information models between different institutions. Even within the same institution, the principles of structuring and coding clinical information and the level of granularity of information can vary depending on the health profession profile (doctors, nurses, physiotherapists etc.) and within these professions, depending on the specialty (cardiology, psychiatry, imaging, biology, etc.) or the activity mode (hospitalization, consultation, hospital medicine, general practice etc.).

In the context of clinical research, currently, the clinicians have to manually copy the results of therapeutic protocols and examinations from an EHR system into the Case Report Forms (CRF) which causes errors and work disruption as well as delays in reporting data. Similarly, the investigators have to manually select the eligible patients from the underlying EHR systems by examining the inclusion/exclusion criteria listed in the study design documents[18].

These challenges derive from the fact that the clinical research and the healthcare domains each use different standards as "models of use". As discussed earlier in this deliverable, CDISC standards are widely used in the clinical research domain, while in healthcare, the most widely used content and messaging standards are by HL7. Additionally, the terminology systems used are different: while MedDRA, WHODD, and CDISC terminology are commonly used in the clinical research domain; the prominent terminology systems in healthcare are SNOMED CT, LOINC, and ICD-10.

Several efforts have tried to bridge the gap between clinical research and healthcare. The EHRCR Functional Profile Working Group defined the HL7 EHR Clinical Research Functional Profile[19] that provides high-level functional requirements necessary for using EHR data for regulated clinical research. It also provides a roadmap towards an evolutionary process of integrating the environment that provides both patient care and data for clinical research. It encourages EHR vendors to incorporate functions into their products that are necessary to utilize the EHRs as a direct data source for clinical studies.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

The Biomedical Research Integrated Domain Group (BRIDG)[20] developed the domain analysis model (DAM), which harmonizes CDISC data standards with the HL7 reference information model (RIM). HL7 Regulated Clinical Research Information Model (RCRIM) Work Group used parts of this DAM as domain message information models (DMIMs) to develop related HL7 message specifications.

Although BRIDG aimed at harmonizing CDISC and HL7 domains, it is still not possible to achieve automated or semiautomated semantic interoperability at the message level, because currently the BRIDG model semantics is available only to human experts: the BRIDG DAM is represented in Unified Modeling Language (UML) to be used by domain experts to build implementation specifications, and the mappings between the model and the standards harmonized by it are available in spreadsheets. Therefore, it cannot help automating end-to-end interoperability at the message level.

Another effort for providing interoperability between clinical care and clinical research systems is the Integrating the Healthcare Enterprise (IHE) Drug Safety Content (DSC) Profile[21] and the Clinical Research Document (CRD) Profile[22] that are defined on top of the Retrieve Form for Data Capture (RFD) Profile[23]. These profiles reuse the available standards in clinical care and research domains to achieve interoperability. However, in the DSC and CRD content profiles, the interoperability is achieved through hard-coded mappings between clinical research and healthcare standards.

FDA highlighted the interoperability problem between clinical research and healthcare in their guidance document, "Use of Electronic Health Record Data in Clinical Investigations," issued in July 2018.

5.3. Harmonisation of content

A big part of cohort datasets are kept and analysed in data silos and not sufficiently shared. If properly integrated into the clinical life cycle, such data collections could offer a unique opportunity to drive scientific discoveries and improve healthcare. Developing the harmonisation of health data-described as the sum of all "efforts to combine data from different sources and provide users with a comparable view of data from different studies"- is urgently needed to improve clinical research and practice.

The benefits of harmonizing and pooling biomedical databases are numerous. Integrating harmonized data from different populations allows achieving sample sizes that could not be obtained with individual studies, improves the generalizability of results, helps ensure the validity of comparative research, encourages more efficient secondary use of existing data, and provides opportunities for collaborative and multi-centre research. Policy makers, funders, publishers and researchers alike have been highlighting the importance of harmonisation and collaborative use of data and biosamples in population health and biobanks over the last 10 years[24–28].

The harmonisation of health data is a complex procedure which involves significant changes in how data are collected, shared and linked. Pezoulas et al. provided an overview of medical data harmonisation in[29]. According to the Deliverable 2.1, harmonisation can be either prospective, when modifications occur in the study design to subsequently render the pooling of data more

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

straightforward, or retrospective, when pooling is performed with data collected previously according to different study designs. In practical terms, harmonisation can be achieved through two distinct but complementary approaches, namely a "stringent" and a "flexible one"[30].

5.3.1. The stringent approach

The stringent approach is an ideal strategy which involves the harmonisation process to cohort data that have been collected under common collection criteria and operating procedures[29,31,32], where the common data collection criteria refer to the adoption of identical study specifications (uniform measures) between the clinical studies that participate in the data harmonisation process. This approach is what in Deliverable 2.1 was also called a prospective harmonisation strategy.

These specifications include: (i) common inclusion and exclusion criteria for the definition of the population subgroups, (ii) common follow-up time periods, and (iii) a common set of qualitative and quantitative measures (e.g., therapies), among others. These specifications together constitute a data collection protocol and are exclusively designed by domain experts who are able to identify (i) the domain of the field of interest (type of study), (ii) the set of measures that should be collected for the specified study, and (iii) the standardized measurement units for the recommended set of measures for the particular type of study[29,31].

According to this approach, the studies that participate in the data harmonisation process must be initially designed to meet these specifications to be harmonized and finally synthesized, otherwise the data harmonisation process won't work. These requirements are strict and limited to only a small portion of cohort data sources that adopt common data collection criteria and standard procedures. The majority of the cohort studies do not follow identical procedures for the data generation process and thus stringent harmonisation remains a conceptual and ideal strategy for the scientific community.

Especially in the case of retrospective studies, the stringent approach is inapplicable. The stringent approach can be meaningful in the case of a prospective study or perhaps in a cross-sectional study which focuses on data that have been obtained at a specific time point although it would require a substantial amount of time to be prosperous[29].

5.3.2. The flexible approach

The stringent method is a strict and a rather ideal approach that significantly limits the statistical power of the data harmonisation process because it obscures the integrity of the produced harmonized data through the underlying information loss and limits the harmonisation to a small portion of data that have been collected under the same standard operating procedure. An alternative approach that aims to deal with the limitations that are posed by the stringent approach is flexible harmonisation[29,31]. The flexible approach allows a certain level of heterogeneity between the data which participate in the harmonisation process.

Therefore, the flexible approach can support the harmonisation of both prospective and retrospective data as far as the level of compatibility between them is well defined. Through this manner, the flexible methodology envisages to enable the harmonisation of data that do not necessarily need to be homogeneous or obtained under a common data collection protocol criterion with equal-sized populations. In flexible harmonisation, the level of heterogeneity of the data directly affects the percentage of harmonized variables across them. This implies that the amount of flexibility is constrained to a specific set of requirements that need to be defined. That

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

is, the set of clinically relevant parameters (factors) that will be common among the heterogeneous data. Of course, the clinical domain where the data that participate in flexible harmonisation belong to must be common. To facilitate flexible harmonisation, the clinical experts must first define a set parameters (variables) that will serve as the core set for the domain of interest allowing for a specific level of flexibility regarding the data collection protocol and the standard operating procedures[29,31].

Therefore, flexible harmonisation is constrained to specific outcomes that are defined by the clinical experts. In the prospective case, the core set of variables is defined and agreed to by the experts so as to allow a specific level of flexibility during the recording of the follow-up data. In the retrospective case, the core parameters are combined together with pairing rules to identify potential associations with those from the heterogeneous data and thus quantify the harmonisation accuracy. The flexible approach has a much higher research value and overall applicability than the stringent approach, although, in both cases, certain compatibility criteria must be carefully defined so that harmonisation can be feasible.

The compatibility criteria are expressed in the form of a set of standard variables, i.e., a core set of variables that describe the requirements of the research domain of interest. In both cases, however, the standard model is defined by the experts in the field in such a way to: (i) be in line with the majority of the parameters within the data that are collected by different research centers and (ii) explicitly describe the domain knowledge of the disease under investigation. This means that the experts select the variables of the standard model by taking into account: (i) the contribution of each variable toward the efficient description of the disease's domain knowledge and (ii) the extent to which these variables are present in the majority of the data that exist under each research center[29,31,32].

It is not in the scope of this deliverable to provide an extensive review on the harmonisation and data integration methods but rather to touch upon the challenges and limitations that these methods can have. For an extensive report on methods for harmonisation and data integration in cohort studies the reader is referred to Deliverable 2.1.

5.4. Data protection issues

Data sharing within a research or healthcare context is a prerequisite for gaining new scientific insights and discovering new medical treatments. At the same time, data sharing raises questions on aspects such as data quality (standardization, harmonisation, comparability, interoperability, methodology), the privacy of the subjects participating in a study, and the rights of these participants regarding their own data. Both the public and patients are in general positive towards data sharing for research and healthcare purposes, but not unconditionally: Who has access to their data and how it is used are recurrent questions[33,34].

The purpose of the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679 of the European Parliament and of the Council) is to protect all EU citizens from privacy and data breaches in today's data-driven society[35]. GDPR came into effect on the 25th May 2018 following a 2-year transitional period granted by the European Parliament and repeals the Data Protection Directive 95/46/EC[36].

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

In the frame of this deliverable, the GDPR and its implications for cohort initiatives (population cohorts, patient cohorts and clinical trials) are briefly discussed highlighting the main legal, ethical and practical challenges of handling data.

5.4.1. The implications of GDPR for cohort initiatives

In the EU, GDPR has changed the landscape of data protection from that outlined under the Directive to a setting that is protected as a fundamental right in Article 8 of the EU Charter of Fundamental Rights, and recognizes that everyone "has the right to the protection of personal data concerning him or her"[36]. In contrast to a Directive, a Regulation is enforceable by law. Central to data protection is the concept of personal data itself. Many of the principles that form GDPR reflect the core principles of the Directive and the definition of personal data, as outlined in Article 4(1) of GDPR, includes "any information relating to an identified or identifiable natural person (data subject)". This includes names, surnames, home address, email address, or an identifier number or data held by a hospital or laboratory that could be used to identify a living individual. In addition, the existence of special categories of personal data, referred to as sensitive personal data, adds another layer of complexity. Sensitive personal data are outlined in Article 9(1) GDPR and include data pertaining to ethnicity, sexual orientation, religious beliefs, trade union membership, and genetic data. Genetic data is defined as "personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained".

Compliance with GDPR starts with awareness, understanding of the data subject rights, choosing the appropriate legal basis for data processing (Article 6 GDPR), and understanding the principles that are embedded in the GDPR, including those relating to the processing of personal data. It is stated under Article 4(2) of GDPR that virtually any use of personal data, from collection and recording to retrieval and dissemination, storage, and finally erasure or destruction, constitutes "processing" and comes with significant accountability.

A data controller is an individual or legal person(s) such as a company, department or organization, which under Article 4 of GDPR "determines the purposes and means of the processing of personal data". In Article 26, GDPR introduces the possibility of having more than one data controller or "joint controllers". Joint controllers can determine the purpose and means of data processing, although this does not always imply equal responsibilities. The data processor is a separate legal entity. According to Article 4 of GDPR, the data processor is a natural or legal person, public authority, agency, or other body that processes personal data on behalf of the controller. Moreover, data processors need to assist controllers in various circumstances when relevant, for example, in a potential personal data breach notification or in considering a Data Protection Impact Assessment (DPIA). The principles of GDPR Article 5, regarding personal data processing, apply to both data processors and controllers. Examples of data processors in health research include transcription services, DNA sequencing/translation services, biobanking/data repositories, etc. Any organization to which sample or data manipulation is outsourced is considered a data processor.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

GDPR allows for exemptions to data processing applicable to clinical and scientific research that have nevertheless caused great confusion among the scientific community and are subject to a continuous debate[38–41]. The GDPR application is inconsistent throughout the European Economic Area (EEA), leading to multiple misunderstandings and adverse effects for the delivery of clinical and scientific research. The European infrastructure for biobanking, BBMRI-ERIC, has initiated a European Code of Conduct for health research which may result in more harmonisation and clarification of the "vague" terms of the GDPR. ECRIN is participating in this activity. Besides, several projects have been working on clarifying the legal landscape across the EU and providing guidance to facilitate cross-border data sharing: Aegle, EOSC-Life, B1MG, SYNCHROS (in WP3), CINECA, EU-STANDS4PM, SIENNA, EuCanImage (launched October 2020), EUCANCan to name but a few.

The interplay between the GDPR and the Clinical Trials Regulation (CTR) is expected to be another point of confusion. The European Data Protection Board (EDPB) aimed to clarify the relation between GDPR and the CTR in their "Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection Regulation (GDPR) (art.70.1.b))". The EDPB has concluded that informed consent (as obtained from participants in clinical trials), should not be confused with "consent" as a legal basis under GDPR. The EDPB has identified issues with obtaining freely given "consent" in the context of a clinical trial, given that data subjects needed to have a free choice and control in whether to give their "consent". Despite the clarifications provided by EDPB, their opinion is not legally binding, and more issues might arise when the CTR comes into effect.

The opinion highlights that "consent" cannot be regarded as a valid legal basis when there is a suggestion of a power imbalance between the data subject and the controller. This is the case in studies involving critically ill patients, who can be considered vulnerable and there might be a stark imbalance of power between the investigator and the patient providing the data. In this clarification, the EDPB suggested that in most cases "consent" is not an appropriate exemption for processing under GDPR. Instead, data controllers should seek to rely on either "public" or "legitimate interests", only relying on "consent" under very specific circumstances, that is when "consent" is freely given, specific, informed, and unambiguous, and withdrawal of the consent will not adversely affect the proposed use of the data. Despite this clarification, many authorities continue to request that "consent" is obtained from participants even in settings where informed consent has not been required - e.g., for cohort studies, which is having a profound effect on the deliverability of a research agenda[42].

One way to ensure data security when processing sensitive data is de-identification, which can be achieved *via* anonymization or pseudonymization. Anonymization involves the complete removal of any information that can lead to the identification of the individual, whereas pseudonymization involves the partial removal of the individual data with an additional storage of information that can indirectly lead to the identification of the individual (e.g., an identifier number). Pseudonymized data are still considered personal data and fall into the scope of the GDPR. Although the GDPR does take into consideration the "means reasonably likely" to be used for re-identification of individuals including "all objective factors, such as the costs and the

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

amount of time and effort required for identification", the question of when should data be considered really anonymous is all but trivial and different aspects need to be taken into consideration[41].

In particular, re-identification of seemingly anonymous data in the era of big data has to be reconsidered and new technological advancements (or even realistically possible future advancements) to be taken into account. Rocher et al. for example demonstrated in a recent publication that using their model it was possible to re-identify 99.98% of American citizens in any de-identified dataset using 15 demographic attributes[43].

As a result, data protection authorities are adapting a strategy of reluctance and consider informed consent mandatory for the processing of healthcare data, although not necessarily required by the GDPR and in a different way than the one recommended by the EDPB in this opinion. This conflation of informed consent and data security provokes uncertainty and inconsistent views from authorities. Ethics boards now demand GDPR statements in advance, which are not readily provided by data protection authorities, which might stall research projects already in their conception phase. Cohort studies, however, play an important role in informing biomedical research and healthcare and should not be hampered.

To highlight the huge variation in obtaining ethical permission for cohort studies in Europe, de Mange et al. surveyed the experience of getting ethics approval in a cohort study including very old intensive care patients[44]. The authors examined the situation in 16 European countries and reported a large variety of the ethical processes and outcomes from either national Ethics Committees, regional Ethics Committees or Institutional Review Boards approval about an identical study protocol. In most countries, more than one level of ethical approval had to be approached. Often a national Ethics Committee needed to assess the research protocol and, once approved, the local Institutional Reciew Board needed to consent with the protocol as well. The time from applications to the decision was unexpectedly and unnecessarily large and only a few countries received feedback within a month from their national Ethics Committee[45,46]. The study also highlights misunderstandings among the research community as to whether or not informed consent is required (in this study patients were critically ill and, in some cases, unable to consent) and concludes that half of the research coordinators were not satisfied with the process and the timelines.

In another study, Kho et al. concluded in a systematic review that mandatory informed consent introduces selection bias and leads to differences in characteristics between participants and non-participants[47]. Moreover, they reported studies where several centers could ultimately not participate due to their ethics board demand for informed consent. This is an issue because research with health data needs to be unbiased, or it will mislead care providers, policy-makers and patients alike.

GDPR has resulted in inefficient distributed analysis of international data. For example, the International Genomics of Alzheimer's Consortium and the U.S.-based Alzheimer's Disease Sequencing Project based at the University of Pennsylvania have been unable to pool personal data on a single server because EU investigators believe that the GDPR prevents them from

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

sharing the European personal data with U.S.-based researchers. This creates a scientifically compromised, inefficient, and more costly distributed analysis of international Alzheimer's disease data because investigators must run identical analyses on segregated pools of data in different locations. This distributed analysis model both slows research and limits the scope of research projects in which they can engage[48].

All this is paradoxical since the GDPR itself highlights the importance of big-scale data collection and analysis. It states that "by coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer, and depression".

To conclude, GDPR presents several obstacles for data sharing between infrastructures, including failing to provide a clear basis for processing personal data for secondary research purposes[41]. The few regulatory pathways that GDPR provides lead to big variations among EU member states and these variations add significant barriers to secondary research uses of data and biosamples. These issues derive from the fact that GDPR was intended as a law of general applicability that would offer protection to personal data when processed in all sectors of the EU economy. Thus, the challenges it has created for scientific research were likely unanticipated and unintentional[38].

Further guidance would be beneficial in areas where GDPR has created confusion for the research community. Especially regarding:

- the concept of anonymization, specifically whether pseudonymized data can be considered anonymized data under certain circumstances (and in that case under which circumstances);
- the basis for processing personal data for secondary research and healthcare purposes;
- the basis for cross-border transfer of personal data for research and healthcare purposes.

The major legal, ethical and practical challenges as stated in the literature are presented below [35-38,41,49,50].

5.4.2. Summary of legal, ethical and practical challenges

• The individuals' personal data must be processed with respect to the individual's rights and freedoms.

• Individual consent forms must be obtained by anyone who wishes to process personal data according to the purposes of the processing. The individual must be informed about all types of processing, which involve his/her personal data and provide his/her informed consent according to the consequences (i.e., the risks) that might arise as a result of the processing of his/her personal data.

• The individuals must be given the right to (i) access, rectify, and erase their personal data, (ii) object and restrict the processing of their data, and (iii) request to obtain their data when they wish to do so.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

• The risks behind the processing of the individual data (i.e., risk assessment) must be clearly stated.

• Any cross-border data flows involving sensitive data must be subject to international legal requirements and data protection principles that require the cooperation of international supervising authorities.

• The sensitive data must not be transferred to third countries (parties) without the fulfillment of adequate data protection requirements and principles under the international data protection regulations.

• Personal data must be de-identified by pseudonymization or anonymization processes.

• Common international standards and definitions must be introduced for the terms data anonymization and data pseudonymization to avoid any confusion during data collection and data processing.

• The heterogeneity of the data protection laws across different countries, i.e., the existence of legal and ethical inequalities across developed and developing countries, as well as ethical issues during the data collection process that is introduced by different countries.

• Additional bioethical regulations in the case of genome-wide studies must be taken into consideration. The health policies regarding the processing of genetic data are usually stricter and harder to follow.

• The negative implications of big data in privacy protection (e.g., the use of big data for the identification of individuals using information from social media or any other information from the internet).

• The negative effect of centralized databases in the case of data breach. It is easier for the hackers to breach centralized data repositories instead of federated databases where the access to the rest of the repositories can be blocked in the case of data breach in a specific repository[49]. On the other hand, federated databases pose significant computational challenges (see subchapter about pros and cons of each system).

• The early detection and prevention of personal data information leaks in large scale-platforms. Large-scale platforms might be hard to breach, but a successful attempt can have serious consequences.

A detailed overview of the ethical, legal and practical challenges on data processing in research is provided in the D3.1 of the SYNCHROS project. Pezoulas et al. also address this issue in their book Medical data sharing, harmonisation and analytics with a focus on technical challenges[27].

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

6. CONCLUSIONS

The present deliverable provided an inventory of 54 cohort initiatives with an identified infrastructure for the data analysis. The initiatives listed here constitute a subset of the initiatives identified in the SYNCHROS repository (https://repository.synchros.eu). Both the data layout within the infrastructures of the initiatives and the software used for the data analysis are discussed in detail. With respect to the data layout within the infrastructure, both the centralized and federated data models are presented and compared. The features of the main software for the management and the analysis of the data are also presented.

Finally, the challenges with regards to the data analysis within the infrastructure are discussed. The first challenge focuses on the data standards used in the different types of studies (population cohorts, patient cohorts, clinical trials). Interoperability of standards should be promoted to avoid creation of non-interoperable data silos. Until this is achieved, harmonisation of content will be necessary: core datasets should be defined and promoted, according to a consensus procedure similar to the one used to define core sets of outcome measures for clinical trials, or patient-reported outcome measures. Ethical challenges and data protection issues are also highlighted, focusing on the obstacles that arose since the implementation of the GDPR.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

7. REFERENCES

1. Larsson A. The Need for Research Infrastructures: A Narrative Review of Large-Scale Research Infrastructures in Biobanking. Biopreserv Biobank. 2017;15(4):375-383. doi:10.1089/bio.2016.0103

2. Pezoulas VC, Exarchos TP, Fotiadis DI. Chapter 1 - Introduction. In: Pezoulas VC, Exarchos TP, Fotiadis DI, eds. *Medical Data Sharing, Harmonization and Analytics*. Academic Press; 2020:1-18. doi:10.1016/B978-0-12-816507-2.00001-3

3. Missikoff M. The Future of Enterprise Systems in a Fully Networked Society. In: Ralyté J, Franch X, Brinkkemper S, Wrycza S, eds. *Advanced Information Systems Engineering*. Lecture Notes in Computer Science. Springer; 2012:1-18. doi:10.1007/978-3-642-31095-9_1

4. Weil P. Don't just lead, govern: How top-performing firms govern IT. *AMIS Quarterly Executive*. 2008; 3(1)

5. Bolman LC, Deal TE. Reframing Organizations, 4th ed. San Francisco, California: John Wiley & Sons. 2008

6. Weill P, Ross J. IT Governance: How Top Performers Manage IT Decision Rights for Superior Results.*Harvard Business School Press*, Boston; 2004.

7. Hevner AR, March ST, Park J, Ram S. Design Science in Information Systems Research. *MIS Quarterly*. 2004;28(1):75-105. doi:10.2307/25148625

8. Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. *J Biol Res (Thessalon)*. 2015;22(1). doi:10.1186/s40709-015-0032-5

9. http://www.cdisc.org

10. https://www.fda.gov/media/80613/download

11. https://www.lexjansen.com/phuse-us/2019/si/SI08.pdf

12. Oemig F, Snelick R. Healthcare Standards Landscape. In: Oemig F, Snelick R, eds. *Healthcare Interoperability Standards Compliance Handbook: Conformance and Testing of Healthcare Data Exchange Standards*. Springer International Publishing; 2016:75-103. doi:10.1007/978-3-319-44839-8_3

13. Do NV, Barnhill R, Heermann-Do KA, Salzman KL, Gimbel RW. The military health system's personal health record pilot with Microsoft HealthVault and Google Health. *J Am Med Inform Assoc*. 2011;18(2):118-124. doi:10.1136/jamia.2010.004671

14. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *Journal of Internal Medicine*. 2013;274(6):547-560. doi:https://doi.org/10.1111/joim.12119

15. Nordo AH, Levaux HP, Becnel LB, et al. Use of EHRs data for clinical research: Historical progress and current applications. *Learning Health Systems*. 2019;3(1):e10076. doi:https://doi.org/10.1002/lrh2.10076

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

16. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *npj Digital Medicine*. 2019;2(1):1-5. doi:10.1038/s41746-019-0158-1

17. Richesson RL, Krischer J. Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions. *J Am Med Inform Assoc*. 2007;14(6):687-696. doi:10.1197/jamia.M2470

18. Laleci GB, Yuksel M, Dogac A. Providing semantic interoperability between clinical care and clinical research domains. *IEEE J Biomed Health Inform*. 2013;17(2):356-369. doi:10.1109/TITB.2012.2219552

19. https://www.hl7.org/implement/standards/product_brief.cfm?product_id=16

20. https://bridgmodel.nci.nih.gov/

21. https://wiki.ihe.net/index.php/Drug_Safety_Content

22. https://wiki.ihe.net/index.php/Clinical_Research_Document

23. https://wiki.ihe.net/index.php/Retrieve_Form_for_Data_Capture

24. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol*. 2009;24(12):727-731. doi:10.1007/s10654-009-9412-1

25. Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. *Nature Genetics*. 2012;44(2):121-126. doi:10.1038/ng.1054

26. Budin-Ljøsne I, Isaeva J, Maria Knoppers B, et al. Data sharing in large research consortia: experiences and recommendations from ENGAGE. *Eur J Hum Genet*. 2014;22(3):317-321. doi:10.1038/ejhg.2013.131

27. Walport M, Brest P. Sharing research data to improve public health. *The Lancet*. 2011;377(9765):537-539. doi:10.1016/S0140-6736(10)62234-9

28. Vickers AJ. Making raw data more widely available. *BMJ*. 2011;342. doi:10.1136/bmj.d2323

29. Pezoulas VC, Exarchos TP, Fotiadis DI. Chapter 5 - Medical data harmonization. In: Pezoulas VC, Exarchos TP, Fotiadis DI, eds. *Medical Data Sharing, Harmonization and Analytics*. Academic Press; 2020:137-183. doi:10.1016/B978-0-12-816507-2.00005-0

30. Geneviève LD, Martani A, Mallet MC, Wangmo T, Elger BS. Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review. *PLoS One*. 2019;14(12). doi:10.1371/journal.pone.0226015

31. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: consolidating data harmonization--how to obtain quality and applicability? *Am J Epidemiol*. 2011;174(3):261-264; author reply 265-266. doi:10.1093/aje/kwr194

32. Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol*. 2017;46(1):103-105. doi:10.1093/ije/dyw075

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

33. Grande D, Mitra N, Shah A, Wan F, Asch DA. Public preferences about secondary uses of electronic health information. *JAMA Intern Med*. 2013;173(19):1798-1806. doi:10.1001/jamainternmed.2013.9166

34. Kalkman S, Delden J van, Banerjee A, Tyl B, Mostert M, Thiel G van. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *Journal of Medical Ethics*. Published online November 12, 2019. doi:10.1136/medethics-2019-105651

35. 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). Off J Eur Union 2016;119:1-88.

36. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Off J Eur Union 1995;281:31-50.

37. Charter of fundamental rights of the European Union. Off J Eur Union. 2012;326:391-407.

38. van Veen E-B. Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. *European Journal of Cancer*. 2018;104:70-80. doi:10.1016/j.ejca.2018.09.032

39. Dove ES. The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era. *J Law Med Ethics*. 2018;46(4):1013-1030. doi:10.1177/1073110518822003

40. Pormeister K. Genetic research and applicable law: the intra-EU conflict of laws as a regulatory challenge to cross-border genetic research. *J Law Biosci*. 2018;5(3):706-723. doi:10.1093/jlb/lsy023

41. Peloquin D, DiMaio M, Bierer B, Barnes M. Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics*. 2020;28(6):697-705. doi:10.1038/s41431-020-0596-x

42. Metnitz PGH, Zajic P, Rhodes A. The General Data Protection Regulation and its effect on epidemiological and observational research. *The Lancet Respiratory Medicine*. 2020;8(1):23-24. doi:10.1016/S2213-2600(19)30411-4

43. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*. 2019;10(1):3069. doi:10.1038/s41467-019-10933-3

44. de Lange DW, Guidet B, Andersen FH, et al. Huge variation in obtaining ethical permission for a non-interventional observational study in Europe. *BMC Medical Ethics*. 2019;20(1):39. doi:10.1186/s12910-019-0373-y

45. Al-Shahi Salman R, Beller E, Kagan J, et al. Increasing value and reducing waste in biomedical research regulation and management. *Lancet*. 2014;383(9912):176-185. doi:10.1016/S0140-6736(13)62297-7

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

46. Goodyear-Smith F, Lobb B, Davies G, Nachson I, Seelau SM. International variation in ethics committee requirements: comparisons across five Westernised nations. *BMC Med Ethics*. 2002;3:E2. doi:10.1186/1472-6939-3-2

47. Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ*. 2009;338:b866. doi:10.1136/bmj.b866

48. Bovenberg J, Peloquin D, Bierer B, Barnes M, Knoppers BM. How to fix the GDPR's frustration of global biomedical research. *Science*. 2020;370(6512):40-42. doi:10.1126/science.abd2499

49. Pezoulas VC, Exarchos TP, Fotiadis DI. Chapter 4 - Data protection. In: Pezoulas VC, Exarchos TP, Fotiadis DI, eds. *Medical Data Sharing, Harmonization and Analytics*. Academic Press; 2020:105-136. doi:10.1016/B978-0-12-816507-2.00004-9

50. Takabi H, Joshi JBD, Ahn G. Security and Privacy Challenges in Cloud Computing Environments. *IEEE Security Privacy*. 2010;8(6):24-31. doi:10.1109/MSP.2010.186

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884