SYNCHROS

WORK-PACKAGE 2
Strategy brief on harmonisation and integration methods, and analytic approaches to maximise the value of cohort data
D2.5

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	**Anastassja Sialm, Albert Sanchez-Niubo**
Contributors	Jerome Bickenbach, Juan R González
Reviewer	Ellen Vorstenbosch
Responsible Partner	SPF, PSSJD
Dissemination Level	Public
Nature	Report
Keywords	Methodology, Data harmonisation and integration, Analytical approaches, Cohort, Optimisation
Due Date	31/12/2020
Actual Submission Date	29/12/2020
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

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

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. CONTEXT

1.1 SYNCHROS Objectives and Specific Objectives

SYNCHROS is an EU Horizon 2020 project that aims to identify the methodological, practical, legal and ethical barriers and opportunities in cohort research. The main aim of SYNCHROS is to formulate a sustainable European strategy for the next generation of integrated cohorts. In this context, SYNCHROS is characterised as a coordination and support action project. Namely, the SYNCHROS project contributes to an international strategic agenda for enhanced coordination of cohorts globally. In particular, SYNCHROS addresses the practical, ethical, legal, and methodological challenges in optimising the exploitation of current and future cohort data. In so doing, SYNCHROS supports developments for a stratified and personalised medicine approach and facilitates health policy.

The present strategy brief is related to the methodological domain of the SYNCHROS project. Namely, we aim to identify the methodological problems faced by cohort researchers as well as provide solutions and recommendations for research practice.

1.1 Strategy Brief

The strategy brief (alternatively evidence brief or policy brief) is an internationally-recognized strategic tool of modern implementation science – which itself has developed from evidence-based medicine and knowledge-transfer methodologies to become the foundation for promoting the adoption and integration of practices and policies for individual clinical care, public health and health research.

SYNCHROS is a coordination and support action with the overall objective of addressing the practical, ethical and legal, and methodological challenges to optimising the exploitation of current and future cohort study data. Because of this aim, it is essential that SYNCHROS rely on implementation science to transfer what tends to be abstract and theoretical issues into practical solutions that can be accomplished in the context of existing research infrastructures and practice. Strategy briefs are the essential first step in implementation, as they provide both the scientific basis and agenda focus for the consensus-based, sustainable and strategic resolution by means of stakeholder dialogues.

The current strategy brief follows a well-established format. We begin describe and contextualize the central issues and a case is made, both for their relationship to the overall objective of the project but also their significance and priority. This is followed by an identification and prioritization of the key issues involved, in this case the methodological obstacles to optimisation and integration of data harmonization. Relying on the evidence that is set out in previous deliverables D2.1 and D2.3, each issue is presented in terms of potential options for realistic and feasible solution. Each option is motivated and evidence and argument presented. Finally, a recommendation for the best option is provided.

1.2 Cohort Study: A definition

A cohort study refers to a panel, longitudinal type of study design. It usually involves a group of people who share a common characteristic, event or habit (e.g., smoking), performing a cross-section at intervals through

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

time. A cohort study can be either retrospective or prospective. In the retrospective case, the study relates to data collected in the past (e.g., medical records). In the prospective case, the cohort study relies on the collection of new data.

1.3 Harmonisation and Integration of Cohorts: A definition

Harmonisation is a practice that improves the comparability of variables from different studies and thus reduces the heterogeneity across cohort studies. Therefore, harmonisation of data from different cohorts allows data to be integrated into the same data set. Integration by contrast, is a process that combines or pools the same data from different cohort studies into a coherent whole.

Harmonisation is not always possible because data may not be necessarily comparable. In such cases, data is aggregated rather than integrated. Data integration methods are closely related to analytical techniques depending on whether the data can be harmonized or not, and on whether individual data can be shared outside their respective institutions.

The value of harmonisation processes derives from its ability to integrate various types, levels and sources of data so that they could become comparable. This increases the sample size and statistical power. Since harmonisation increases participants' diversity, it extends the generalizability of results and allows to explore effect heterogeneity in depth. Harmonisation allows to ensure both the validity of comparative research and its reproducibility. Harmonisation processes support the use of existing data and resources and thus boost research efficiency. This opens opportunities for collaborative and multi-centre research. Finally, harmonisation brings together expert knowledge from across a range of disciplinary boundaries.

Harmonisation relies on three strategies:

- *Prospective Harmonisation*: Studies share the same study design, questionnaires and instruments for collecting biological, psychological and social measures.
- *Ex ante retrospective Harmonisation*: Studies use standard collection tools and standard operating procedures.
- *Ex-post retrospective Harmonisation*: Studies try to achieve commonality through data processing procedures.

2. PROBLEMS, BARRIERS AND SOLUTIONS FOR IMPLEMENTATION

2.1 TEMPORALITY

(a) Nature of the problem:

In general terms, harmonisation is hard to implement because it requires a considerable amount of **time.** The lag time is especially extensive because cohort study collaborators have to collect the required data, to include researchers with expertise in cohort research and to validate data *before* the harmonisation process starts. This makes harmonisation unsuitable for short time duration projects.

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

(b) Current and potential obstacles

Lengthy maturity lengths of large cohorts:

A general barrier to harmonisation implementation is that big cohorts are not easy to establish. They require complex logistical infrastructure, extensive financial resources and investment and represent a significant participant and research burden. Extended cohorts with data from many thousand participants need a significant time to mature before providing crucial research insight. The relatively **lengthy maturity rates** of such resources (their potential research value is fully visible only over time) means that even extensive financial and personnel engagement have no positive impact in the short term.

Harmonisation-Integration relationship:

In practice, there is some confusion regarding the **order between harmonisation and integration**. In practice, most researchers harmonise data first and integrate it second. Such an approach however, is damageable because the integration process reveals errors, differences and problems that the harmonisation process was not able to filter.

Extended time lags for data access and availability in EU based cohorts:

Harmonisation processes require a straightforward access to data. However, most current EU based cohorts include local approval procedures whose implementation requires a significant amount of time. In order to obtain local scientific and ethical approvals for data access, it is necessary to rely on cohort-specific local access committees. In practice however, information about such committees is available only after initial access requests.

This means that researchers do not have the means that adjust their proposals and harmonisation methodology especially since local ethical committees often tend to ask for complementary information. Future uses of data are subjected to the same constraints, as local ethical research committees have to decide on any secondary data uses. Since the time needed for approval is quite considerable (from 3 months to 1 year), this means that researchers do not have the time to develop an appropriate harmonisation strategy.

(c) Potential Solutions

In order to clarify the harmonisation-integrative relationship, it may be advisable to adopt an **iterative approach**. Integration can be applied before harmonisation, so that the feedback from integration processes can be used to better understand data and to correct the process of harmonisation.

The harmonisation-integration relationship can be clarified trough **federated analysis**. For federated analysis, the issue of whether harmonisation or integration should be applied first is not the primary concern. Instead, federated analysis allows each institution to control their data themselves. As a result, data can be either integrated with federated analysis (via previous harmonisation) or directly processed with **meta-analysis** (without harmonisation).

Extended time lags for approval and access to data for harmonisation can be partly solved through **FAIR** (Findability, Accessibility, Interoperability and Reusability) **principles**. There is a need for an EU-wide strategy for harmonizing protocols, adding meta-analytical levels to harmonisation methods and coordinating effort to keep data under the FAIR principles.

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

(d) Strengths and weaknesses of the solutions

Federated analysis allows each institution to control its data, which gives more flexibility for the parallel applications of harmonisation and integration processes. However, federated analysis and infrastructure requires considerable efforts for organizing and curating the data.

While FAIR principles are crucial for data sharing and data access for further harmonisation, they do not guarantee a good and sustainable quality of data. There is thus a risk that harmonisation and integration efforts may be based either on incomplete datasets or of bad data altogether.

(e) Our recommendations

In order to solve temporal issues (e.g., lengthy maturity rates, harmonisation-integration timing, time lags for data access), we recommend **internet-based networking technologies and database management systems** (e.g., DataSHIELD cf. *https://www.datashield.ac.uk/*). While data access approval requires extensive time, these networking technologies can readily provide the necessary support background for collaborative, multi-centre research in the meantime. Such technologies interconnect harmonised datasets and perform joint statistical analyses without pooling or sharing individual data.

Networking and database management technologies create a federated infrastructure that encrypts remote connections, user authentication and control and user access arrangements. This effectively enforces data privacy and confidentiality but also ensures that the data are compatible with FAIR requirements.

We recommend that data should be **accessible beyond the duration of the EU funding** for the original project, including data acquisition (e.g., NIH funds cf. *https://www.nih.gov/grants-funding*).

2.2 HARMONISATION PROTOCOLS AND DOCUMENTATION

(a) Nature of the problem:

Many researchers leading cohort initiatives do not make the harmonisation process transparent and rigorous enough for assessing its validity and ensuring its reproducibility in future initiatives. Researchers are often more focused on results of their cohort initiatives projects rather than on documenting their harmonisation process in detail. Such a situation is often due either to a lack of established documentation or to a lack of knowledge about how to document harmonisation information properly. As a result, there is often a lack of sustainable, documented protocols for data harmonisation processes, which means that such processes can be neither verified nor reproduced.

(b) Current and potential obstacles

Data Quality Issues in Ex-post retrospective harmonisation methods:

Ex-post retrospective harmonisation methods are particularly susceptible to problems arising from a lack of appropriate documentation. Namely, such harmonisation methods produce incomplete data and tend to misclassify data because the terminologies used are either ambiguous or not specific enough.

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

Lack of Data Documentation in Post Harmonisation phase:

Crucial information on the harmonisation process is often not documented. For instance, the origins and validations of specific cohort variables are often missing.

Pre-set standards and protocols in prospective harmonisation:

Prospective harmonisation relies on strict standards and protocols from the onset in order to maintain comparability from the first phases of the harmonisation processes onwards. In practice, this means that all the cohort studies involved share the same study design, survey and meta-data. Such an approach is possible for multi-centre studies where researchers have the possibility to agree upon study design and data collection strategies. However, a reliance on multi-centre studies may be problematic: there may still be differences between study designs and data collection for cultural reasons.

Heterogeneous standards and documentation types across cohorts (consequence on repository):

Researchers may have problems in accessing and locating the appropriate cohort data for harmonisation because internal standards may differ across cohort studies. In practice, research projects involved in cohort studies may differ in their standards for medical procedures, record keeping and quality arrangements. This has harmful effects on the repositories' catalogues for cohort data: they tend to include only a general description of the collection content and summary statistics (such as observations number and variables). However, this is not sufficient because precise descriptions and indications of availability of values for each variable and each sample on an individual basis are missing. As a result, researchers are no longer able to identify the appropriate data for harmonisation in repositories catalogues.

Lack of information for catalogues:

Generating catalogues for a repository is crucial for harmonisation research: researchers need to know where to find the appropriate data for harmonisation. However, it is difficult to establish a *catalogue for samples and data* because of the lack of transparency in harmonising methodologies and processes. More precisely, cohorts often do not communicate detailed information on their research resources in harmonized ways. Namely, EU cohorts catalogues often do not contain sufficient information in order to devise viable research proposals, identify relevant samples and data for the research project and design an appropriate approach to cohort harmonisation.

(c) Potential Solutions

Ex-post retrospective harmonisation methods have the potential to misclassify data. In order to avoid such problems, the Maelstrom Research initiative advises to **classify all variables in a study into a standard variable classification taxonomy of 19 information areas and 148 sub-domains**. The information areas include all types of information collected by cohort studies.

The CINECA project assigns **ontologies to the data of each cohort study but in an automatic way**. It defines ontologies in standardized terms which have a unique identifier. These standardized terms are part of a hierarchy of relationships with other clearly defined ontologies.

Harmonisation (as a process) can also happen on two levels namely *(i)* **harmonized original data indexation** (e.g., biospecimens) and *(ii)* **harmonisation of variables and descriptors**. We think that in order to achieve these two levels, we need to collect data sample from several resources (e.g., repositories, biobanks) into a single

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

infrastructure. The format of such data integration and submission consist in either ''original'' or harmonized vocabularies.

Harmonisation efforts can rely on **two data formats** namely *(i)* vocabularies and *(ii)* availability information format. According to the vocabularies format, both harmonized and original variables can be mapped between vocabularies and thus be used for annotation of samples. Such an approach assumes that the grammar for description of terms is universal: it is possible to link terms across studies (and thus across vocabularies). This allows to integrate external shared vocabularies and ontologies to internal (and local) repository vocabularies.

(d) Strengths and weaknesses of the solutions

Maelstrom Research (cf. *https://www.maelstrom-research.org/*) can alleviate classification problems in ex-post retrospective harmonisation methods. However, the classification process of the variables for all the cohorts included in an initiative makes the harmonisation process much longer and thus, more expensive. Maelstrom Research promotes a very high-resolution type of cataloguing information that is particularly time consuming. Hence, it is necessary to find a balance between catalogue quality and documentation range. Other open-source solutions for cataloguing data (e.g., Molgenis cf. *https://www.molgenis.org/*) can provide an interesting alternative as they are used in large infrastructures.

(e) Our recommendations

Future initiatives aimed at harmonising and integrating data from different cohorts should follow systematic guidelines in accordance with the type of methodological scenario and the availability of data shared by the different cohorts. The process should be transparent, rigorous and well documented in order to justify its validity and to support other initiatives.

For avoiding the lack of documentation in the post harmonisation phase, we advise to follow the **Maelstrom Research requirements**. That is, the origin of the specific variables of the cohort studies and the validation of the harmonized variables should be thoroughly documented. In practice, this means that researchers should provide: ***(a)*** a definition of the variable to be harmonized ***(b)*** a description of the specific variables of the cohort study ***(c)*** an outline of the data process of harmonisation ***(d)*** a statistical description and validation of the harmonized variable and ***(e)*** an evaluation of the quality of the harmonized variable.

The problematic reliance on multi-centre in prospective harmonisation can be alleviated through funding agencies. Funding agencies may be best suited to initiate and coordinate prospective data harmonisation initiatives because of their comprehensive knowledge of research in a particular area, potential to leverage additional resources, ability to encourage collaboration among researchers and unique perspective on goals that extend beyond individual research projects.

The use of ontologies (cf. e.g., CINECA *https://www.cineca-project.eu/*) would help to standardize datasets and increase interoperability.

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

2.3 STANDARDIZATION AND COMPARABILITY

(a) Nature of the problem:

Standardization is one of the most commonly used data processing methods in harmonisation. However, its application is problematic because it assumes constructs can be measured in standardized ways across cohort studies. There is no way to ensure that such assumptions can be viable in practice. For instance, culturally specific and multifaceted constructs (such as depression for instance) should not be measured in standardized ways across studies. There are also strong indications that standards may be impractical for harmonisation processes:

- As technology evolves at a high rate, there is an important risk that a standard can become obsolete, as it does no longer correspond to technology developments.
- Standardization is not compatible with research as it strongly limits innovation and prevents finding valuable variables and relationships outside of pre-set parameters.
- Paradoxically, given the excessive number of standards available there is a need to standardize these standards. There is however, no ground for developing such a general, ''universal'' standard in the first place.

Comparability is a corollary of standardization (as the one is not possible without the other). Comparability is in fact, a central aim for harmonisation efforts as they attempt to make data as comparable as possible. In data processing, the dilemma is that variables are supposed to be harmonised for equivalent content information across studies but that this equivalence often ''sacrifices'' the construct concerned by simplifying it (e.g., depression). In practice, comparability is not always possible because data collection across different studies is characterised by its heterogeneity. This means that data derived from different data collection approaches are not comparable: any harmonisation approach would lead to biases. As a result, harmonisation is no longer possible and researchers have to settle for the data integration and integrative analysis instead. However, there is still no agreed ways on how to pool key measures across studies for simultaneous analysis.

(b) Current and potential obstacles

Inferential Equivalence issues:

A crucial point in any harmonisation issue is to decide on whether data from the studies chosen are inferentially equivalent. Inferential equivalence stipulates that constructs should be combined only if they are already comparable enough in terms of the meaning, format and function beforehand. This means that the central question is not whether data and constructs can be combined but rather whether they *should* be combined in the first place. To a large extent, the type of research questions and analysis depends on judgements on inferential equivalence, on the possibilities for harmonized variables across studies and on the format these harmonized variables can take. However, determining inferential equivalence is difficult because there is a lack of transparency in harmonisation methodologies (c.f. Documentation and Protocols section). As a result, there are significant problems with the replication of analyses and the evaluation of the validity of results.

Difficulties for epidemiological and biological studies:

Harmonisation processes in general and standardization processes in particular, are hard to implement in epidemiological studies. First, researchers in this field are not used to share data and they rarely use the same

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

standards for harmonisation. Second, the harmonisation process for epidemiological research is rather complex as it relies on questionnaires and depends on the cultural context of the research. Documentation on the harmonisation process is either lacking or not recorded (c.f. Documentation and Protocols section). As a result, researchers taking over a particular epidemiological project have to start from scratch in order to re-design an appropriate harmonisation process. For studies with biological data, harmonisation (and thus comparability and standardization) is often not possible. Namely, such cohort research projects tend to use different techniques in different waves for measuring the same variable. In such conditions, harmonisation and comparison become impossible and data pooling analyses may be advisable instead.

Quick technological changes for standards (omics):

Standards tend to change rapidly in relation to technologies. This can be harmful for prospective harmonisation methods within the omics domain. Namely, prospective harmonisation cannot be applied for imaging and genetic data in omics related studies. This is because technology in the genomics domain is changing and developing rapidly, which makes any prospective harmonisation process either irrelevant or redundant. Using the same standard and technology for collecting data thus becomes impossible: the standards do no longer correspond to the technological changes and as such, lose their value and relevance. This in turn, results in the collection of bad quality, incomplete data.

Ex ante retrospective harmonisation-Different standards for same measures:

The main problem is that there can be many different standards for the same measures, whereas the level of granularity or precision depends on the importance of that information in a certain research context. Therefore, the challenge is to unify standards and find unbiased conversions among different standards. However, the use of standardized terminology either requires extra steps or costs or may be difficult to understand because of overlapping terms, terminologies, data elements, and questionnaires.

Standardizations of harmonisation strategies according to data types-Ambiguity of Retrospective Strategies:

Retrospective strategies are not clear on whether they use ex-ante or ex-post harmonisation approaches. In fact, both ex-ante and ex-post harmonisation methods can be used depending on the subset types of data between cohorts. For example, socio-economic data may be constructed more arbitrarily (ex-post) while data on psychological or clinical measures may be well standardised in advance (ex-ante). This complicates the determination of an initiative as either ex-ante or ex-post, as sometimes both approaches are used in the same initiative.

(c) Potential Solutions

Hospital records use a standard data model for cardiovascular research, publish it and encourage healthcare practitioners to use it further (so that feedback from users can be incorporated and update the data model~~).~~

If harmonisation (and thus standardization and comparability) is no longer possible, then it is possible to use a form of **aggregated data meta-analysis** for data integration instead. This is an attractive alternative: first it does not require as much time to mature as newly created cohorts and the research time-frame is thus significantly reduced (cf. Temporality section). This means that primary data collection requires less financial and logistical resources. Second, disaggregated data meta-analysis allows for the funders and financial bodies to maximise their investment returns because it contributes to an effective use of existing data.

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

Centralized collation of data results (cf. pooled analysis and integration) can compensate for standardization and comparability problems. First, it allows data to be analysed at the individual level. Second, it relieves the burden from individual partners and thus supports their engagement in the research project.

When there are issues of harmonisation and comparability, **meta-analysis** can be a solution to get integrated results.

(d) Strengths and weaknesses of the solutions

Standard data model for hospital records may be not applicable for cohort research in general. This is because hospital records are not designed for cohort research and have challenges of their own.

Aggregated data meta-analyses are vital for cohort data integration in cases where harmonisation and comparability are no longer possible. However, the complexity of data management and administration depends on whether data are physically relocated (e.g., within a particular infrastructure) or retained within the host institution. It is not always possible to track the extent to which data has been relocated. Data pooling strategies may thus operate under uncertainty (cf. Infrastructure). Another important problem is that data meta-analysis is often assumed to be equivalent across cohorts, which is likely to cause problems as data may be located in different places and infrastructure.

A centralized collation of data results is central for data integration (when comparability is not possible). However, it overburdens the research group chosen for the stewardship and cataloguing of data. In practice, a centralized collation of data results in a collection of studies that is either historic or too project-specific. Such studies stand little chance of being used outside of the project specific context, which causes difficult in obtaining information on the collection or derivation of particular variables. This results in incompleteness in data dictionaries as far as these studies are concerned.

(e) Our recommendations

It may be more practical to have good practice documents than standards. Initiatives should consider recommendations rather than standards.

A practical solution is to standardize descriptions of datasets in a short identification document that includes the characteristics, motivations and potential biases of the dataset.

We recommend to create transformation standards or links between different standards on the same measured variables whenever possible. Such practices provide equivalent scores for different scales that measure the same health construct. This means that initiatives that intend to harmonize data from different cohorts with standardized tools or measuring instruments can rely on the documentation of standardized descriptions of datasets and, if available, the links between different standards.

We recommend to aim for as high level of detail as possible for harmonisation procedures. That is, the cohorts involved in the research projects should provide specific information about each variable (i.e., what is measured, how the measurement was performed, who performed the measurement and the associated factors). This allows to compare what was measured in research projects. More importantly, it allows to make adequate judgments about inferential equivalence: it will be possible to determine whether variables are

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

equivalent for analysis or whether these variables should be converted and transformed into an equivalent status beforehand.

2.4 DEFINITION AND VALIDATION OF VARIABLES

(a) Nature of the problem:

Harmonisation decisions are to some extent subjective; researchers may disagree with the content of existing variables. In practice, it is difficult to enable researchers to access study-level data, allowing them to create new harmonised variables to their preferences. However, there are often no explicit definition and validation strategies for variables in place.

(b) Current and potential obstacles

Trade off and balances between precision and quantity:

The harmonisation process seeks a balance that is difficult to reach in practice. Namely, the aim is generally to select the appropriate variables' target for the research concerned in order to answer the research question and to harmonise variables according to this initial target later on. At the same time, it is necessary to ensure that no potentially valuable variable is rejected in the process. In order to achieve this however, it is necessary to harmonise an important number of variables. Thus, before an appropriate balance in the harmonisation process is reached, researchers require a lot of time before starting the harmonisation process in earnest.

In fact, such a balance in harmonisation process includes a trade-off between the integration potential for different cohort studies (and thus increased validity) and the specificity, context and ecological validities. Namely, context specific information from studies tends to be lost during the harmonisation process. For instance, ex-post retrospective harmonisation does not have the tools to balance adequately between precision and quantity.

Measurement differences over time:

Variations in the developmental period under study, the historical timing of the study, and the target populations may all result in measurement differences. This happens because as each study attempts to select instruments that are maximally valid for the developmental period and population under study based on the state of knowledge at that time.

(c) Potential Solutions

Variable validation is mostly useful to see if the data falls into normative models. Outside of this domain, recurrent data validation may be not necessary.

The same variable could be harmonised in different ways depending on the importance of that variable in the research (e.g., the variable ''education'' can be harmonised with four categories for a few studies or with one single category for a large number of studies). When studies implement different methods or use different instruments (e.g., questionnaires) for measuring the same characteristics or constructs, the observed variables need to be harmonized in order to obtain equivalent content information across 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 design of the research delineates variables of interest (VOIs), which may be different from the variables recorded by the original questionnaires and measurement protocols followed during sample collection. This means that data heterogeneity is tackled on a project-by-project basis: first, the aims of a research project are defined and then it is possible to identify and extract the data to be harmonised as well as the appropriate standardization strategy.

It is possible to use subject-specific correlation coefficient in order to measure harmonisation strength.

When harmonising scales, constructs or measurement instruments, data processing methods based on latent variable models are often used. This requires the use of equating procedures to account for the specificities of each cohort's data.

Some initiatives (such as Interconnect cf. *https://www.mrc-epid.cam.ac.uk/interconnect/global-network/*) structure their harmonisation strategy in terms of an importance order. Namely, depending on an initiative's research objectives, the harmonisation of variables may begin by outcomes first, then focus on exposure/intervention variables second and finally conclude with covariates.

(d) Strengths and weaknesses of the solutions

Using a subject-specific correlation coefficient allows to measure harmonisation strength. The problem is that harmonisation and processing strategies very much depend on the research question. It is thus necessary to start from the research question and see if other studies have similar questions and only then proceed with data processing and harmonisation. Using a subject-specific correlation coefficient for data processing is not always suitable because it restricts the harmonisation process from the onset.

(e) Our recommendations

We recommend the use of equating procedures across cohort's data alongside subject-specific correlation coefficients.

If a study collects data on a single construct, we recommend to establish an order of preference along with procedures for dealing with missing or inconsistent data. This allows to create multiple harmonised variables for each construct, balancing the often competing demands of resolution and coverage (i.e. number of included studies). This results in higher resolution variables that can rely on detailed data (when available) and lower resolution variables (for the inclusion of the large number of studies). This approach allows to create harmonised variables to reflect the different components of multi-dimensional constructs.

Harmonisation processes should be adapted according to the particular characteristics of each construct. That is, there are constructs that are generally consistent in terms of definition (e.g., demographic variables) and some constructs that are more conceptually complex (e.g., ethnicity). In the first case, harmonisation should be conducted by the core group of researchers (those in charge of the investigation). In the second case, harmonized variables can be created through an iterative process, with contributions from the investigative team(s), panel of experts and members from the steering committees.

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

2.5 DATA ACCESS AND DATA AVAILABILITY

(a) Nature of the problem:

Data access regulations in the cohort studies domain are an important challenge in most initiatives, especially when sharing data. On the one hand, there is EU legislation and the GDPR (General Data Protection Regulation) that restrict what should and should not be shared. On the other hand, local regulations and internal arrangements of the infrastructures where data is located generate additional requirements that may contradict the GDPR framework. This is harmful for cohort research because researchers experience considerable difficulties in devising appropriate strategies for harmonisation and integration strategies without proper data access.

(b) Current and potential obstacles

***Lack of coherence between different types and levels of cohort harmonisation and different levels of data sharing**:*

An important issue is that certain data types differ not only in the ways in which they can be harmonised but also in the extent to which they can be shared. In general, data (e.g., imagining data) can be anonymised and thus harmonised accordingly. As a result, these data will fall outside the GDPR framework and can be freely shared. However, some types of data (e.g., genetic data) are not susceptible to anonymization processes. This means that increasingly complex procedures are used in order to meet GDPR's requirements for data sharing: if a participant decides to withdraw, every researcher who had access to the data should be notified so that this participant will not be included in any future studies and in any future data processing.

There is also no sustainable mechanism for cohort integration and data access mechanism for depletable data (e.g., blood, serum). In practice, in order to access such type of data researchers would need to justify which percentage of the data they would use. This does not solve the issue because they may need more quantities from this depletable data as the research goes on (e.g., according to the results obtained, they may need more serum). Alternatively, asking for more depletable data may be impossible, as there may be no data left with very few recontacting options.

Material Transfer Agreement Heterogeneity:

In practice, access can be hampered by *MTAs* (material transfer agreement). Namely, since rules of cohort access are heterogeneous, it is quite difficult to devise an MTA for cohort data governance. The issue is institutional in the sense that most EU cohorts still prefer to rely on their own MTA whenever possible (rather than using a more general EU legislation). Moreover, there are generally differences in the ways in which cohort projects devise their routines for retrieving, preparing and transferring samples.

In projects involving multiple cohorts such differences may result in significant confusion in relation to data transfer. In technical terms, cohorts also differ in terms of the possibilities for automatic sample manipulation and personnel availability. This means the potential of cohort data transfer is not the same across cohorts, which in turn, causes difficulties in design and rationale for harmonisation strategies.

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

(c) Contradictory data access procedure across cohorts:

There are two main hurdles for access (*i*) heterogeneity in access governance and (*ii*) extensive amount of time needed for local approval procedures. For access heterogeneity, the main hurdle is to devise a concrete access strategy for individual cohorts. Namely, individual cohorts each have their own procedures and rules for organizing access to their samples and data. As a result, the researchers attempting to access multiple cohorts are faced with multiple and contradictory access procedures.

Binary types of data access:

Cohorts data sometimes rely on infrastructures with binary options. Namely, there is either a choice between highly restricted data access or open access data arrangements with little nuance in between. There is still a lack of reliable systems that could provide essential information across cohorts. In this sense, it is still challenging to implement power calculations for successful grant applications. Such difficulties are exacerbated in large meta-studies as the processing of data access application takes longer than the data analysis and harmonisation themselves.

(d) Potential Solutions

Access and availability issues can be alleviated by the enforcement of FAIR principles (data should be findable, accessible, interoperable and reusable). The aim is that computations systems should be able to find access and reuse data with none or minimal intervention.

A rigorous process to automate some harmonisation processes and validation, providing rules about minimum, maximum and usual percentages.

It is possible to design appropriate methods for tracking samples availability (especially in the omics domain). For instance, information could be provided for each sample (e.g., whether there is or is not a value for a given phenotypic or genotypic variable) but without revealing true values. As a result, it is possible to achieve a formalised methodological framework for the integration of data across cohorts.

(e) Strengths and weaknesses of the solutions

A formalised methodological framework allows making power calculations in order to get data access from ethical committees. However, providing information for each sample without revealing true values is possible for genetic data/omics but not for culturally bound data (e.g., depression).

FAIR does not guarantee data quality.

(f) Our recommendations

Europe needs to invest in a few centralized, ideally mechanisms for storing maintaining and enabling ongoing data sharing.

We recommend using linked samples and sampling descriptions. This would significantly improve the search for available samples in EU contexts. We consider that it makes sense to first create harmonised variables and then to make searchable information and annotations on data available in cohorts according to study relevant

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

categories (e.g., phenotype categories). This allows to solve privacy challenges related to the handling of real values (as real values are harmonised) and to provide an informational basis for research in the planning phase.

2.6 SUSTAINABILITY OF DATA INFRASTRUCTURES

(a) Nature of the problem:

Infrastructures are crucial for cohort data because it influences the type of harmonisation, integration and software techniques used. There are three main types of infrastructures for sharing individual data within the initiative namely: *(i)* the individual data is centralised in one institution or server (i.e., central location of data) *(ii)* the individual cohort datasets reside in different institutions (federated), mostly on the server of origin (i.e., data are in different locations) and *(iii)* mixed location types (some data are located centrally and some data locally).

On the EU level however, infrastructures of all types suffer from a lack of provision for sustained data keeping and for insufficient sustainability mechanism for cohort data. This is explained by EU's exceptionally heavy financial burden concerning cohort projects and initiatives.

(b) Current and potential obstacles

Heterogeneity in Governance Structure:

There is considerable heterogeneity in governance for infrastructures. Namely, relatively mature cohorts involve close collaborations with local scientists while recent cohorts with a service-oriented access governance structure require only limited scientific involvement in administrative tasks. Both types of governance involve hurdles for harmonisation and integration processes. Cohorts with limited scientific involvement in access governance have generally strong research support and a task-oriented team for administrative tasks. This is a possibility that is not open to all cohort types (because it requires resources).

More importantly, such a service-oriented approach can limit the scientific scope and potential value of research projects by limiting them to strict access rules (the view may be too general). By contrast, while the involvement of local scientists in access for mature cohorts may be useful for taking the particularities of cohorts in account, it also includes limitations because local scientists do not necessarily have the time to handle administrative tasks.

Centralized data, access and governance infrastructure may involve privacy issues:

Access to cohorts still presents important privacy issues, especially when cohort governance and access is centralized. In particular, the handling of real values triggers real privacy concerns, which in turn, prompts pseudo-anonymising and abstract-oriented strategies. However, this requires considerable harmonisation effort for researchers especially when the development and the description of the constructs at hand are concerned.

Federated Infrastructures-Difficulties in finding appropriate curating strategies:

Curating and organising data requires a lot of effort and there are for the time being, considerable difficulties in finding an appropriate curating strategy. The issue becomes even more complex because there are

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

considerable **data quality issues**. Namely, federated infrastructures include little control on the quality of data in the studies included in cohort studies.

***Federated infrastructures-Lack of interface with other infrastructures types (management problem)*:**

Federated infrastructures lack common interface with other infrastructure types (e.g., with software solutions, other federated infrastructures). As a result, specific tools, approaches and practices may well solve one particular problem but for researchers, it is vital to have a comprehensive solution for a complete range of data annotation and harmonisation options. Some federated platforms' interfaces may have partial overlap in functionality and methodology for harmonisation of data schema, but some aspects of the process will be different (e.g., central curation vs. distributed curation).

***Interoperability between stand-alone data harmonisation platforms and frameworks are relatively rare*:**

Generic evaluation mechanisms for interoperability projects and methodologies are still not available (e.g., except for BBMRI-ERIC cf. *https://www.bbmri-eric.eu/*).

(c) Potential Solutions

Software applications such as Opal (cf. *https://www.obiba.org/pages/products/opal/*) rely on infrastructures that are close to the federated mode. Data is based centrally or locally and researchers usually ask to each data infrastructure/website where data is located. Once the required data is found, researchers generally are able to integrate and collect data as well as to apply the appropriate analysis.

If harmonisation processes are well documented in repositories and infrastructures, it is possible to fully understand the conversion process. For instance, effective documentation allows to relate the harmonised variable back to the original question within the questionnaire that was harmonised in the first place.

For infrastructures, it is possible to share entire virtual machines (that contain the particular software version used). The aim is to minimize differences in the implementation of such software while enabling automated analysis reproduction.

It may be useful to establish small cohorts in parallel. These cohorts can be located in different geographical locations and will share a common methodology. Parallel cohorts may be easier to manage and less costly to maintain.

A cloud platform is not only able to give access to multiple cohorts regardless of the infrastructure used but also provides standardisation for the design of a common format for datasets.

(d) Strengths and weaknesses of the solutions

Small parallel cohorts evenly distribute resource management and work assignment among each study centre. However, reliance on multi-centre arrangements (cf. Protocols and Documentation section) is problematic because they require a clear consensus on methodology from research collaborators. Data collection is also more difficult to implement and requires significant financial resources. Finally, much like large cohorts, smaller parallel cohorts have a maturity rate problem. That is, they require a significant amount of time in order to mature and thus reveal research insights and outcomes only after a significant amount of time

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

Cloud platforms can provide solutions for infrastructure problems by providing *(i)* categorisation for terminology descriptions and *(ii)* harmonisation through ontology alignment. Cloud platforms can also perform a range of useful background technical tasks for harmonisation such as data curation and data imputation (i.e., automated methods for missing values). However, cloud platforms have difficulties to clearly define the primary data collectors (i.e., data providers) and secondary analysts (i.e., data processors).

Federated structures do not resolve privacy issues. However, they allow researchers to mitigate privacy-related obstacles and procedures in the planning phase. For large projects, this represents significant time-saving opportunities. In the later phases of research projects however, when the real data are to be exchanged and the application for the data access is to be filed, the procedural constraints involved in full data access applications are unavoidable (but a harmonisation strategy may be ready by this time).

(e) Our recommendations

We recommend using federated infrastructure or mixed infrastructures types. In order to streamline data location for harmonisation search, we recommend using cloud interfaces.

3. RECOMMENDATIONS SUMMARY

Temporality: We recommend ***internet-based networking technologies and database management systems*** (e.g., DataSHIELD *https://www.datashield.ac.uk/*).

Harmonisation Protocols and Documentation: We advise to follow the ***Maelstrom Research requirements*** (*https://www.maelstrom-research.org/*). Namely, the origin of the specific variables of the cohort studies and the validation of the harmonized variables should be thoroughly documented. Harmonisation processes should be transparent, rigorous and well documented in order to justify their validity.

Standardization and Comparability: We recommend to create ***transformation standards or links between different standards*** on the same measured variables whenever possible. We also advise to aim for as high level of detail as possible for harmonisation procedures.

Definitions and Validation of Variables: We recommend using ***equating procedures*** to account for the specificities of each cohort's data, alongside a ***subject-specific correlation coefficient***. We also advise to adapt harmonisation processes to the particular characteristics of the constructs concerned.

Data Access and Data Availability: We recommend using ***linked samples and linked sample descriptions***. We advise to first create harmonised variables and then to make searchable information and annotations on data available in cohorts according to study relevant categories.

Sustainability of Data Infrastructure: We recommend using ***federated infrastructure*** or mixed infrastructures types in combination with ***cloud interfaces***.

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. GLOSSARY

Algorithmic (data processing method): Harmonizes the same measure (categorical, continuous variables or both) with different but combinable ranges and categories.

Calibration (data processing method): Harmonizes to the same metric measure.

Central location (infrastructure type): Data from all studies are stored in the same server.

Central & Local location (infrastructure type): Some studies share their datasets to be stored in the same server while other studies store their datasets in their local server.

Data analysis types (integrative methods): Cf. Meta-analysis, Pooled Analysis and Federated Analysis.

Data processing methods: Cf. Algorithmic, Calibration, Standardization, Latent variable model and Multiple Imputation.

Different locations (infrastructure type): Data from each study is stored in their local server. Each study imposes its data restrictions.

Ex-ante retrospective harmonisation: Studies use standard collection tools and standard operating procedures.

Ex-post retrospective harmonisation: Studies try to achieve commonality through data processing procedures.

FAIR: FAIR are principles for the scientific management and stewardship of data developed in 2016. FAIR specifies that data should be findable, accessible, interoperable and reusable (more on: *https://www.go-fair.org/*)

Federated Analysis: Centralized analysis with individual-level data remaining on their local servers.

GDPR: Refers to *General Data Protection Regulation* (Regulation (EU) 2016/679). The GDPR is a regulation developed by the European Commission, the European Council and the European Parliament to reinforce data protection of individuals living in the European Union. More on*: https://gdprinfo.eu/*

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

Harmonisation: Practices that improve the comparability of variables from separate studies and reduce study heterogeneity.

Infrastructure types: Cf. Central location, Different locations and Central & Different locations.

Integration: The act or process of combining the same data from different sources into one unified whole.

Latent Variable model (data processing method): Harmonizes the same constructs measured using different scales with no known calibration method but with bridging items present.

Meta-analysis: Combines the result of multiple studies addressing the same variable.

Multiple Imputation (data processing method): Harmonizes datasets (not variables) with the same set of variables using bridging variables.

Pooled Analysis: Analyses can be carried out at individual-level after pooling data.

Prospective Harmonisation: Studies share the same study design, questionnaires, and instruments for collecting biological, psychological and social measures

Prospective Cohorts: Include two types of cohorts: mature and contemporary. Mature cohorts involve extensive follow-up of several decades while contemporary cohorts include relatively recent exposure information.

Standardization (data processing method): Harmonizes the same constructs measured using different scales with no known calibration method or bridge items.

Vocabularies: Refer to taxonomically structured sets of parameters used for annotating samples. *Original vocabularies* are descriptors and terms used for annotating samples at the biobanks and collections. *Harmonised vocabularies* refer to common representation of several varieties of original sample descriptors.

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