SYNCHROS

WORK-PACKAGE 6

Developing a sustainable strategy for optimisation of cohort data in Europe and to define a strategic agenda for cohorts globally

D 6.2

Report on standards for improving future 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	Laura Fincias
Contributors	Ellen Vorstenbosch, Jerome Bickenbach
Reviewer	Anastassja Sialm
Responsible Partner	Parc Sanitari Sant Joan de Déu, PSSJD, Spain
Dissemination Level	Public
Nature	Report
Keywords	Standards, Cohort Data, Study Design, Data Collection, Data Sharing, Ethics, Legal, Common Data Models, Metadata, Data Protection, Data Storage, Data Structure, Data Transfer, Federated Analyses
Due Date of Deliverable	April 30th 2022
Actual Submission Date	April 30th 2022
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

SYNCHROS	SYNergies for Cohorts in Health: integrating the ROle of all Stakeholders
CDMs	Common Data Models
CDEs	Common Data Elements
FDA	Food and Drug Administration
SCCs	Standard Contractual Clauses
GDPR	General Data Protection Regulation
SD	Stakeholder Dialogues
DDI	Data Documentation Initiative
DCC	Data Curation Centre
EDPB	European Data Protection Board
CIOMS	Council for International Organizations of Medical Sciences

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

DOCUMENT INFORMATION 1
ACRONYMS AND ABBREVIATIONS 3
TABLE OF CONTENTS 4
1. BACKGROUND 5
2. GENERAL CHALLENGES IN COHORT RESEARCH 6
3. STUDY DESIGN 9
3.1. CONTINUITY, SCOPE AND GOVERNANCE OF CONSENT 9
3.2 AUTONOMY AND SOCIAL VALUE 13
3.3 DATA PROTECTION SAFEGUARDS 17
4. DATA COLLECTION 22
4.1 STANDARDISATION 22
4.2 DATA COLLECTION PROCESSES: OPPORTUNITIES AND CHALLENGES FROM EMERGING DATA COLLECTION TECHNOLOGIES 27
5. METADATA 30
6. DATA SHARING 36
6.1 INTEROPERABILITY 36
6.2 INTEROPERABLE INFRASTRUCTURES 37
6.3. ALIGNING PROCEDURES 39
6.4 EUROPEAN COORDINATION 41
6.4. RESEARCH COMMUNITY AND COMMITMENT 43
8. REFERENCES 45

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. BACKGROUND

The main objective of the SYNCHROS Project is to establish a sustainable European strategy for the development of the next generation of integrated cohort studies thereby addressing the practical, methodological, ethical and legal challenges to optimizing the exploitation of current and future cohort data. Through the course of the project, SYNCHROS:

1) assessed the current-state-of-the-art concerning cohort data access, sharing, harmonisation and integration, thereby focusing on the methodological, practical, ethical and legal implications as well as potential issues related to the use of new data collection technologies;
2) defined the way-forward regarding implementation of potential solutions to overcome the identified obstacles, thereby presenting windows-of-opportunities and barriers; and
3) will formulate a robust, stable and sustainable strategic agenda for optimising the benefits and exploitation of health-related cohort data [due May 2022].

A broad plethora of stakeholders was involved in SYNCHROS and they played a key role in the outcomes of the SYNCHROS project. Concretely, they were involved in two main phases: 1) the evidence-based synthesis and priority setting, and 2) the evidence-informed policy-making. During the first phase, the evidence-based synthesis and priority setting, SYNCHROS organised 3 days of *stakeholder consultations* with PIs of cohort studies or cohort harmonisation initiatives, biostatisticians, methodological experts, ELSI representatives, etc. The aim of the stakeholder consultations was to corroborate and complete the challenges previously identified through literature studies. The outcomes of the stakeholder consultations have been translated into two strategy briefs (i.e. D2.5 Strategy brief on harmonisation and integration methods, and D3.3 Strategy brief on practical, legal and ethical issues in the optimisation of cohort data in Europe), describing the main challenges in the context of cohort data harmonisation from a methodological and ethical/legal point of view, respectively.

During the second phase, evidence-informed policy-making, stakeholders were invited to discuss opportunities and barriers regarding implementation. By means of *stakeholder dialogues*, representatives of funding agencies and EU policy-makers were invited to reflect upon potential solutions for the most important methodological and ethical/legal challenges (i.e. *Data Interoperability between Cohorts and Infrastructures* and *Continuity and scope of consent, Preserving confidentiality and Benefit for patients and society: the domain of justice*, respectively). The stakeholder involvement as well as the outcomes obtained through their involvement have been described in a previous SYNCHROS Deliverables (D5.1 Report on Stakeholder Meetings and D6.1 Report on Stakeholder Outcomes).

In this document, we will direction to overcome the main challenges in the context of cohort research. For the sake of the readability of this report, we categorize the challenges in accordance with three stages of cohort research: study design, data collection and data sharing (including harmonisation and integration). Per phase, we will present indications, solutions or

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

successful initiatives that can help to improve future strategies for cohort research. With this document, SYNCHROS aims to set a standard for an improved coordination and harmonisation of future cohort research.

The sources of information to report the standards to improve the future cohort studies include scientific literature from different fields such as health research, ethics, legality, cohort studies and the following SYNCHROS Deliverables:

- WP2: D2.4 (Inventory of infrastructures providing support for integration of cohort data); and D2.5 (Strategy brief on harmonisation and integration methods, and analytic approaches to maximise the value of cohort data);
- WP3: D3.1 (Report on practical, legal and ethical issue scoping exercise), D3.2 (Report on ethical guideline proposals) and D3.3 (Strategy brief on practical, legal and ethical issues in the optimisation of cohort data in Europe);
- WP5: D5.1 (Report on the two stakeholders meetings);
- WP6: D6.1 (Report on stakeholder outcomes)

2. GENERAL CHALLENGES IN COHORT RESEARCH

Although this document aims to provide standards and solutions for future cohort studies, some challenges are inherent to (international) cohort research and, (for the moment) do not come with an outstanding solution. Nevertheless, these issues may directly and indirectly affect the design and data collection of a cohort study as well as the possibilities for cohort data sharing, and therewith determine the opportunities for future cohort harmonisation and integration. Therefore, worth to consider here are (a) involvement of a wide array of stakeholders, (b) type of study design, (c) scientific and technological progress, (d) replicability, (e) loss to follow-up, (f) transnational ethical dilemmas, and (g) multiple jurisdictions.

(a) Involvement of a wide array of stakeholders

To start, large and enduring cohort studies are not easy to establish. Most large-scale cohort studies involve a consortium of researchers from different countries, which means variations in jurisdictions, policies, expectations about cohort data collection, as well as different cultural aspects. Additionally, they may require participation of many disciplines (e.g. researchers, physicians, data managers, legal officers, biostatisticians, and representatives of ethical committees), resulting in a multidisciplinary team with **many stakeholders who belong to different (international) organisations and areas of responsibility**. To ensure effective collaboration among members, a well-organized network is required (Kayaba, 2013). Moreover, large-scale cohort studies need **complex logistical infrastructures**, which requires extensive financial **resources** and **investment.** The latter can generate **different expectations** between researchers and for instance, funding agencies. Funding agencies might be expecting yearly study results whereas the valuable outcomes of a cohort study will only become available after

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

many years (Kayaba, 2013). Cohort data need significant time to mature before providing crucial research insights and their potential research value is only fully visible over time. In addition, in health research, it should be considered that in order for a disease to occur time has to pass. Therefore, to investigate a relation between timing and occurrence (White et al., 1998), **temporality** is needed. This is an implicit aspect in the design of a cohort study.

(b) Type of study design

Naturally, each **type of study design** has its own challenges. For instance, a randomized clinical trial would not be indicated in some cases, as it would be unethical to randomize participants by the exposure of interest (e.g. to study the risk of smoke tobacco in development myocardial infarction). A prospective cohort design is more time-consuming and costly than retrospective cohort design, as the follow-up of study participants in a cohort study is a major challenge and the failure to collect outcome data from all members of the cohort will affect the validity of the study results. On the other hand, a retrospective design is more susceptible to biases: the exposure has not been randomly allocated, and the associations found with the outcome between exposed and not-exposed could be explained by confounders, overestimating or underestimating the association (Euser, 2009).

(c) Scientific and technological progress

Another inherent aspect of cohort research is that due to **scientific and technological progress**, new methodological tools or measurements become available; hence, researchers must decide between consistent measures over time (i.e. waves, evaluations) and the quality of measures. Although it is said that *"if you want to measure change, do not change the measure"* (Otis Dudley Duncan), in practice this might have serious consequences for the quality of data. For example, if a cohort persists in using the same technology (e.g. array-based gene expression instead of RNAseq) throughout different data collections over the years, the data will lose value compared to data from more recent cohorts using newer technologies. In such situations, specific methods must be developed to harmonise data obtained from different technologies (Rudy et al 2011; Borisov et al 2017).

(d) Replicability

Also the concept of **replicability** (i.e. obtaining consistent results in different studies that pursue the same research question, but with different data) might be challenging. Related to the longitudinal character, replicated studies across years could be affected by added confounding variables (contextual, cultural, society, policy-makers…) arising during time. Hence, these could interfere with the results of different replicated studies conducted in different years. Likewise, the abovementioned scientific and technological progress may influence the replicability of cohort studies.

(e) Loss to follow-up

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

Habitual in cohort studies is also the **loss to follow-up**. Loss to follow-up may occur for many reasons: death, loss of engagement, loss of contact, loss of motivation to keep going on in the study, rejection of consent, etc., and can seriously affect the cohort's sample size. Small sample sizes can make it is difficult to draw robust conclusions as spurious associations could be found. On the other hand, a large sample size is a costly endeavour due to long-term commitment. It can be argued that the use incentives for participation could increase the sample size or overcome the loss of engagement, but paying participants for their engagement is not free of ethical controversy:

1) it facilitates that inequalities of foundation and resources across studies emerge;
2) payment could affect the quality of informed consent and a commercial exploitation of participants; hence, value, method and duration of payment should not be coercive neither inductive (Ashcroft, 2001); and
3) payment in relation to the assumption of risk for participants, could be in conflict with the ethical principle of non-maleficence (i.e., there are differences between participating in a study that re-tests a well-studied medicine than in a Phase I study with a novel pill)(Menikoff,2001).

(f) Transnational ethical dilemmas

Although not enforceable, ethical challenges can also be an obstacle during the design of cohort studies. Unchallengeable, the ethical domain is paramount in all research involving people, and is the first step to be taken into account at the beginning of studies. It is what governs the initiation of all research per se and if this domain is not covered and justified, then the study cannot proceed. The challenge during the design phase may lay in **differences in ethical considerations** between countries. For instance, in certain cultures, women do not have the autonomy to consent on behalf of their children or even on their own behalf, without the authority of the spouse. Another example is the differences in minimum age for being entitled to sign an informed consent. As such, these differences in ethical considerations can lead to transnational ethical dilemmas and may have repercussions on the design and data collection of international cohort studies.

(g) Multiple jurisdictions

Cohort research is rarely restricted to a single jurisdiction and in cohort studies across Europe, there are significant tensions and outright contradictions between **national legal frameworks** - most of which are themselves bound by the international and European conventions and guidelines, such as the EU Clinical Regulation nº 536 / 2014, and the Council of Europe Oviedo Convention and the 2004 Universal Declaration on Bioethics and Human Rights (Salokannel et al., 2019). When research is international, there is the inevitable problem of the interplay and potential contradictions between the **legal rules governing consent and confidentiality** (Townend, 2018). The lack of consistency between legal standards for protection of confidentiality and the terms of informed consent for health research generally across the EU countries; and the fact that the potential solution to this diversity, the General Data Protection

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

Regulation (GDPR) is an extraordinarily complex and relatively new regulatory document, may continue to not resolve the issues.

3. STUDY DESIGN

3.1. CONTINUITY, SCOPE AND GOVERNANCE OF CONSENT

Informed consent is generally required for the duration of a given research project and has to be constantly renewed and updated in order to inform participants of changes in the purpose, data sharing and data uses of the research. For cohort data studies, the main challenges are to adequately inform participants, the nature and temporal scope of trust and the governance of control.

At the design stage, the continuity scope and governance of consent are determined by seven factors namely (a) information required for consent, (b) the level of community engagement, (c) broad consent arrangements, (d) the nature and the temporal scope of trust, (e) motivations sources for consent, (f) governance frameworks and control mechanisms, and (g) GDPR related safeguards.

(a) Information required for consent

The **kind and extent of information necessary to adequately inform participants** for the purposes of autonomous consent is often considered a major hurdle. The highly technical and complex information of many cohort studies and the specific (medical) knowledge involved, makes it difficult to determine what should be communicated to participants and how. Attempts to educate participants and to engage them in the research process are increasingly common but they do not fully solve the difference in knowledge between the researchers and the participants mainly because models of participant engagement remain under-developed. Notwithstanding, researchers have access to established strategies – e.g. decision aids, workshops and training sessions – to ensure that participants fully understand the content of consent (Shaban-Nejad et al. 2018; Price & Cohen, 2019; Miao et al., 2020). Of note, only information directly relevant to support the decision to participate is required; the participant does not need to become an expert in the background science. Therefore, it is not plausible that, with patience and time, comprehensible information about the nature of the research, its risks and benefits cannot be communicated successfully. Techniques of 'supported decision-making' in which more active counselling and education is used can also overcome this informational inequality.

Moreover, the emerging use of 'big data' data collection tools, such as wearable devices prompts the question of whether participants can be adequately informed at all. Namely, in studies using digital data collection devices, it is the participants who are creating the 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

Paradoxically, this makes it increasingly unlikely that a participant can be informed of the extent of the data that will be collected (Dobrick, 2018) or the future uses of their data (Ahmad, 2019). This is because data intensive research (or big data) relies on real-time collection of large volumes of data from different sources (Hunter et al., 2018). A large amount of data is collected over extended periods with no clarity of what should be tested exactly (Firchow & Mac Ginty, 2020), so that, there is no way to foresee how a research project could be "fit for purpose" because the research purpose is often unclear during data collection (Marelli et al., 2020).

(b) Community Engagement for Consent

Another common criticism in the consent area is that community engagement remains hard to implement. Due to the longitudinal nature of cohort studies and the potential of integrating data with other cohort studies, obtaining informed consent raises the key question of how a participant can **meaningfully consent to the use of data that are unforeseeable at the time of consent**. Future uses of data may not be known, or knowable, not just by the participant but by the researchers neither. If the participant is informed of this possibility, he or she may withhold consent, which creates an obstacle to research. To address this problem, some have argued that cohort studies in principle create the ethical requirement that informed consent must be continuously renewed and updated in order to give participants the opportunity to respond to changes in the course of the research caused by unforeseen events, and in particular the perceived advantage of future data sharing or new uses of the data that the participant initially consented to (McMahon & Denaxas, 2019; Borry et al., 2018).

(c) Broad Consent Arrangements

Others have mentioned that a **broad consent** is generally the best option for cohort studies, at least in terms of ensuring the scope and durability of consent. The ethical acceptability of broad consent (sometimes called 'prospective consent') has recently been incorporated into the 2018 US Department of Health and Human Services criteria for Institutional Research Board approvals of research (Code of Federal Regulations: Part 46 – Protection of Human Subjects)[1]. Broad consent requires the researcher to provide participants with:

- a description of the types of secondary research that may be conducted;
- a description of the private information that might be used in research, whether sharing of the information might occur, and the types of institutions or researchers that might conduct research with the information;
- information on how long the information will be stored, maintained, and used;
- a statement that the participant will or will not be informed of the details of any subsequent research and research results.

Although broad consent is generally the best option in the cohort studies domain, it is still associated with many uncertainties related to future data re-uses. Of note, broad consent does

[1] https://www.ecfr.gov/on/2018-07-19/title-45/subtitle-A/subchapter-A/part-46#se45.1.46_1111.

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

not necessarily reflect the preferences of data subjects and cannot fully guarantee the protection of data subjects' rights. A common concern is that once the consent has been given, the **control over data is diluted as data are shared** between institutions. Ensuring data controlled access arrangements at the research institutions level, ensures that control over data is not lost but rather, evenly redistributed to those institutions and infrastructures best able to preserve it. It is thus necessary to ensure that the control over data does not disappear but rather, that it is equally distributed to research institutions that participants can trust.

(d) Nature and temporal scope of Trust:

The main issue is both the **nature and the temporal scope of trust**. There is always a risk potential for data misuse after consent has been granted. The general public is well aware of this aspect and knows that commercial institutions (such as pharmaceutical companies) may not necessarily have their best interests in mind. Therefore, it is necessary to be as open and specific about research purpose and future data uses as possible in order to generate trust. Indeed, participants do not cede their autonomy per se. Rather, they make an autonomous decision of granting some control over their data to the researcher and the governance structure the study is part of.

When giving consent, participants need to believe that the representatives of the research institutions have integrity and benevolent motivations. In that case, patients or participants are likely to trust their institutions (and their alleged respect for human dignity and democratic values) and thus give their consent for the use of their personal data. However, this does not mean that they will trust the same institutions a decade later. There is no guarantee that research institutions won't turn rogue later on.

Hence, **trust is not warranted by time.** Future benevolent motivations or interests of governments, pharmaceutical, etc., are not guaranteed for the future, neither are changes on individual and collective principles of the subjects. The issue is not trivial because consent in current cohort research implies a lifelong donation of personal and potentially sensitive data. A solution would be to create a platform where participants have the **possibility to opt-out** from studies they consider ethically objectionable (e.g. data are re-used for purposes such as discriminatory insurance schemes or employment discrimination).

(e) Motivations for consent

Data subjects' decision to give their consent is not value-free. That is, data subjects have their own motivations for granting control over their personal data. **Dynamic consent** takes this into account and offers data subjects the option to refuse data use for specific studies. Although dynamic consent is technologically assisted by consent apps that monitor consent continuity (Bilkey et al., 2019, Manzoni et al., 2018; Bialke et al., 2018; Budin-Ljøsne, 2017), in practice there are substantial practical problems of returning to the original participants in a study to renew consent.

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

More importantly, the decisions to refuse data use are generally made because the study's subject matter relates to sensitive topics such as mental health or ethnicity. However, in some cases, the motivations underlying decisions for not granting permission for data use are ethically questionable. For instance, a data subject may refuse to give his data to a research project because of the lead researcher's ethnicity. These underlying reasons are not identifiable and thus, ethically objectionable biases can shape data use for some studies at the expense of others. Dynamic consent is therefore, only recommended in limited cases.

A way of controlling for the abovementioned underlying biases is to use a **meta-consent** model where the moral status reflects moral preferences. In practice, this means that choices for giving or not giving consent to data use is structured according to a pre-defined set of options. While adopting a meta-consent model is a viable solution, it should be noted that data subjects' attitudes to data sharing and re-use are not known. Studies on the subject do not measure knowledge, attitudes or practices related to data sharing and re-use in comparable ways. Hence, it is impossible to make any inference about data subjects' preferences for data sharing across contexts, diseases and populations. To a large extent, data subjects' attitudes for data use are still a black box.

(f) Governance frameworks and control mechanisms

Nevertheless, an adequate governance structure is key to ensure patients' and participants' trust. Determining how this governance framework looks like depends on the distribution of responsibilities: it cannot be simply that patients give up their autonomy, while the consequences for doing so remain unclear. The broad consent option, for example, would require an **adequate governance framework with regular control mechanisms** over time. Governance structures in this context should install checks and barriers to data misuse. They may include access committees and patients' representations but in general terms, it is necessary to classify datasets types in terms of whether they require broad consent and associated governance structures or not. Some data and datasets (such as data with a high social value, or aggregated and anonymised data) do not require broad consent and use a different legal basis.

Also important to consider from the start, is that an appropriate governance framework for consent should include **(i) tools and pathways for data sharing and (ii) controlled data access arrangements** (cf. 3.3 Data Protection Safeguards). For instance, when a hospital receives data (and consent) from patients, it has to ensure that this data will be used only for legitimate purposes and by legitimate users. Doctors have to provide controlled access arrangements that monitor access applications by data users. Such an approach relieves infrastructures from administrative burdens and reinforces future data-sharing safety. The only purpose of infrastructures in this context is to maintain data safety by minimizing risks (e.g. putting restrictions on data downloads).

(g) GDPR-related safeguards for consent

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 way to implement safeguards compliant with the GDPR is to ensure that informed consent or/and ethical approval covers all potential data uses (Kirwan et al., 2021; Ducato, 2020). The GDPR offers detailed specifications on the subject in Article 6.1, Recital 62 and Recital 33 on the obligation to inform study participants. Namely, personal data use can be justified on two legal grounds namely (i) informed consent (with a subsequent approval from ethical review boards) and (ii) approval from ethical review board based on the designation of the research project as being in the area of public interest. Recital 62 is of particular interest for retrospective (cohort) research where data subjects can no longer be contacted for a renewed consent: there is no obligation to inform data subjects if these subjects already possess study related information and if the provision of information to study participants requires a disproportionate effort (Ducato, 2020).

For cohort research, the assumptions of the GDPR regarding the right of information and broad consent focuses around a core distinction between research projects whose purpose are general (e.g. rare disease research) and research projects where the designs and the level of processing may involve possible risks and privacy breaches for the data subjects (Hansson, 2021). For the former, a general description of the research purpose is sufficient while for the later, descriptions of the research purposes have to include information about the data controller, the nature of research, the parameters of data sharing (e.g. whether data will be shared across borders), partnerships frameworks (e.g. with commercial partners), the implementation of results feedback (including incidental findings), measures for data protection about unauthorized use and potential linkages to registry data (Stommel and Rijk, 2021).

Finally, the new Data Governance Act, within the framework of a European Strategy for Data aims to introduce a uniform "European data altruism consent form" for altruistic data re-use to allow the use of their non-personal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services." (Shabani, 2021).

3.2 AUTONOMY AND SOCIAL VALUE

In health research, the principle of autonomy supports the requirement that people who participate in research freely consent to do so in light of sufficient information to make an informed decision, and that information derived from their participation is confidential. Respecting autonomy means that researchers must ensure that participants' choice over matters concerning themselves is respected. However, the value of autonomy (i.e. individual choice) is not absolute and must be balanced against the potential value of the study (both for the participant and for society as a whole). In order to achieve such a balance, it is necessary to take in account three research dynamics namely (a) a progressive reduction of autonomy through digital data collection technologies (b) the tensions between individual autonomy and social

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

justice and (c) the identification of the social value of research.

(a) Reduction of Autonomy in cohort research: the impact of digital data collection technologies

The **dynamics of data intensive and cohort research do not favour participants' autonomous choices**. In studies using digital data collection technologies for instance, it is not always easy to determine what participants are exactly consenting to (Kitchin, 2020). Many participants do not read consent forms and consent notifications related to certain apps properly, and thus are willing to give away personal information and agree to privacy settings (Ogunseye et al., 2021; Aiello et al., 2020). While such behaviours are primarily motivated by the preference to use apps and digital services, they also point towards a lack of information and education about how confidentiality breaches operate in the post-digital world (Marelli et al., 2020).

In any case, the **ability of participants to give an autonomous and fully informed consent is significantly reduced in digital communication settings**. In the world of new technology, informed consent is simply an unrealistic requirement that cannot be implemented in practice (Shaban-Nejad et al., 2018). For geolocation technologies (e.g. GPS located or wearable sensors), for instance, consent can never cover the full extent of possible data uses. Namely, the data collected can include personal information that was outside the remit of the initial purpose (e.g. information related to personal habits). There is much opacity in this context (e.g. proprietary regimes, inaccessibility of algorithmic codes, lack of readability of machine learning outputs) but the impact is clear: participants are alienated from the processing, analysis and collection of their own data. In order to circumvent such limitations to autonomy, research relying on digital technologies uses a differentiation between volunteered (VGI) and contributed information (CGI) in order to clarify the extent to which data subjects can control the parameters of their contributions.

(b) Individual autonomy vs. Social value of justice

In the context of health-related cohort studies, researchers must consider the balancing exercise to ensure that for the individual the potential benefit from participating in research or the social value outweighs the potential harm. A research that lacks social value lacks ethical rustication as well, for the simple reason that, in light of scarce resources, socially valueless research is a waste of resources. This consideration does not involve the **individual values of autonomy and privacy, rather it involves the social value of justice**. This ethical consideration is clearly reflected in international ethical conventions and guidelines. For example, the opening guideline of the 'International Ethical Guidelines for Health-related Research Involving Humans' states that "The ethical justification for undertaking health-related research involving humans is its scientific and social value: the prospect of generating the knowledge and the means necessary to protect and promote people's health." (CIOMS, 2017: Guideline 1).

Individuals can fully participate in group decisions to restrict autonomy (e.g. by means of democratic processes that are open to all and fully transparent), then the individual has in

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

effect consented to having his or her autonomy limited. A stronger, and more contentious, version of this argument is that, under these conditions of full participation, that **if research has a good chance of producing genuine value for everyone, then the individual participant has a duty to consent**. Some points to justify the loss of individual autonomy in favour of social value in cohort studies are:

- Autonomy must always be linked to the public good, and while participants may plausibly argue that they should be fully autonomous over some highly personal information, that does not mean they are over all information (Vergallo et al., 2020; McLennan et al., 2019; Richterich, 2018)
- Unconditional autonomy can only be justified by an implausible application of methodological individualism that in practice can lead to health inequalities and stigmatization (Dove & Garattini, 2018).
- When participants, because of intellectual impairments or mental health problems, do not have the full decision-making capacity, autonomy is limited and, as long as the research promises to produce social value and the participant is not harmed, consent is being presumed (Wolf, 2020).
- Limiting the impact of autonomy in the case of cohort research makes sense because participants are not only limited in their capacity to understand scientific information, they do not necessarily have an adequate idea of what constitutes their best interests, or to evaluate the risks and benefits of their participation (Cech, 2018). They are thus not necessary qualified to evaluate the *risks and benefits* of their participation. They may over-estimate the likelihood of potential benefits to research and underestimate the risks of research procedures.
- A related issue is the therapeutic misconception problem where participants accept to take part in clinical research because they expect to receive the same individually focused treatment that they would receive in a non-research clinical context (Lidz, 2002). The question therefore, is whether researchers should specify that the research is calibrated for gaining generalizable knowledge rather than for patients' benefits.

(c) Identifying the social value of research

To determine social value of research is challenging and a potential ethical obstacle to the value of cohort data. Evaluating the social value of research is fraught with controversy, and skepticism about whether, and how, social value can be ascertained. The best option would be to use appropriate blind peer-reviewing by relevant and conflict-free scientific experts to make a decision to approve or fund research solely on scientific merit. **Scientific consensus becomes a proxy indicator of social value**. There are many examples of research that did not initially appear to have any further social aim, but later did. By the view that social value is often unknowable, the best we can do is rely on scientific consensus utilizing objective criteria of scientific soundness.

With respect to the second option, it is relevant to mention that most middle- and high-resource countries rely on research councils, institutes or funding agencies that seek to evaluate

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

both the scientific and social value of research, especially in the health area in which the potential for direct implementation at the clinical level, or even commercialization, is often a requirement of funding. The idea is that such **agencies describe what is in the public interest**. For this reason, participatory involvement and transparency are pre-requisites: hidden agendas or special interests do not reflect the public interest or identify public goods. Also, it is important to take into account that determination of the social value of research inevitably involves reconciling and negotiating relevant interests of individuals and groups that are often in conflict (Richards et al., 2015).

Community engagement is important and in order to be effective needs to be encouraged at the outset, for example when research funding objectives are being developed, so that the public is not brought in after the research is a fait accompli. However, it would be naïve to assume that the general public would have the ability to be able to fully understand the scientific background or have the specialized knowledge to be able competently evaluate all examples of health research (Wiggins and Wilbanks, 2019). Equally problematic is the phenomenon of patient advocacy groups whose interests in promoting research to benefit specific groups may complicate agreement on the ultimate social value of health research (Umbach et al., 2020). The feasibility of this type of agency in the health research field very much depends on several open questions, perhaps the most important of which is the extent to which **active participation and ethical engagement are possible for a wide range of stakeholders** with potentially conflicting agendas (Aiello et al., 2020; Wiggins & Wilbanks, 2019). Moreover, it is critical that the agency has the public perception of neutrality, fairness and respect for different interests.

To maximize patients' autonomy another solution is to give patients the possibility to approve access for each of the data parts provided. For instance, cooperatives as MIDATA and SALUD COOP facilitate citizens' control and cooperation by providing them access to manage the use of their own data, therewith **facilitating a certain sovereignty** over their own data (Shabani, 2021).

Finally, health research and thus, health-related cohort studies should also consider social ethics in which the underlying social value is fair for the population as a whole, future generations or the biosphere or environment at large. This means that, both **procedural justice** and **distributive justice** must be taken into account. Procedural justice involves the fairness of processes and protocols underlying decisions that may have consequences for individuals. If an individual is prohibited from participating in research because of a consideration that in the context of the research is scientifically and ethically irrelevant (gender, race, religious belief), then the sampling frame is unjust as it is discriminatory. Distributive justice, as the name implies, means in the context of research, that the benefits and burdens of research are fairly distributed across the population. So specific regulations on international level have to be defined in order to ensure that research does not always benefit the same groups, and that achieve funding do not be influence on which collectives are benefit of the research (e.g., there is generally less funding for malaria research). A **holistic approach is required to quantify and asses social value** examining: burdens, benefits and harms, policy and impact levels.

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

3.3 DATA PROTECTION SAFEGUARDS

The GDPR was intended to harmonise data privacy laws across Europe and specify the legitimate modalities of data collection, storage, sharing and use for categories of data that are potentially "private" (Salokannel, 2019). In general terms, the GDPR "strengthens individual control of data subjects over their data in this digital age" especially in light of growing scepticism about the reliability of anonymisation as a technique for protecting confidentiality (van Veen, 2018).

The GDPR aims to achieve "privacy by design" where safeguards must be built in the research infrastructure from the onset of research (Pagallo, 2021). The GDPR (Article 89) also makes clear that research in the public interest should include considerations on what "necessity" and "proportionality" mean in a research context with personal data (Guinchard, 2018). This means, that the collection and storage for personal data are lawful as long as there is a necessity to retain data for public interest, or scientific and statistical purposes (Slokenberga et al., 2021). In the context of cohort research, to determine what is necessary and proportionate might be difficult because of the integration of data from multiple data sources, which can potentially lead to non-compliance with the GDPR (while national bodies are respected and vice versa). Hence, considerations about necessity and proportionality are determined by the possibility to apply **appropriate safeguards and derogations for the protection of data subjects' rights** (including confidentiality and autonomy) (Duguet et al., 2021; Verhennman et al., 2020).

(a) Focus on personal data: the impossibility of full anonymization

It should be noted that the GDPR is more favourable to cohort research at the design and consent stages than previous directives on data protection (Duguet et al., 2021). For instance, Recital 33 opens the possibility of broad consent arguing, "it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection". Data subjects can thus, give their consent according to the intended purpose of the research. This does not relieve researchers from the obligation to inform subjects and from providing a detailed description of the research purpose (cf. Recital 33). This description can be more general if the purposes for personal data processing cannot be fully outlined from the onset (Peloquin et al., 2021).

But how is this relatively favourable arrangement for consent translated into the design domain, especially in terms of confidentiality and privacy issues? The processing of personal data should be made in a way that would ensure its confidentiality and security (General Data Protection Regulation, Recital 39, 2016). Such an approach means that **in terms of privacy and confidentiality, the GDPR applies to only personal data rather than to fully anonymized data**. According to its principles (i.e. Purpose limitation; Accuracy; Storage limitation; Integrity and confidentiality; and Data minimisation), the GDRP requires that data should be fully anonymised when it no longer serves any scientific or statistical purposes. This leads to ambiguous

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

conceptualizations as to what anonymized data really means in legal terms. For instance, it is often claimed that pseudonymized data cannot be converted into an anonymized version because it is linked to an individual citizen subjected to his/her legal regulations. Other perspectives state that anonymized versions are possible in certain circumstances. Regardless of the position adopted, the legal basis for the conversion of pseudonymized data into anonymized versions remains unclear. This is because de-identification is not irreversible. Namely, while pseudonymised data is not directly relatable to individual data subjects, additional information can still potentially lead to their re-identification. Indeed, *"there are some studies, for instance, rare diseases studies, where it is almost impossible to completely anonymize data , because it could be personal data obtained from one institution"* [Stakeholder].

In the context of cohort studies, anonymization is more a declarative commitment towards data security than a viable practical solution. The longitudinal aspects of cohort research and some research fields (i.e. genetical studies), makes full anonymization difficult to implement, especially as it degrades research value. Researchers are often confronted with the impracticability of anonymization per se. For biological data, for instance, anonymization and de-identification are impossible because it is related to genetic analysis. In such circumstances, it is not possible to have fully de-identified data without destroying its scientific value. First, full anonymization techniques use algorithms that make data unusable. Second, *"Some full anonymization techniques are likely to make data useless altogether"* [Stakeholder]. The value of cohort data is often dependent on the amount of location and temporal information (such as dates). Suppressing such kind of information in a longitudinal context is impossible unless one is willing to sacrifice the scientific value and impact of the cohort study. As a general rule of thumb, the level of details in data is inversely proportional to its affordability for anonymization.

(b) Pseudonymization as safeguarding measure: issues in identifying personal data

Pseudonymization is thus the preferred solution to respect confidentiality and privacy rights according to the GDPR. Pseudonymised data can still identify an individual but only through a key that is safely kept by the data controller. Pseudonymization is a measure for safeguarding privacy and confidentiality and hence ensures that personal data would be processed only for a specific purpose and for a specific research question. This linearity of processing personal data is an expression of the GDPR's proportionality principle: Article 89 suggests that research interest and research objectives should be balanced against data subjects' interests and the right for personal data protection.

However, it remains **unclear whether so –called "pseudonymised" data is personal or not** (van Veen, 2018). If data is personal data under the terms of the GDPR, the collection data, storage, sharing and use require informed consent that is "freely given, specific, informed and unambiguous by a clear and affirmative act"[2]. In this direction, focusing on the risk of re-

[2]Regulation (EU) 2016/679 of the European Parlament and the Council. Article 4(11). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

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

identification, some researchers suggest that informed consent may not be a sufficient protection of the rights of individuals, when for example, the data controller is a public authority with sufficient power to convince the data subject that he or she would suffer some detriment were they to refuse[3]. Related to data sharing, namely under the GDPR, each participant has to be informed individually when their data is retroactively transferred to databanks if the controller intends to further process the personal data for a purpose other than that for which the personal data were collected (Article 13 (3)).

Re-identification, furthermore, is an inherent risk in cohort research where the increasing dimensionality of data, the availability of open-access databases and the extensive sharing practices of researchers make it hard to avoid. Re-identification naturally implies a privacy breach, but it also highlights the confusion between privacy and data protection. While they are often used interchangeably in the literature, privacy is a human right while data protection (and by extension the right to privacy) is a legal provision. In this sense, data protection is a "negative freedom" in the sense that it sets to limit the behaviour of others and thus directly concerns activities that lead to de-identification. By contrast, privacy is a prerogative that enables the behaviour. Re-identification in this sense, has little to do with privacy rights per se: instead, it is the product of interactions between what the data affords (some data are more prone to de-identification than others) and what the negative freedom of data protection forbids.

Identifying what is or is not personal data may be quite difficult in cohort research. In the context of digital data collection technologies boundaries between healthcare and social domains are increasingly blurred (Marelli et al., 2020). Moreover, social media data require more triangulation and ontology-based strategies before being pseudonymized, stored and harmonized for subsequent sharing (Correia et al., 2020). There is thus considerable ambiguity as to which information should/could be pseudonymised and encoded in practice. Some datasets structures are more affordable and "prone" to pseudonymization than others (e.g. this seems to be the case for wearable data) (Muzny et al., 2020).

(c) Non-EU privacy considerations

Different countries across the world take another approach to privacy than the EU. For example, the USA use a sectoral approach that relies on a mix of legislation, regulation, and self-regulation [4]. Data protection safeguards can be used as a ground for data transfers from the EU to third countries (i.e. outside the EU/EEA) by **Standard Contractual Clauses** (SCCs) according to GDPR. Nevertheless, some third public institutions are legally barred from agreeing to some clauses on liability, jurisdiction and governing law provisions contained in the SCCs. As a result, EDPB approved more flexible alternative quasi-judicial mechanisms (e.g. mediation, independent review, commitment to be liable for compensation of damages following such review) (Mitchell et al., 2020).

[3] https://www.eu-stands4pm.eu/lw_resource/datapool/systemfiles/cbox/331/live/lw_datei/d3-1_v1_sep2020_compact_public.pdf

[4] https://www.privacyshield.gov/article?id=OVERVIEW

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

To ensure effective safeguards and personal data protection when data is transferred to non-EU countries to US, the Privacy Shield Framework Principles have been developed: Notice, Choice, Accountability for Onward Transfer, Security, Data Integrity and Purpose Limitation, Access and Recourse, Enforcement and Liability[5]. In that direction, Schrems II considers the validity of SSCs in the context of the EU-US Privacy Shield and surveillance programmes in the US. The level of protection the third country provides impacts upon the protection that any SSC within its jurisdiction provides (Mitchell et al., 2020). So, SchremsII now requires that European companies are to conduct individual assessments of each data transfer to a non-EU country in order to ensure compliance[6].

(d) Solutions

Pseudonymisation is thus a measure to ensure that personal data would be processed only if necessary for a specific purpose and in relation to a specific research question (Guinchard, 2018). This logic extends not only to the collection of personal data but also to its storage and its modalities of access (Hansson, 2021; Pagallo, 2021). Much health research and cohort research projects rely on pseudonymised data where a level of participants' traceability (e.g. genomic research) and re-contacting opportunities are needed (e.g. longitudinal studies). This means that organisational and technical measures are required in order to ensure that (pseudonymised) personal data cannot be related to identifiable persons (Molnár-Gábor and Korbel 2020). The **options for solving confidentiality issues** are governance structures, compute-to-data (e.g. federated analyses) and codes of conduct; although, the latter could be more considered a guideline than an actual safeguard.

A **governance structure** could balance and coordinate the individual concerns with the private interests in a transparent way. The allocation of responsibilities is implemented through various models of access such as controlled access and registered access for sharing data (cf. 6.3 Aligning procedures).

In order to solve the contradiction between various jurisdictions and enable the scaling of research, researchers may rely on "compute-to-data" methods where the data is not physically shared. Regardless of whether datasets are held by custodians or requested remotely, "compute-to-data" frameworks allow to combine individual level analysis of harmonised data from various EU cohorts (Nurmi et al., 2019). 'Compute-to-Data' is a technical means for exchanging data while preserving privacy by allowing the data to stay with the data controller (the individual or individual responsible for the generation, harmonization and storage of data) and allowing data consumers to run computation tasks on the data: rather than sending data to the algorithm, the algorithm runs where the data is. Here we should think of client-server architecture such as DataSHIELD for cohort studies

[5] https://www.commerce.gov/tags/eu-us-privacy-shield#
[6] https://dis-blog.thalesgroup.com/security/2021/04/29/what-is-schrems-ii-and-how-does-it-affect-your-data-protection-in-2021/

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

(Wilson et al., 2017) or for biobanks BioSHaRE-EU (Kaye et al., 2016). This is a useful solution in case the governance scheme in place prevents data release or forbids the combination of multiple datasets (Conley & Pocs, 2018). Federated networks, for instance, increasingly compensate the limitations of controlled governance models (Keane et al., 2021). Another, related possibility is that federated networks support DACs overview with automated review tools (Cabili et al., 2021). Federated networks assume that they solve partly the confidentiality issue as data can be shared in an anonymized and secured way. However, while "methodologically and technologically creating an absolutely secure environment is eminently doable, no environment is 'absolutely' bullet-proof" [Stakeholder].

It should be noted, that the technical progress of **big data analytics outpaces the changes in legal norms**. Law proceeds slower than advances in data collection technology which could facilitate a huge volume of data (i.e. big data). Besides, big data processing and sharing procedures are often dependent on ''outdated'' regular norms, to the extent of being oblivious to the ethical and social impact of their work (Minari et al., 2018; Molnar-Gabor and Korbel, 2020). For instance, database structures often focus on privacy and security rights at the expense of the long-term social impact of data results. In other words, while big data analytics use legal frameworks to implement a structurally global but a conceptually local understanding of science. The solution in this context, is to take soft law into account (and not only in the context of broad consent). **Soft law solutions** (i.e. quasi-legal measures) could be the solution on that dynamic field which needs immediate and flexible procedures instead of long-term legal provisions. It may help databanks and societies alike to adapt more quickly to new risks and benefits of cohort research.

Finally, another safeguard, if we refer to the GDPR, are **codes of conduct for joint research projects** (Commandé et al., 2021; Molnár-Gábor and Korbel., 2020). In this regard, the GDPR gives some indications on data protection (e.g. Recital 98 and 77) but these indications are too vague and general to be of practical use (Shabani et al., 2021). Recital 77 makes clear that such codes of conduct are closely related to the various specific contexts where personal data are used, which makes any conceptualization of general guidelines difficult (Stommel and Rijk, 2021). Nevertheless, individual initiatives may be a source of inspiration for the design of generalized guidelines. For instance, they generally include an interface coupled with a secure database that allows the collection, storage, real-time analysis, reuse and integration of linked data (e.g. RD-Connect) (Hansson, 2021). In such cases, codes of conduct mainly concern the terms and roles according to which users can access such platforms.

Summary and recommendations:

- The conditions and modalities of consent should be carefully considered and identified in relation to the participants. In particular, researchers should use established strategies such as decision aids to ensure that participants fully understand the content

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

of consent.

- Research participants do not understand the language of broad consent for future use. They tend to think that by accepting consent for future use, they will be informed on incidental findings and thus receive some benefits. Crucially, it is not the consent for future use that is important but rather the fact that research participants don't consider all future uses in the same way (e.g. they may be more concerned with commercial uses than for future uses).
- Broad consent is the most relevant option for cohort studies because it allows re-contacting the patient more easily and with fewer risks. Nevertheless, broad consent remains associated with uncertainties related to future data re-uses and this should be contemplated.
- Adequate governance structures should be developed to (i) offer participants the option to opt-out from studies they consider ethically objectionable (e.g. data are re-used for purposes such as discriminatory insurance schemes); (ii) to create regular control mechanisms to prevent data misuse.
- The lack of trust of the general public and participants can be partly explained by the lack of transparency of commercial and non-commercial institutions doing the research. The main task, therefore, is to counterbalance and identify the power structures and to attribute belief back to the institutions.
- In the context of data intensive cohort research (e.g. in digital communication settings), participants are alienated from the processing, analysis and collection of their own data; and hence, can impossibly consent in an informed and autonomous way. Participants, but also researchers, should be educated in the ethical and legal implications of new data collection technologies.
- There is no parameter for the social value of research and hence, we depend on scientific consensus or agencies defining public interest. Community engagement and active involvement of a wide array of stakeholders should be encouraged to maximise the public perception of neutrality, fairness and respect for different interests.
- The legal basis for the conversion of pseudonymized data into anonymized versions remains unclear; this is because de-identification is not irreversible in the context of cohort research. Hence, anonymization is more a declarative commitment towards data security than a viable practical solution.
- Options for solving confidentiality issues in cohort research are governance structures and compute-to-data (e.g. federated analyses).
- Soft law solutions (i.e. quasi-legal measures) could be the solution when technical progress of, for instance big data analytics, outpaces the changes in legal norms, and immediate and flexible procedures instead of long-term legal provisions are needed.

4. DATA COLLECTION

4.1 STANDARDISATION

Cohort research could profit from a certain level of standardisation, at least, when in a later stage data harmonisation and integration are the objective. With standardisation in this context

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

we do not necessarily refer to using common measures or common standards, which might be more problematic. Standardisation implies that constructs can be measured in standardised ways. However, for culturally specific and multifaceted constructs (e.g. depression), this assumption might be unviable and measuring these in standardised ways across cohort studies would jeopardize the quality of data. Moreover, there can be many different standards for the same measures, whereas the required level of granularity or precision depends on the importance of that information in a certain research context (Fortier, 2011). Hence, instead of focussing on the standardisation of measures or standards, a solution would be to focus on standardisation of the data and their representation[7]. This is possible by using (a) Common Data Models (CDMs) and (b) Common Data Elements (CDEs); however, (c) consensus finding will depend on the active engagement of the research community.

(a) Common Data Models (CDMs)

Common Data Models (CDMs) build and standardise different data content for a same data structure format. Structuring data in a more consistent way could:

- Organize and structure data collection from different formats, sources and measurements into a standard structure;
- Promote data sharing;
- Make data more comparable across studies;
- Facilitate the harmonisation and integration of data;
- Provide an opportunity for re-use of data;
- Reproducibility.

In terms of scope, four types of CDMs are proposed[8]:

1. **Protocol-based:** Source data extracted, transformed, and loaded to the CDM is limited to that required for a specific protocol or set of protocols. It is the subset of the source data required to answer a predefined set of study questions.
2. **Protocol independent:** Source data extracted, transformed, and loaded to the CDM is minimally a subset of the source data and maximally the entirety of the source data. In case of a subset, this subset is not limited according to data deemed relevant to a study question or set of study questions. Rather, it is potentially applicable to as yet undefined study questions.
3. **Syntactic harmonization:** Is the arranging of data elements into a common structure without altering their content or meaning. Source data is extracted, transformed, and loaded to a CDM harmonised in terms of structure across data sources. The content of the tables and columns of the data in the CDM remains in its original format and is therefore allowed to remain heterogeneous amongst data sources.

[7] https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:background
[8] https://www.imi-conception.eu/wp-content/uploads/2020/10/ConcePTION-D7.5-Report-on-existing-common-data-models-and-proposals-for-ConcePTION.pdf

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. **Semantic harmonization:** Is defined as of or relating to meaning in language, or the meaning or relationship of meanings of a sign (data element) or set of signs. Semantic harmonization is the derivation of common variables from the combination or restructuring of various data elements. Source data is extracted, transformed, and loaded to a common data which is harmonised in terms of structure and content across data sources. The content of the tables and columns of the data in the CDM must be mapped to a set of predefined concepts from a common vocabulary or set of vocabularies.

There are three important factors that guide the choice of a CDM for a healthcare database: **suitability** for the research question, **transparency** to allow reproducibility, assessment of **validity** and easy-use for speed and scalability (Schneeweiss , 2020). Several clinical networks have worked in building CDMs such as Clinical Data Interchange Standards Consortium (CDISC), The National Patient-Centered Clinical Research Network (PCORNet), Informatics for Integrating Biology & the Bedside (I2B2), The Observational Medical Outcomes Partnership (OMOP) and Sentinel Common Data Model (Sentinel CDM - FDA).

(b) Scope of Common Data Models

Of these, CDISC develops standards in collaboration with experts in pharmaceutical organisations and aims to improve standards for pharmaceutical organisations internationally and I2B2 allows to explore and query clinical data that has been de-identified and aggregated[9]. PCORNet is used to support pragmatic trials and observational research supporting recruitment, data collection, monitoring, and follow-up (e.g. of studies that use PCORNet CDM: ADAPTABLE, the PCORnet Obesity Observational Study, and the PCORNET Bariatric Study)[10]. Sentinel is mostly used for post-marketing drug safety surveillance but increasingly also for effectiveness research and OMOP CDM has been used to study treatment pathways, comparative effectiveness, safety, and patient-level prediction (Kent et al., 2021). For instance, the CNODES initiative has used Sentinel CDMs (Platt, 2019) because of its purpose to respond to urgent drug safety and effectiveness questions posed by our government partners. And, the EHDEN initiative currently has 140 data partners from 26 different countries which are mapping their data to the OMOP common data model. This includes several EHDEN project partners who have also mapped their data to the OMOP CDMs for use in a federated network[11].

Contextual changes over time do not affect the suitability of CDMs; hence, they are sensitive to the challenges of changes over time. For instance, due to the health emergency during the COVID-19 pandemic many studies with different research questions and designs were carried out across countries. Studies collected clinical data without any CMD standards specifically for COVID-19. An initiative run by the Reagan-Udall Foundation (for the FDA) in collaboration with Friends of Cancer Research, now aims to harmonise a list of identified and reviewed COVID-19

[9] https://ukhealthdata.org/wp-content/uploads/2021/12/211124-White-Paper-Recommendations-of-Data-Standards-v2-1.pdf
[10] https://www.healthaffairs.org/do/10.1377/forefront.20170606.060415/full/
[11] https://www.ehden.eu/datapartners/#

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

data elements (*see COVID-19 Mapping spreadsheet on the website :* https://www.fda.gov/drugs/coronavirus-covid-19-drugs/covid-19-real-world-data-rwd-data-elements-harmonization-project) with several CDMs (from those mentioned above) and open standards.

In addition, the use of CDMs allows to control and assess the quality of data (completeness, validity, accuracy, uniqueness, and consistency) and compare data quality across centres or studies (Kim, 2021), what also means a future improvement for data collection, data structure, data storage and data sharing. An active engagement and agreement of the research community is required to reach consensus on CDMs.

(e) Common data elements

Common language is also needed in order to identify health measurements, observations, documents and variables through the use of codes and names, which are enablers to create a common terminology to facilitate data structure and data sharing. Initiatives as *Loinc* (https://loinc.org/about/) or ATHENA OHSI have developed standard vocabularies in order to facilitate data sharing.

To validate and define variables for cohort studies CDEs are needed. Common and minimal datasets cannot be large. They are generally very small and are approved by the research institutions conducting the research. CDEs are possible for certain kinds of data, for instance, adverse events. Namely, there is a common way to record adverse event data even if specific adverse events are different in each study. That is, it is still possible to structure the data despite inter-study differences so that the data will be organised in a similar manner and therefore, be more comparable. For each adverse event, it is possible to choose a common data element that indicates the start date, the end date and the ongoing status of the event. This common data element becomes the main code for structuring each adverse event data. **The aim, therefore, is to obtain fixed data items and apply them to structure the data for further comparison**.

The more heterogeneity in data collection measures, the more complex the harmonisation exercise, if not impossible. For instance, in the ATHLOS project (Sanchez-Niubo, 2017) a harmonised dataset with over 341,000 individuals from 20 existing longitudinal studies of ageing was created through a rigorous process. The dataset contained harmonised variables of health status and functional limitations, lifestyles, social environment, psychological areas, among others. However, in the psychological domain, there are differences between the varieties of instruments in data collection across studies. For depression, the 20 cohort studies used GDS(2), CES-D(20 items-10 items)(10), DSM-V criteria(2), EURO-D(2) and The Composite International Diagnostic Interview (CIDI)(1) whereas, on the cognitive domain, the instruments used by the cohort studies were much more consistently: MMSE(5), MEC(1) and MOCA(1). In such cases, it is not possible to fix the data items themselves, it is only possible to harmonize data organisation, and rely on prefixes and suffixes to clearly relate data points to definitions.

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

Thus, "*except for a few instances where core fixed data are possible (e.g. demographics), each study will have its own data points with specific data items. An explanation for this can be found in the very nature of research. That is, there is likely to be no established standards for innovative questions*" [Stakeholder]. In such conditions, it is only possible to categorize, structure and define data points.

(d) Consensus finding: active engagement of the research community for minimal datasets

However, such activities can only be done through **active engagement and agreement of the research community**: a single researcher cannot fully know what should be a minimal dataset. Apart from the fact that it can be quite challenging to reach consensus among researchers, the **scope of *consensus finding* is also limited only to a few variables** because "each database has a very strong focus on certain data elements that are not redundant" [Stakeholder]. For instance, in prospective data collection for two brain injury studies, researchers could agree only on 13 of the 273 variables that were initially considered. The same phenomenon was observed in rare diseases where researchers tried to extend common data elements to domain-specific elements. So, in the same research field, the more specific the research question, the less overlapping variables.

Minimal datasets are also dependent on the particular field of the research. For instance, epidemiologists are likely to require different variables (and thus different fixed data items) than chemical specialists. Minimal datasets do not emerge out of anywhere: it is only possible to generate them once there is sufficient metadata about what other researchers are using (in terms of common data elements). Usually, the right strategy is to see how others use metadata and generate common codes for data structure and make an informed selection of them for a minimal dataset. However, it is also possible on some occasions to start from a common minimum dataset. The Joint Research Center for Rare Disease in EU defined common data elements that are now referenced by all the 24 European reference networks for data comparison.

The National Institutes of Health (NIH) convened a Working Group of scientific investigators to develop recommendations for CDEs for studies researching COVID-19 in paediatric participants[12]. The initiative proposed 49 Biomedical CDEs and measures across 8 domains (baseline child health, clinical / laboratory / cardiopulmonary diagnostic assessment / and imaging manifestations, diagnosis, treatment and outcomes). In addition, the initiative has recommended 50 CDs across 10 domains of Psychosocial aspect (Social Determinants of Health + Educational Factors, Community/ Family/ and Peer Factors, Social Media/Screen Time, Well-being Factors, COVID-19 Stress and Worry, COVID-19 Attitudes, Behaviours and Experiences, Health-related Behaviours, Mental/ Behavioural Health and Health Care). The CDEs are classified into two categories: highly recommended and recommended, and also recommendations of instruments to measure them are included.

[12] https://www.niehs.nih.gov/research/programs/disaster/database/promoting_data_harmonization_to_accelerate_covid_19_pediatric_research_508.pdf

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

4.2 DATA COLLECTION PROCESSES: OPPORTUNITIES AND CHALLENGES FROM EMERGING DATA COLLECTION TECHNOLOGIES

New data collection technologies can offer great opportunities to collect data in previously unreachable populations or geographical areas, and can thus provide new insights to human behaviour, physical action, social activity, and mental state. However, apart from the issues raised with respect to informed consent (cf. 3.1 Scope, continuity and governance of consent) and protection of personal data (cf. 3.3 Data Protection Safeguards), some other aspects be considered here. Data collection and data use are not neutral and hence, the decisions researchers make can (a) either support or harm **vulnerable communities**. A special category of data collection in the context of cohort research are the new and emerging data collection technologies; however, **digital technologies have a representativeness problem** due to (b) digital illiteracy and self-selection and (c) lack of specificity.

(a) Vulnerable populations

Vulnerable populations may be more reserved in granting their consent for many reasons, including stigmatization risks. Researchers may also be reluctant to involve vulnerable populations in an attempt to protect them from harm. As a result, vulnerable populations are not fully represented in cohort studies. Evaluating the scope and extent of consent in vulnerable populations may be hard to determine. The difficulty is increased in health cohort studies, as (in some occasions at least) they cannot directly test vulnerable populations, even if the research purpose concerns these populations in the first place. For instance, many treatments cannot be tested in cohorts related to disability for both ethical and legal reasons. As a result, **individuals with disabilities are *de facto* excluded from appropriate treatment** because the outcomes of research (e.g. medicines, drugs etc.) do not reflect their needs.

In the context of **mobile data collection**, there is a potential harm for vulnerable populations. In emergency situations, people are more willing to give away their personal and confidential data (place of living, psychological needs) or their rights to privacy and data monitoring/access (Merchant and Lurie, 2020). Online contexts exacerbate this problem because they offer structured choices that ''force'' participants to consent (Chancellor et al., 2019). The bargaining power of participants in the context of mobile data collection is thus limited because they have little say on how their data will be shared and used (Hand, 2018). Hence, the need for problems to be solved far outweighs any concerns about privacy (Copes et al., 2018). Mobile data collection has no real safeguards against the misuse of personal data in such circumstances, which means that private, intimate data collected during emergency times or from vulnerable populations can potentially be used later in a discriminatory manner (e.g. health benefits) (Ali et al, 2019).

Passive data collection can be harmful to vulnerable populations by increasing health related

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

disparities. Namely, passive data devices restrict care only to those communities that have the means, the time, the literacy and the cognitive ability to use them (Anaya et al., 2018). This means that vulnerable, undeserved communities are excluded from the health benefits related to personalised medicine. In this context, measures should be taken to make passive data collection as inclusive as possible (e.g. training programs free of charge) (Nebeker et al., 2019).

(b) Representativeness Issues and potential for in-built biases

Digital technologies have a representativeness problem: they tend to privilege a relatively young, middle class, population with satisfactory digital literacy and stable resources (e.g. a stable internet connection) at the expense of a socially and financially vulnerable older populations[13]. There is also the possibility of a **technological gap** associated with the social-economic class of the participants may make it difficult for them to continue to participate in the study in case of changes in data collection with new technologies if they do not have access to Internet, computers, smartphones, specific software or other electronic devices. Moreover, some individuals are more willing to give their personal data and comply with the conditions of the web service used for data collection (e.g. ''Google Maps"). Hence, the **results based on such data do not represent individuals who are more privacy-conscious**. Applying the results from communication devices data indiscriminately to the population as a whole does not only exclude such individuals from research. It is also prone to selection bias.

As a results, when **collecting data through social media**, the most important challenges are self-selection and representativeness (Arigo et al., 2018). In the former, some types of social media users are recruited at the expense of others (Khazaal et al, 2007), meaning that only **participants who manifest an interest in the study** because of their occupation or/and professional profiles take part. As mentioned above, there is a risk that participants with higher socio-economic status will be recruited at the expense of other social categories (Copes et al., 2018; Arigo et al., 2018). In the context of social media data collection, bias can also occur because individuals who like to reveal information about themselves via social media are overrepresented.

When it concerns **mobile data collection**, researchers should be aware that it can result in **partial data sets** and hence, can either distribute or repeat existing biases (Martinez-Martin et al. 2018; Olteanu et al., 2019). A related problem is that mobile data collection relies on information artefacts and software that potentially pre-define the data to be analysed. For instance, some software may prompt researchers to identity patterns that do not exist (Dixon, 2012).

(c) Potential for the disclosure of personal data and confidentiality breaches

Passive data collection unavoidably targets not only the participants themselves but **also the bystanders and family members who interact with them** (Ambrosini et al., 2018). Naturally, such an intrusion into a participants' private environment makes informed consent excessively difficult to obtain (Martinez-Martin and Kreitmair, 2018). Collecting data in patient care

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

contexts can also be highly problematic; there is uncertainty about the passive data applications (e.g. wearables to evaluate patients' health status) used, which often lack external validity (Corwin et al., 2019). Moreover, participants may experience psychological stress and unease due to fear for revelation of data related to their private sphere and may cope by changing their behaviour, which will distort the data collected (Geneviève et al., 2019; Anaya et al., 2018).

Also the **secondary use of passive data** is not straightforward because research proposal and frameworks become increasingly technical (Resnik, 2019). Namely, the focus is not much on secondary use per se but rather on the platforms, algorithms and diagnostic tools which make this secondary use possible both in legal and in practical terms (Corwin et al., 2019). In this context, institutional review boards experience increasing difficulties not only in determining the scientific worth of the proposals but also in evaluating the potential risks and benefits of participation (Janacek, 2018, Di Matteom 2018). The increasing technicality within the passive data collection and analysis domains is also confusing for the participants themselves. In particular, participants have difficulties in determining the significance of their contribution or the potential benefits of their participation (Anaya 2018; Nebeker et al., 2019). There are strong indications that researchers still do not have a clear strategy for fully integrating participants within the research process.

Finally, **raw media datasets** are often analysed according to distributed learning principles (i.e. artificial intelligence). This can reveal overarching patterns that could not have been visible with standard technologies, which in turn extends the scope and depth of data analysis (Merolli et al., 2013). However, raw data analytics can also reveal patterns that can be either biased or incorrect (Olteanu et al., 2019; Izmailova et al. 2018). Raw data sets are vulnerable to biases because their content is unstable from the onset (Taylor and Pagliari, 2018). Twitter data, for instance, is produced by different actors (bots, humans and institutions) who each have different motives and incentives for their online interventions (Tene and Polenetsky, 2012). In this sense, what counts is not only what is in the data (e.g. biases) but also what is not (e.g. excluded communities). Raw media data should thus be complemented by an in-depth exploration about the nature of data and the conditions of its production. In any case, **no health policy** should be based on the patterns derived from raw data alone (Nebeker et al., 2018).

Summary and recommendations:

- Whereas standardisation of data collection might be too challenging; standardisation of data and their representation is a realistic goal. Research consortiums with equal health-topics should join their forces in developing Common Data Models (CDMs).
- CDEs are enablers to create a common terminology to facilitate data structure and data sharing; hence, they should become a standard vocabulary for an optimal exploitation of cohort data.
- Minimal datasets require actively engagement of researchers; depending on the field of research, consensus finding might be more or less feasible.
- Researchers are reluctant to involve vulnerable populations in an attempt to protect them from harm. As a result, vulnerable populations are not fully represented in cohort studies. Researchers should be aware of this and apply appropriate measures to be as

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

inclusive as possible.

- Mobile data collection has no real safeguards against the misuse of personal data, which means that private, intimate data collected during emergency times or from vulnerable populations can potentially be used later in a discriminatory manner (e.g. health benefits).
- New data collection technologies have representativeness problems, and a potential for the disclosure of personal data and confidentiality breaches; implementing technical capabilities for effortless and retroactive retraction of data should be explored.
- Analyses of data collected through new data technologies should be complemented with an in-depth exploration about the nature of data and the conditions of its production. In any case, no health policy should be based on the patterns derived from raw data alone.

5. METADATA

Metadata is used to define variables that are in databases with information about the context in which that data have been produced. It is the sort of **item by item description of what's in a data set**, providing valuable descriptions. Metadata can provide the formal structures to govern data. Collecting metadata and developing metadata standards is important in order to provide such contextual information for clear interpretability, and to facilitate data management and data pooling across different research and cohort studies.

Hence, if the aim is optimal exploitation of cohort data, proper metadata and metadata documentation are prerequisites for future cohort data re-use and integration. Therefore, standardisation of metadata (documentation) by the scientific community should at least be encouraged. Nevertheless, there are some hurdles to overcome. To start, (a) there is a lack of open documentation, (b) the current landscape of existing standards should be mapped, (c) there should be consensus on a minimal appropriate level of metadata documentation, (d) determine standard metadata content, (e) metadata collection should be encouraged, and (f) incentivised.

(a) Lack of open documentation and description of metadata

Including metadata effectively is not as easy as it seems, mainly because the harmonisation processes related to metadata are rarely made publicly available. Many researchers leading cohort initiatives do not report the harmonisation process transparently and rigorously enough for assessing its validity and ensuring its reproducibility in future initiatives. Researchers seem often more focused on the results of their cohort initiative projects rather than on documenting and publishing 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 validated nor reproduced. With respect to metadata, it means that no functional examples or

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 become available for other researchers to follow.

This lack of transparency and details in harmonising methodologies complicates the creation of a catalogue for samples and data. Cohorts often do not communicate detailed information on their research resources in harmonised ways. Namely, EU cohort 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. As a result, researchers will experience significant difficulties in finding the appropriate variables and data, if they wish to answer a certain research question by means of a cohort harmonisation process. The latter might especially be the case where sample sizes are small due to the very nature of the population under study (e.g. rare diseases).

The general lack of information in catalogues means that metadata descriptions remain underdeveloped within the field of health-related cohort research. This is because the level of description and documentation depends on the nature and use of harmonized data. Namely, it is not always clear what is harmonized in the first place because it is impossible to fix the data items themselves. It is only possible to harmonize data organization, and rely on prefixes and suffixes to clearly relate data points to definitions. Thus, except for a few instances where core fixed data are possible (e.g. demographics), each study will have its own data points with specific data items.

(b) Cataloguing existing standards

While there are already many standards in use, there is not enough empirical data about them. It is thus difficult to determine what standards people are using in practice. There are already standard (descriptive) metadata available such as the Data Documentation Initiative (DDI). DDI is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences[14]. Or the Library and Information Science (LIS), involved in examining the issue of metadata quality, and describing and evaluating characteristics of metadata records in order to inform and assess the processes by which it is created and the functionality it can support (Robertson, 2005). Also the Data Curation Centre (DCC)[15] provides digital objects, administrative, descriptive, structural and technical archival metadata, which helps to find links of standards depending on the area of research, Moreover, DCC facilitates software to capture, store, share, and validate metadata (e.g.: Bio-Formats (microscopy data), CKAN). However, there is still need **for a list of appropriate standards that researchers could follow**.

Use of standards of metadata could encourage data re-use, valid interpretations of research findings and study reproducibility (O'Neill et al., 2019) and the generation of interoperable catalogues. Here there is no point in reinventing the wheel. For surveys (such as the European Health Survey, EU-SILC), there are already existing modules that are defined as recommended

[14] https://ddialliance.org/
[15] https://www.dcc.ac.uk/guidance/standards/metadata

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. However, their actual use is both wide and fragmented so researchers may have difficulties finding them. The solution would be to make a catalogue for collecting those standards inputs. This would enhance standardization and facilitate harmonisation since researchers would no longer have to generate new modules from scratch and would be able to use already existing modules instead. Such an approach may be especially helpful for the design of minimum datasets.

(c) Appropriate level of metadata

The level of detail has resource implications but also depends on the nature and use of the (harmonized) data. "***Basic minimum metadata documentation*** *should be suggested for other cohort studies to follow. However, this basic minimum documentation is also determined by funders' needs and expectations. Namely, the question is which kind of details the funders should specify when they put out the request for proposals*" [Stakeholder]

Discovery metadata refers to its general topic, its source, its authors, the type of data searches involved and its access arrangements (e.g. license). **Descriptive metadata** refers to a detailed item description of the dataset content. Descriptive metadata include item type categories, control vocabularies and definitions (common language and terminologies). Descriptive metadata provides a detailed description but it is advisable to go further and to describe how this data has been collected in the first place. The data collected through questionnaires is relatively easy to document. However, in most cohort studies data collection implies different types of data such as tabular data (questionnaires/surveys), non-tabular data (EEG experiments, (f)MRI scans…), and/or biological samples. Moreover, the method for data collection includes **objective measurements** and **contextual metadata** that require even more detailed descriptions (e.g. the kinds of devices and calibration used, of they were machine-generated etc.). As mentioned above, listing the methods for standard metadata descriptions currently used in different catalogues (e.g. in the BBMRI, the EUPHA) could help researchers to determine an adequate standard of metadata descriptions. These could furthermore, be used by the European Commission in their strategy for standards definition.

In practice, however, contextual metadata is not documented in a consistent way. This is in strong contrast to the situation for clinical trial studies where researchers can rely on an established **consistent system for documentation** (CDISC define-XML or OMOP system for observational data).

(d) Standard metadata content

In general terms, the standard metadata content depends on the research question. In order to determine what standard metadata is, however, it is necessary to determine what type of metadata it concerns. As we have noted in the previous section, metadata can concern discovery data, descriptive data but the development of standards will differ for both of these cases. That is, **standardizing descriptive metadata** is in principle easier than standardizing

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

discovery metadata.

Developing metadata standards requires defining the metadata to collect, the vocabulary to represent the metadata, and how to encode the data for transmission (Badawy et al., 2019). Metadata from clinical research, for instance, have to include information about (Badawy et al.,2019):

- Time: comprises relevant metadata regarding the temporal information about the data collected during the course of primary data collection; predefined document date, time (based on standard time zone) to order them chronologically.
- Person: relevant metadata pertaining to the individuals who were the subjects or those associated with a subject but not under study.
- Context of collection: encapsulates relevant metadata relating to a study, its technology, conduct, processes, and procedures.
- Observation: captures relevant metadata relating to data collected during a study or interpreted through post-study analyses, including those reported from a digital health technology or a human being.

Specific domains of metadata standards employ different vocabularies across research fields and study designs. For example, in the YOUTh cohort study, data are organized by means of four characteristics: wave, experiment code, pseudocode and version (Zondergeld et al.,2020), and use JSON Schema6 which determines whether the values of extracted metadata are valid. Maelstrom Research cataloguing toolkit, initiated in 2004, has explored existing catalogues and standards in order to interpret and analyse cohort data (e.g.: biological samples), creating a comprehensive and user-friendly web-based metadata catalogues. It promotes classifying all variables in a study into a standard variable classification taxonomy of 18 information areas and some 135 sub-domains (Bergeron et al., 2018). Related to software uses for metadata, Mica[16] facilitates a portal for individual epidemiological studies or multi-study networks. It helps investigators and data custodians to efficiently disseminate information about their studies and metadata.

Also noteworthy is the CINECA project that currently explores existing data representation as variables recorded, variable values and coding systems used of ten cohorts to construct a common minimal metadata model aligned with output from international standard activities (e.g.: ELIXIR, GA4GH, BBMRI, EOSCpilot and P3G)[17].

(e) Metadata collection

In the field of metadata collection, academic health data is rarely recorded with detailed structural information (unless they are transferred to the FDA or EMA). There is thus a need for guidance as to the detail needed. A lot depends on the use preferences from statisticians.

[16] https://www.obiba.org/pages/products/mica/
[17] https://www.cineca-project.eu/harmonized-metadata

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

Namely, while the IT staff and the data managers are often willing to use the proposed standards for data (recording), statisticians are more conservative and refuse to use standards they are not familiar with. The main issue is, therefore, to train and prepare the staff with different backgrounds so that the **acceptance for standards** could be increased. Such a process cannot be implemented in a top-down manner, however. Instead, the acceptance for standards should come from the bottom up or more precisely, from the research community. That is, if metadata collection and websites are commonly used, it is because researchers, IT people and statisticians are committed to them through practice.

A more general point is that one generally aims to structure data in consistent ways so that it could be comparable. This is why the level of detail in the structural information for data is crucial. For instance, if we want to investigate the side effect of medication, it would be very useful for us to already have a definition of when and how side effects are recorded. That is, in order to do a consistent comparison, it is necessary to know what each of the researchers involved means by certain terms such as “dizziness”.

(f) Incentivising proper metadata documentation

Now we have made a case for improving metadata standardisation and documentation in the context of cohort research. How could this be achieved? Incentivising on proper metadata description could be a major source of motivation for detailed documentation standards. Setting a minimum set of data/metadata, however, is to a major extent a **cost issue**. Funders tend to not adequately allocate resources for such activities while researchers tend to neglect the issue. It is thus necessary to find effective incentives. There are three sources of such incentives: research community, funding agencies, and research infrastructures. Proper metadata documentation should also be rewarding in some way (e.g. crediting the effort).

First, there is the possibility to convene a **research community** so that a common researcher-based consensus can be reached. If journals are set to publish the harmonisation process in the detail, this may incentivize researchers to use the same metadata. However, there should be clarity about the type of standards that researchers want to reach a consensus on.

With respect to **funding agencies**, their necessities and expectations could determine metadata documentation, but basing funding on proper metadata descriptions could exclude all research projects that do not have the resources to provide a whole set of metadata. The consensus achieved by the researcher community can be reinforced by transnational and international agencies (such as the World Health Organisation or the European Commission). Such agencies have the mandate and the power to implement this researcher community consensus.

Research Infrastructures should step up to provide services for researchers so that consistent metadata can be created. Expertise can be concentrated in one place and researchers can rely on a central service. Though this would mean still a need for funding to facilitate documentation and creation, it would facilitate and harmonise the metadata documentation across cohort studies. Hence, a Research Infrastructure that effectively credits, supports, and guides

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

researchers in publishing their metadata would be indicated/urgently needed. Moreover, Research Infrastructures would also be indicated to take on the task of generating a catalogue for the collection of recommended standards. This means an institution at EU level should be in the position to guide, steer and structure such a catalogue. An example of such an active engagement can already be found at the WHO (World Health Organization level) with GATHER guidelines for global health estimates. GATHER guidelines are now de facto requirements for publication in high impact journals. Moreover, a standard collecting catalogue should specify the scope of the standardization to be done and the type of research community concerned. By concentrating all these facilities in one service, Research Infrastructures would also be the indicated entities to provide training on proper metadata descriptions and documentation to researchers and others involved in data management.

"*It should be noted that basing funding on proper metadata description is a double-edged sword. On the one hand, it can be a major source of motivation for detailed documentation of standards. On the other hand, it can become dangerous because it would exclude all research projects that cannot provide a whole set of metadata*" [Stakeholder]. Documenting metadata consistently requires a lot of time and effort so researchers should be credited for publishing their metadata. However, it is not clear if researchers' careers will be really advanced through metadata publication. More importantly, while social sciences libraries of education research (such as ERIC) actively support and credit researchers (by, for instance, supporting them to upload their material, giving them reference for the uploaded data description), there is no similar infrastructure in the health domain/health cohorts.

Summary and recommendations:

- Due to a lack of openly and properly documented information about harmonisation processes, no functional examples or standards become available for other researchers to follow. The general lack of information means that metadata descriptions remain underdeveloped within the field of health-related cohort research. Efforts should be made to train researchers in publishing detailed information about harmonisation processes.
- Although it depends on the nature of the study itself, how to structure and organize the data, developmental standard protocols, catalogues or documentation are primordial to facilitate later interoperability and consistency of variables. Likewise, standards of metadata could encourage data re-use, valid interpretations of research findings and study reproducibility, and the generation of interoperable catalogues. Existing standards for metadata and best practices in metadata documentation should be listed, so researchers can identify appropriate standards to follow.
- Basic minimum metadata documentation must be suggested and set as a standard to follow. This might be determined by funders' needs and expectations, and the level of details they require. Funding agencies should incentivise proper metadata documentation (e.g. provide funding for technical support).
- Documenting metadata consistently is a costly and time-consuming endeavor;

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

researchers should be credited for publishing their metadata.

6. DATA SHARING

6.1 INTEROPERABILITY

In the context of cohort research, there is still no generic mechanism for interoperability of projects and methodologies in place. In general terms, interoperability refers to the ability and the potential of information systems to share data (EDPS, 2021; Geraci, 1991). Interoperability depends on the design and content of the building blocks (i.e. storage, structure, data transfer and data integration) of data infrastructures.

Moreover, data interoperability (both at the individual data level and at metadata level) focuses on the types of data, data flows and core standards that are the most likely to facilitate data sharing and data access. The aim is to achieve FAIR data on a consistent basis, that is, data thatis findable, accessible, interoperable and reusable.

However, there are many barriers to achieve data interoperability. For example, the **structure** (i.e. how the data is organised) in which the source data are stored may be heterogeneous and unclear across cohorts (e.g. the data sources are not always tracked, the rationale for variables' definition is lacking), which can result in a loss of interoperability. Also, **data transfer** may be compromised by contradictory data access and data transfer procedures across participating institutions providing source data. The same applies for data from different cohort studies. The integration of different cohort studies into a European cohort repository may be prevented by the heterogeneity in used measures and instruments to assess a certain construct, as well as the use of ambiguous terminology. Finally, interoperability is further impeded through **incompatibility due to the use of specific software** for different cohort projects and their respective infrastructures.

The challenge resides in organizing data (i.e. data **storage**) in a manner that optimises findability and retrieval through a comprehensive data **structure**. Hence, a comprehensive **data structure** is fundamental. This can be achieved by means of characteristics such as wave, experiment code, pseudocode (participant identification code) and version generating an internal unique classification (code) for all datasets.

(b) Query strategies

An important factor for the interoperability of data infrastructure is the extent to which researchers are able to make appropriate queries in the database. In general terms, the solution is to adopt a **user-centric view** (in this case this means a focus on the cohort data holders and data providers). In practice, for queries purpose, users have standard analytical tools, good

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

practices derived from federated analysis and Bayesian analysis when they have to engage in data assembly themselves. However, this says little about what these users actually need. In order to attend to the needs of the users, it is necessary to suggest query arrangements and methods and evaluate the subsequent response of the user community as to what measures are feasible.

Another solution is to adopt a **hybrid approach to query strategy**. Namely, while some data centres are well equipped for complex queries, others are hampered by discoverability problems. Users should form a research committee to examine such query capability in databases as well as to assess if queries can be carried out in a safe environment. The variability of query organisation is demonstrated in various databases across Europe. In practice, there are already many instances of these hybrid approaches where some databases/initiatives use federated approaches while others used centralized approaches. HDS (European Head Status) rely on infrastructure nodes for queries arrangements and BMRI (with their virtual colorectal cancer cohort) are able to pool data in parts. In Finland by contrast, researchers tend to use a federated approach. As a rule of thumb, it makes sense to see which approach database setters have taken in certain local contexts and use such approaches as benchmarking when input fields in large-scale cohorts are concerned.

A hybrid approach to data queries thus pays special attention to the national infrastructures. In France, Germany, in Wales, researchers are creating national hubs of health data (mixed with social science data) that require various data safe haven schemes. Data safe heaven schemes allow users to access the database under certain conditions and preserve the potentially use of data resources of national hubs.

6.2 INTEROPERABLE INFRASTRUCTURES

There are three main **types of interoperable infrastructures** for sharing individual data within an initiative namely:

(i) the individual cohort datasets reside in different institutions (federated networks), mostly on the server of origin (i.e., data are in different locations),
(ii) the individual data is centralised in one institution or server (i.e., central location of data, e.g. data lakes), and
(iii) mixed location types (some data are located centrally and some data locally).

Both, (a) decentralized and (b) centralized types of governance involve hurdles for harmonisation and integration processes and thus hamper interoperability.

(a) Decentralized - Federated infrastructures

Federated networks are a technical solution that includes linked databases while providing a

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

secure, real time environment for data access (e.g. in the case of data breach the access to the rest of the repositories can be blocked (Pezoulas et al., 2020)), data storage (e.g. does not have centralization problems such has the amount of resources needed for secure data storage in a single platform) and data use (e.g. institutions can control what data they want to be shared in in a de-identified form).

Generally, federated networks are considered to be the most viable strategy for the interoperability of data infrastructures. However, while federated infrastructures are increasingly being used for cohort data, they also present some flaws.

Curating and organising data for federated infrastructures requires a lot of effort and there are considerable difficulties in finding appropriate curating strategies. The issue becomes even more complex because there are considerable **data quality issues** (e.g. incomplete datasets, lack of appropriate documentation, or misclassification of data). Federated infrastructures, namely, include little control on the quality of data included in cohort studies. The problem is that federated infrastructure come with a cost and it is not always clear who has to assume these costs. For instance, who will pay for the software or recruit managers for using the software on the national nodes? Unless such issues are solved, it is difficult to develop infrastructures that will assure data quality. Alternatively, it is always possible to empower the individual resources to do professional data quality management.

In the context of dataset queries, there are also issues of analytical compatibility and interoperability because some elements are not practically feasible in federated analysis.

With respect to **confidentiality**, the main issue is the increasing importance of federated infrastructures in cohort research. Namely, federated approaches are not created equally since some systems rely on more resources than others. The fact that researchers assess the data without ever seeing it, prompts the question about their ability to prevent the reproduction of biases. Data re-use is not neutral and the linguistic structure of the law in the privacy domain can lead to discriminatory limitations on data sharing. The structure of federated governance has to be rethought in terms of the ethical and legal contexts where the data subjects find themselves in.

Effective use of federated analysis would also require an open dialogue between the developers of software systems and the actual end-users, in order to make federated solutions as user friendly as possible.

(b) Centralized - Data lakes

Data lakes in the context of interoperability of data infrastructures, is not a completely satisfactory solution for many reasons. Data lakes are “centralization types” solutions that secure data in vast secure data centres systems. However, they can be quite difficult and costly to implement depending on how they are implemented in practice. If data lakes refer simply to the practice of putting data in one place, then it is necessary to consider both legal and ethical

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

issues that may arise from centralization. If by contrast, data lakes refer to curating data directly so that it can be placed into a single platform, then they require a huge amount of resources. Moreover, data lakes require a centralized compute-to-data process, which is a costly endeavour.

Thus, the issue boils down to the pros and cons for centralized and decentralized approaches for governance. Despite its scalability problem, the **federation approach remains the preferred solution**, especially since we have the technical tools to apply it to cohort research. However, on a wider scale, it is necessary to ensure that potential impact of each scientific strategy, such as data lakes and federated networks, is approved by the research community. That is, "a scientific strategy cannot be implemented if the researcher community does not see any practical improvement from its use" [Stakeholder]. It is also necessary to ensure that "*what is being done during the project will have a positive impact in the future*" [Stakeholder]. Therefore, there should be a mechanism to check the impact of the scientific strategy once the project has been completed. In general terms, unless the boundaries of usefulness for federated analysis are firmly established, it is difficult to know if the investment required is worth it and what to expect from this investment.

6.3. ALIGNING PROCEDURES

In the context of cohort research, data sharing is also hampered by the heterogeneity in data access procedures across cohorts. There are two main hurdles for access namely (a) heterogeneity in data access governance and (b) an extensive amount of time needed for local approval procedures.

(a) Data access governance

The interoperability of data transfer and data integration procedures are hampered by the high degree of heterogeneity in governance structures. In general, data (e.g., imagining data) can be anonymised and thus shared accordingly. However, for types of data that cannot be anonymised (e.g. genetic data) increasingly complex procedures are being used in order to meet GDPR's requirements for data sharing. Data access governance structures can rely on controlled access and registered access for sharing data. In the former case, data controllers grant access to datasets only to approved data users under certain conditions (Shabani et al., 2021). As such, controlled access is the exact opposite of open access and is suitable for cases where privacy concerns are particularly acute (e.g. genomic research). By contrast, registered access concerns low risk personal data (e.g. non-stigmatizing data from healthy individuals who have consented to data sharing). The model assumes that since the processing of data would not create risks of re-identification, it would be enough to simply restrict data access to trusted users (Peloquin et al., 2020; Slokenberga et al., 2021). The policy in EU countries for **controlled access models** is generally to rely on legally binding 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

access agreements overseen by DACs (Data Access Committees –DACs; Article 89 of the GDPR). The main issue here is that DACs often do not possess adequate oversight tools for monitoring potential breaches in data access agreements (Cheah and Piasecki, 2020; Shabani et al., 2021).

With respect to the **registered access model**, it requires numerous steps (i.e. authentication, authorization and attestation) but data users do not have to sign a data access agreement in a paper-based format with DACs (online agreements with DACs are sufficient). This is a significant advantage over the controlled model whose mode of operation is administratively and technically burdensome (Slokenberga et al., 2021). "*However, other components of the governance structure, for example, data access committees will still be needed because you're still internally doing research and the data and the data access community is very useful in ensuring that subsequent uses are in line with what the participant was told and what was indicated in the consen*t" [Stakeholder].

Both types of governance involve hurdles for data sharing and thus hamper interoperability. 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. 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).

Some EU initiatives build a virtual platform that contains a **federated data discovery ecosystem**. In practice, this means that users can get access to a variety of linked disease databases for rare diseases at different levels of specificity. Queries are adapted to the degree of specificity. For the lowest specificity level, automatic queries arrangements quickly give a response without asking for detailed permission. At the higher specificity levels, users have to get in contact with the resource and ask for permission from data access committees.

(b) Data access approval procedures

In the context of cohort research, 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. Oftentimes, it is difficult and time consuming to identify the correct contact point for access correspondence. Researchers must contact each cohort separately to request information about data access and fulfil a heterogeneity of data access applications.

Most cohorts have local scientific access committees for evaluating access requests and dealing with ethical, legal and administrative issues. Each local committee might have its own criteria for granting access or not. Moreover, countries have different data acquisition procedures and these differences impose restrictions not only on the type of data available but also on 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

negotiations between countries for access (Ienca et al., 2018). The time required and the lack of technical staff could further explain the difficulties in data sharing.

In practice, data sharing can also 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. Hence, when necessary approvals for obtaining access have been approved, there are still significant administrative obstacles that cause additional delays due to establishing the necessary Material Transfer Agreements (MTA).

Extended time lags for approval and access to data can be partly solved 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; nevertheless, such technologies come with a cost and require resources.

To encourage the re-use of data, researchers should be facilitated in obtaining access. This can be done through the involvement and support of trained staff, but what really is needed is an overall alignment of data access procedures. In this sense, Europe should invest in a few centralized, ideally framework for storing maintaining and enabling ongoing data sharing.

6.4 EUROPEAN COORDINATION

If it comes to data sharing, we should step up into practical rather than theoretical uses of standards and methodologies for interoperability. To start, (a) a clear overview or information is needed about what tools and standards are being used, and what their respective (dis)advantages are. Another approach would be to start doing by trying, and thus, (b) start integrating existing cohort

(a) EU level inventory

A first necessary step would be to get a **picture of the current landscape of EU member states' infrastructures** and data hubs. This should be complemented with an exploration of the current situation for interoperability. In particular, there should be a focus on the standard metadata available, data standards, software for federated analysis and possible future data descriptors.

In terms of data infrastructure in general and federated structures in particular, it is advisable to explore, utilize and exploit the existing landscape of infrastructures currently in use in EU member state countries. For instance, a common practice is to rely on **automated responses for**

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

queries as long as the data is not shared. Data use, by contrast, warrants compliance to the requirements of the cohort concerned. In parallel to an exploration of current queries practices, it is necessary to create a state of the art of national initiatives and project collaborative research initiatives.

Likewise, would it be crucial to have a clear picture of the pros and cons of the available software for federated software, including their assumptions, properties, analytical possibilities (e.g. Bayesian) and the requirements for central hubs and individual nodes. We should know if the technology and software currently in use are scalable enough for a large number of cohorts and nodes. A certain level of standardization has to be reached so that links between federated systems and cohort data can function seamlessly.

A solution is to develop a repository/inventory that would give an overview of the software available. It would help us to determine what kind of requirements the software has in relation to federated systems and the kind of analysis they are able to perform. This material can be linked to the landscape of use for specific software by different projects. "*If we see that specific software is commonly used, we can use it further for scientific projects with similar structures and/or properties*" [Stakeholder]. That is, "*we should focus on big cohorts (as opposed to small) ones, identify the software used and evaluate which standardisation strategy is used in each case*" [Stakeholder].

(b) Piloting

Once bottlenecks and practices of the current initiatives are identified, it is advisable to create a forum and a pilot proposal/study on the basis of specific small user cases. This would allow identifying possible valuable approaches, the limits of these approaches and the architecture that would best suit users' needs.

A next step is to utilize the initiatives landscape by starting to integrate existing cohorts (after getting agreements with owners). This will allow identifying the real challenges and facilitators in practice and uncovering unexpected problems. On a more general level, researchers from different domains have to test the whole concept field. The results and practical solutions of such enquiries should be integrated in a consortium to check their feasibility.

In addition, it is necessary to create a state of art of national initiatives and project collaborative research initiatives, identifying the current initiatives (e.g. to take references of them by starting to integrate existing cohorts) to create a forum and a pilot proposal/study on the basis of specific small user cases. Besides, to take references of other initiatives to utilize landscapes by starting to integrate existing cohort.

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

6.4. RESEARCH COMMUNITY AND COMMITMENT

(a) Commitment

Dorey et al. (2018) note that researchers are mostly committed to data sharing as they consider that it adds social value to databases and Electronic Health Records (EHRs). However, they are warry of collaborations with experts from non-medical disciplines and transparency, and influenced by funding and grants requirements.

Data sharing is not an easy task because dissenting forms of governance, often generated by the public itself, play an increasingly pro-eminent role (e.g. patient driven registries for COVID 19). The public is increasingly willing to organise, create and share its own data registries independently of the research community. The proliferation of such patient-driven registries is an indictment of the data sharing scientific community as it has failed to timely share data in a way that would be relevant for the general public. Namely, the public does not share the researchers' community concern with consent formalities and are wary of the GDPR constraints.

Researcher communities in transnational initiatives' contexts (such as EMR and the European Health Data space), should thus assess how the increasingly independent role of the general public in the creation, governance and access arrangements of its own registries may redefine the common data sharing practices in the research space. In particular, the research community should assess its "meta-sharing" capabilities. Namely, how do their data sharing platforms and registries interact with the general public? Such considerations are crucial in supporting the trust of the public towards cohort data.

Furthermore, the rationale for sharing data is rarely determined by the researchers themselves but rather by grant terms, funding, peer-review requirements, country regulations and ethical committees' approval (Kiehntopf, 2019).

(b) Training needs

Training for researchers depends on the software they already use since different tools require different training. Training can be managed in terms of short and long term goals in accordance with different factors. Thus, in the short term, it is possible to make an inventory of what tools are available and what training needs should be met. In the long term, it is possible to identify how methodologies developed with federated infrastructures in mind can be extended to other infrastructure types. Of particular importance are the interactions of these methodologies in relation to ethical issues (such as privacy) and the legal framework. At this point, it is possible to design workable data lakes for a secure hub of private data.

Summary and recommendations:

- There should be an exploration of the current situation for interoperability. In particular, there should be a focus on the standard metadata available, 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

standards, software for federated analysis and possible future data descriptors. We should create an inventory of standards.

- We should focus on end-users and user communities and evaluate their needs. In particular, we should actively encourage them to use standards either through funding, publications or workshops.
- The types of data queries and training needs can be determined by a pilot study approach. In this context, we should adopt a learning-by-doing approach with a future user community in mind.
- It is impossible to generate appropriate methodological approaches to the interoperability of data infrastructure, without considering what kind of governance is needed. Institutions (such as the W.H.O) should be committed to the definition and enforcement of appropriate standards.
- We should step up into practical rather than theoretical uses of standards and methodologies for interoperability. We are currently missing information about tools, standards and about the distributions of projects. Namely, a lot of projects tend to run in parallel and while their output is meaningful, the interaction between them remains scarce. As a result, projects are not able to learn from each other.
- We should engage in a forum with research communities, institutions and stakeholders in order to determine the feasibility of federated infrastructures. This can be done by running small use cases (as pilots) in order to test the limits of the structure adopted. The propriety is that we should be certain about what kind of architecture could best fit the user community's needs.

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

8. REFERENCES

Ahmad, O. F., Stoyanov, D., and Lovat, L. B. (2020). Barriers and pitfalls for artificial intelligence in gastroenterology: ethical and regulatory issues. *Techniques and Innovations in Gastrointestinal Endoscopy*, *22*(2), 80-84.

Aiello, A. E., Renson, A., and Zivich, P. (2020). Social media-and internet-based disease surveillance for public health. *Annual review of public health*, *41*, 101.

Ali, J., DiStefano, M. J., Coates McCall, I., Gibson, D. G., Al Kibria, G. M., Pariyo, G. W. and Hyder, A. A. (2019). Ethics of mobile phone surveys to monitor non-communicable disease risk factors in low-and middle-income countries: A global stakeholder survey. Global public health, 14(8), 1167- 1181.

Ambrosini, A., Calabrese, D., Avato, F. M., Catania, F., Cavaletti, G., Pera, M. C., ... and Pareyson, D. (2018). The Italian neuromuscular registry: a coordinated platform where patient organizations and clinicians collaborate for data collection and multiple usage. Orphanet journal of rare diseases, 13(1), 176.

Anaya, L. S., Alsadoon, A., Costadopoulos, N., and Prasad, P. W. C. (2018). Ethical implications of user perceptions of wearable devices. Science and engineering ethics, 24(1), 1-28.

Arigo, D., Pagoto, S., Carter-Harris, L., Lillie, S. E., and Nebeker, C. (2018). Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery. Digital health, 4, 2055207618771757.

Ashcroft, Richard E. (2001). *Money, Consent, and Exploitation in Research. American Journal of Bioethics, 1(2), 62–63.* doi:10.1162/152651601300169158

Badawy, R., Hameed, F., Bataille, L., Little, M. A., Claes, K., Saria, S., ... and Karlin, D. R. (2019). Metadata concepts for advancing the use of digital health technologies in clinical research. *Digital biomarkers*, *3*(3), 116-132.

Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I (2018) Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS ONE 13(7): e0200926. https://doi.org/10.1371/journal.pone.0200926

Bialke, M., Bahls, T., Geidel, L., Rau, H., Blumentritt, A., Pasewald, S., ... and Hoffmann, W. (2018). MAGIC: once upon a time in consent management—a FHIR® tale. *Journal of translational medicine*, *16*(1), 1-11.

Bilkey, G. A., Burns, B. L., Coles, E. P., Bowman, F. L., Beilby, J. P., Pachter, N. S., ... and Weeramanthri, T. S. (2019). Genomic testing for human health and disease across the life cycle: applications and ethical, legal, and social challenges. *Frontiers in Public Health*, *7*, 40.

Borry, P., Bentzen, H. B., Budin-Ljøsne, I., Cornel, M. C., Howard, H. C., Feeney, O., ... and Felzmann, H. (2018). The challenges of the expanded availability of genomic information: an agenda-setting paper. *Journal of community genetics*, *9*(2), 103-116.

Budin-Ljøsne, I., Teare, H. J., Kaye, J., Beck, S., Bentzen, H. B., Caenazzo, L., ... and Mascalzoni, D. (2017). Dynamic consent: a potential solution to some of the challenges of modern biomedical research. *BMC medical ethics*, *18*(1), 1-10.

Cabili, M. N., Lawson, J., Saltzman, A., Rushton, G., O'Rourke, P., Wilbanks, J., ... and Philippakis, A. A. (2021). Empirical validation of an automated approach to data use oversight. *Cell*

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

Genomics, *1*(2), 100031.

Candilis, P. J. (2002). Distinguishing law and ethics: a challenge for the modern practitioner. *Psychiatric Times*, *19*(12).

Cech, M. (2018). Genetic Privacy in the Big Biology Era: The Autonomous Human Subject. *Hastings LJ*, *70*, 851.

Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., and De Choudhury, M. (2019, January). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 79-88).

Cheah, P. Y., and Piasecki, J. (2020). Data access committees. *BMC medical ethics*, *21*(1), 1-8.

CoE (1997). (Oviedo Convention) Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine. Available at: https://www.coe.int/en/web/conventions/full-list/-/conventions/rms/090000168007cf98.

Conley, E., and Pocs, M. (2018). GDPR compliance challenges for interoperable health information exchanges (HIEs) and trustworthy research environments (TREs). *European Journal of Biomedical Informatics*.

Copes, H., Tchoula, W., Brookman, F., and Ragland, J. (2018). Photo-elicitation interviews with vulnerable populations: Practical and ethical considerations. *Deviant Behavior*, *39*(4), 475-494.

Correia, R. B., Wood, I. B., Bollen, J., and Rocha, L. M. (2020). Mining social media data for biomedical signals and health-related behavior. *Annual review of biomedical data science*, *3*, 433-458.

Corwin, E., Redeker, N. S., Richmond, T. S., Docherty, S. L., and Pickler, R. H. (2019). Ways of knowing in precision health. *Nursing outlook*, *67*(4), 293-301.

Council for International Organizations of Medical Sciences(CIOMS). (2017). International ethical guidelines for health-related research involving humans. International ethical guidelines for health-related research involving humans. [Guideline 1]

Di Matteo, D., Fine, A., Fotinos, K., Rose, J., and Katzman, M. (2018). Patient willingness to consent to mobile phone data collection for mental health apps: structured questionnaire. JMIR mental health, 5(3), e56.

Dixon, D. (2012). Analysis Tool or Research Methodology: Is there an epistemology for patterns?. In *Understanding digital humanities* (pp. 191-209). Palgrave Macmillan, London.

Dobrick, F. M., Fischer, J., and Hagen, L. M. (Eds.). (2018). *Research Ethics in the Digital Age: Ethics for the Social Sciences and Humanities in Times of Mediatization and Digitization*. Wiesbaden: Springer Fachmedien Wiesbaden.

Dorey, C. M., Baumann, H., and Biller-Andorno, N. (2018). Patient data and patient rights: Swiss healthcare stakeholders' ethical awareness regarding large patient data sets–a qualitative study. *BMC medical ethics*, *19*(1), 20.

Dove, E. S., and Garattini, C. (2018). Expert perspectives on ethics review of international data-intensive research: Working towards mutual recognition. *Research Ethics*, *14*(1), 1-25.

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

Ducato, R. (2020). Data protection, scientific research, and the role of information. Computer Law and Security Review, 37, 105412.

Duguet, A. M., and Herveg, J. (2021). Safeguards and Derogations Relating to Processing for Scientific Purposes: Article 89 Analysis for Biobank Research. In *GDPR and Biobanking* (pp. 105-120). Springer, Cham.

Euser, A. M., Zoccali, C., Jager, K. J., and Dekker, F. W. (2009). Cohort studies: prospective versus retrospective. *Nephron Clinical Practice*, *113*(3), c214-c217.

Firchow, P., and Mac Ginty, R. (2020). Including hard-to-access populations using mobile phone surveys and participatory indicators. Sociological Methods and Research, 49(1), 133-160.

Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., … Burton, P. R. (2011). Is rigorous retrospective harmonization possible? Application of the dataSHaPER approach across 53 large studies. International Journal of Epidemiology, 40(5), 1314–1328.

Geneviève, L. D., Martani, A., Wangmo, T., Paolotti, D., Koppeschaar, C., Kjelsø, C., ... and Elger, B. S. (2019). Participatory disease surveillance systems: ethical framework. *Journal of medical Internet research*, *21*(5), e12273.

Guinchard, A. (2018). Taking proportionality seriously: The use of contextual integrity for a more informed and transparent analysis in EU data protection law. *European Law Journal*, *24*(6), 434-457.

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now?. *Big data*, *6*(3), 176-190.

Hansson, M. G. (2021). Striking a balance between personalised genetics and privacy protection from the perspective of GDPR. In *GDPR and Biobanking* (pp. 31-42). Springer, Cham.

Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., ... and McKinney, E. F. (2019). From big data to precision medicine. *Frontiers in medicine*, 34.

Hunter, R. F., Gough, A., O'Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., ... and Kee, F. (2018). Ethical issues in social media research for public health. *American Journal of Public Health*, *108*(3), 343-348.

Ienca, M., Vayena, E., and Blasimme, A. (2018). Big data and dementia: charting the route ahead for research, ethics, and policy. *Frontiers in medicine*, *5*, 13.

Izmailova, E. S., Wagner, J. A., and Perakslis, E. D. (2018). Wearable devices in clinical trials: Hype and hypothesis. *Clinical Pharmacology and Therapeutics, 104*(1), 42–52.

Janeček, V. (2018). Ownership of personal data in the Internet of Things. Computer law and security review, 34(5), 1039-1052.

Juengst, E. T., and Meslin, E. M. (2019). Sharing with Strangers: Governance Models for Borderless Genomic Research in a Territorial World. *Kennedy Institute of Ethics Journal*, *29*(1), 67-95.

Kayaba K. (2013). Overcoming the difficulties of cohort studies. *Journal of epidemiology*, *23*(3), 156–157. https://doi.org/10.2188/jea.je20120225

Kaye, J., Briceño Moraia, L., Mitchell, C., Bell, J., Bovenberg, J. A., Tassé, A. M., and Knoppers, B. M. (2016). Access governance for biobanks: the case of the BioSHaRE-EU cohorts. *Biopreservation and Biobanking*, *14*(3), 201-206.

Keane, T. M., O'Donovan, C., and Vizcaíno, J. A. (2021). The growing need for controlled 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

access models in clinical proteomics and metabolomics. *Nature Communications*, *12*(1), 1-4.

Kent, S., Burn, E., Dawoud, D., Jonsson, P., Østby, J. T., Hughes, N., ... and Bouvy, J. C. (2021). Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics*, *39*(3), 275-285.

Khazaal, Y., Chatton, A., Bouvard, A., Khiari, H., Achab, S., and Zullino, D. (2013). Internet poker websites and pathological gambling prevention policy. *Journal of Gambling Studies*, *29*(1), 51-59.

Kiehntopf, M. (2019). Biobanking–current questions and positions. *Journal of Laboratory Medicine*, *43*(6), 287-290.

Kim, K. H., Choi, W., Ko, S. J., Chang, D. J., Chung, Y. W., Chang, S. H., ... and Choi, I. Y. (2021). Multi-Center Healthcare Data Quality Measurement Model and Assessment Using OMOP CDM. *Applied Sciences*, *11*(19), 9188.

Kirwan, M., Mee, B., Clarke, N., Tanaka, A., Manaloto, L., Halpin, E., ... and McElvaney, N. G. (2021). What GDPR and the Health Research Regulations (HRRs) mean for Ireland:"explicit consent"—a legal analysis. Irish Journal of Medical Science (1971-), 190(2), 515-521.

Kitchin, R. (2020). Civil liberties or public health, or civil liberties and public health? Using surveillance technologies to tackle the spread of COVID-19.*Space and Polity*, *24*(3), 362-381.

Lehner-Mear, R. (2020). Negotiating the ethics of Netnography: developing an ethical approach to an online study of mother perspectives. *International Journal of Social Research Methodology*, *23*(2), 123-137.

Lidz, C. W., and Appelbaum, P. S. (2002). The therapeutic misconception: problems and solutions. *Medical care*, V55-V63.

Manrique de Lara, A., and Peláez-Ballestas, I. (2020). Big data and data processing in rheumatology: bioethical perspectives. *Clinical Rheumatology*, *39*(4), 1007-1014.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, *19*(2), 286-302.

Marelli, L., Lievevrouw, E., and Van Hoyweghen, I. (2020). Fit for purpose? The GDPR and the governance of European digital health. Policy studies, 41(5), 447-467.

Martinez-Martin, N., and Kreitmair, K. (2018). Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. *JMIR mental health*, *5*(2), e9423.

Martinez-Martin, N., Insel, T. R., Dagum, P., Greely, H. T., and Cho, M. K. (2018). Data mining for health: staking out the ethical territory of digital phenotyping. NPJ digital medicine, 1(1), 1-5.

McLennan, S., Shaw, D., and Celi, L. A. (2019). The challenge of local consent requirements for global critical care databases. *Intensive care medicine*, *45*(2), 246-248.

McMahon, C., and Denaxas, S. (2019). A novel metadata management model to capture consent for record linkage in longitudinal research studies. *Informatics for Health and Social Care*, *44*(2), 176-188.

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

McRae, L., Ellis, K., Kent, M., and Locke, K. (2020). Privacy and the ethics of disability research: Changing perceptions of privacy and smartphone use. *Second international handbook of internet research*, 413-429.

Menikoff, Jerry (2001). *Just Compensation: Paying Research Subjects Relative to the Risks They Bear. American Journal of Bioethics, 1(2), 56–58.* doi:10.1162/152651601300169121

Merchant, R. M., and Lurie, N. (2020). Social media and emergency preparedness in response to novel coronavirus. *Jama*.

Miao, L., Zhang, J., Yi, L., and Huang, S. (2020). The Ethical, Legal, and Regulatory Issues Associated with Pharmacogenomics. In *Pharmacogenomics in Precision Medicine* (pp. 219-239). Springer, Singapore.

Minari, J., Brothers, K. B., and Morrison, M. (2018). Tensions in ethics and policy created by National Precision Medicine Programs. *Human genomics*, *12*(1), 1-10.

Mitchell, C., Ordish, J., Johnson, E., Brigden, T., and Hall, A. (2020). The GDPR and genomic data—the impact of the GDPR and DPA 2018 on genomic healthcare and research. *PHG Foundation*.

Molnár-Gábor, F., and Korbel, J. O. (2020). Genomic data sharing in Europe is stumbling—Could a code of conduct prevent its fall?. *EMBO Molecular Medicine*, *12*(3), e11421.

Mozersky, J., Walsh, H., Parsons, M., McIntosh, T., Baldwin, K., and DuBois, J. M. (2020). Are we ready to share qualitative research data? Knowledge and preparedness among qualitative researchers, IRB Members, and data repository curators. *IASSIST quarterly*, *43*(4), 952.

Muzny, M., Henriksen, A., Giordanengo, A., Muzik, J., Grøttland, A., Blixgård, H., ... and Årsand, E. (2020). Wearable sensors with possibilities for data exchange: Analyzing status and needs of different actors in mobile health monitoring systems. *International journal of medical informatics*, *133*, 104017.

National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and replicability in science.

Nebeker, C., Torous, J., and Ellis, R. J. B. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC medicine*, *17*(1), 137.

Ning, M., and Lo, E. H. (2010). Opportunities and challenges in omics. *Translational stroke research*, *1*(4), 233–237.

Nurmi, S. M., Kangasniemi, M., Halkoaho, A., and Pietilä, A. M. (2019). Privacy of clinical research subjects: an integrative literature review. *Journal of Empirical Research on Human Research Ethics*, *14*(1), 33-48.

O'Neill, D., Benzeval, M., Boyd, A., Calderwood, L., Cooper, C., Corti, L., ... and Park, A. (2019). Data resource profile: cohort and longitudinal studies enhancement resources (CLOSER). *International journal of epidemiology*, *48*(3), 675-676i.

OECD, Health in the 21st Century: Putting Data to Work for Stronger Health Systems, OECD Health Policy Studies, OECD Publishing, Paris; 2019. Available at: https://doi.org/10.1787/e3b23f8e-en.

Ogunseye, S., Parsons, J., and Afolabi, D. (2021, August). Training-Induced Class Imbalance in Crowdsourced Data. InVLDB 2021 Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale.

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

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

Pagallo, U. (2021). On the principle of privacy by design and its limits: Technology, ethics and the rule of law. In *Italian Philosophy of Technology* (pp. 111-127). Springer, Cham.

Peloquin, D., DiMaio, M., Bierer, B., and Barnes, M. (2020). Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics*, 28(6), 697-705.

Pezoulas, V., Exarchos, T., and Fotiadis, D. I. (2020). *Medical data sharing, harmonization and analytics*. Academic Press.

Platt, R. W., Henry, D., and Suissa, S. (2019). The Canadian Network for Observational Drug Effect Studies (CNODES): Reflections on the first eight years, and a look to the future.

Price, W. N., and Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature medicine*, *25*(1), 37-43.

Resnik, D. B. (2019). Citizen scientists as human subjects: Ethical issues. *Citizen Science: Theory and Practice*, *4*(1).

Richterich, A. (2018). *The big data agenda: Data ethics and critical data studies* (p. 154). University of Westminster Press.

Robertson, R. John (2005). Metadata quality: implications for library and information science professionals. Library Review, 54(5), 295–300. doi:10.1108/00242530510600543

Rudy, J., and Valafar, F. (2011). Empirical comparison of cross-platform normalization methods for gene expression data. *BMC bioinformatics*, *12*(1), 1-22.

Salokannel, M., Tarkkala, H., and Snell, K. (2019). Legacy samples in Finnish biobanks: social and legal issues related to the transfer of old sample collections into biobanks. *Human Genetics*, *138*(11), 1287-1299.

Sanchez-Niubo, A., Tyrovolas, S., Moneta, M., Prina, M., Panagiotakos, D., Caballero, F., and Fortier, I. (2017). DATA HARMONIZATION OF LONGITUDINAL STUDIES ON HEALTHY AGEING: THE ATHLOS PROJECT. *Innovation in Aging*, *1*(Suppl 1), 1315. https://doi.org/10.1093/geroni/igx004.4818

Schneeweiss, S., Brown, J. S., Bate, A., Trifirò, G., and Bartels, D. B. (2020). Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. *Clinical Pharmacology and Therapeutics*, *107*(4), 827-833.

Servoli, L., Meroli, S., Passeri, D., and Tucceri, P. (2013). Measurement of submicrometric intrinsic spatial resolution for active pixel sensors. *Journal of Instrumentation*, *8*(11), P11007.

Shabani, M. (2021). The Data Governance Act and the EU's move towards facilitating data sharing. *Molecular systems biology*, *17*(3), e10229.

Shaban-Nejad, A., Michalowski, M., and Buckeridge, D. L. (2018). Health intelligence: how artificial intelligence transforms population and personalized health. *NPJ digital medicine*, *1*(1), 1-2.

Shadbolt, N., O'Hara, K., De Roure, D., and Hall, W. (2019). *The theory and practice of social machines*. New York: Springer International Publishing.

Slokenberga, S., Tzortzatou, O., and Reichel, J. (2021). *GDPR and biobanking: Individual rights,*

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

public interest and research regulation across Europe (p. 434). Springer Nature.

Stommel, W., and Rijk, L. D. (2021). Ethical approval: none sought. How discourse analysts report ethical issues around publicly available online data. *Research Ethics*, *17*(3), 275-297.

Stone, C. J., Skinner, A. L., and Unit, M. I. E. (2017). New technology and novel methods for capturing health-related data in longitudinal and cohort studies. In *Abstracts of presentations at the knowledge exchange workshop*

Taylor, J., and Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges?. *Research Ethics*, *14*(2), 1-39.

Tene, O., and Polenetsky, J. (2012). To track or do not track: advancing transparency and individual control in online behavioral advertising. *Minn. JL Sci. and Tech.*, *13*, 281.

Townend, D. (2018). Conclusion: harmonisation in genomic and health data sharing for research: an impossible dream?. *Human genetics*, *137*(8), 657-664.

Umbach, N., Beißbarth, T., Bleckmann, A., Duttge, G., Flatau, L., König, A., ... & Schweda, M. (2020). Clinical application of genomic high-throughput data: Infrastructural, ethical, legal and psychosocial aspects. European Neuropsychopharmacology, 31, 1-15

Van Der Wel, K. A., Östergren, O., Lundberg, O., Korhonen, K., Martikainen, P., Andersen, A. M. N., and Urhoj, S. K. (2019). A gold mine, but still no Klondike: Nordic register data in health inequalities research. *Scandinavian Journal of Public Health*, *47*(6), 618-630.

van Veen, E. B. (2018). Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. *European Journal of Cancer*, *104*, 70-80.

Vergallo, G. M., Shapiro, D. L., Walker, L. E., Mastronardi, V., Calderaro, M., Ferrer, C. I. S., ... and Zaami, S. (2020). Health care providers ethical use of risk assessment to identify and prevent terrorism. *Ethics, Medicine and Public Health*, *12*, 100436.

Verhenneman, G., Claes, K., Derèze, J. J., Herijgers, P., Mathieu, C., Rademakers, F. E., ... and Vanautgaerden, M. (2020). How GDPR enhances transparency and fosters pseudonymisation in academic medical research. European Journal of Health Law, 27(1), 35-57.

White, E., Hunt, J. R., and Casso, D. (1998). Exposure measurement in cohort studies: the challenges of prospective data collection. *Epidemiologic reviews*, *20*(1), 43-56.

Wiggins, A., and Wilbanks, J. (2019). The rise of citizen science in health and biomedical research. *The American Journal of Bioethics*, *19*(8), 3-14.

Wilson, R. C., Butters, O. W., Avraam, D., Baker, J., Tedds, J. A., Turner, A., ... and Burton, P. R. (2017). DataSHIELD–new directions and dimensions.

Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V et al. Smart Medical Information Technology for Healthcare (SMITH): Data Integration based on Interoperability Standards. Methods of information in medicine; 2018., 57(Suppl 1), e92.

Wolf, S. M., Ossorio, P. N., Berry, S. A., Greely, H. T., McGuire, A. L., Penny, M. A., and Terry, S. F. (2020). Integrating rules for genomic research, clinical care, public health screening and DTC testing: creating translational law for translational genomics. *Journal of Law, Medicine and Ethics*, *48*(1), 69-86.

Zondergeld, J. J., Scholten, R. H., Vreede, B. M., Hessels, R. S., Pijl, A. G., Buizer-Voskamp, J. E., ...

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

and Veldkamp, C. L. (2020). FAIR, safe and high-quality data: The data infrastructure and accessibility of the YOUth cohort study. *Developmental cognitive neuroscience*, *45*, 100834.

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