Environment International 165 (2022) 107334

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Position paper on management of personal data in environment and health research in Europe

Govarts Eva [a,*], Gilles Liese [a], Bopp Stephanie [b], Holub Petr [c], Matalonga Leslie [d], Vermeulen Roel [e], Vrijheid Martine [f,g,h], Beltran Sergi [d,h,i], Hartlev Mette [j], Jones Sarah [k], Rodriguez Martin Laura [a], Standaert Arnout [a], Swertz Morris A. [l], Theunis Jan [a], Trier Xenia [m], Vogel Nina [n], Van Espen Koert [o], Remy Sylvie [a], Schoeters Greet [a,p]

[a] VITO Health, Flemish Institute for Technological Research (VITO), Mol, Belgium
[b] European Commission, Joint Research Centre (JRC), Ispra, Italy
[c] BBMRI-ERIC, Graz, Austria
[d] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
[e] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, Netherlands
[f] ISGlobal, Barcelona, Spain
[g] Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Spain
[h] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[i] Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain
[j] Faculty of Law, University of Copenhagen, Copenhagen, Denmark
[k] GÉANT, Glasgow, United Kingdom
[l] Department of Genetics & Genomics Coordination Center, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[m] European Environment Agency (EEA), Copenhagen, Denmark
[n] German Environment Agency (UBA), Berlin, Germany
[o] Apogado CVBA, Mechelen, Belgium
[p] Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

ARTICLE INFO

Handling Editor: Adrian Covaci

Keywords:
Data protection
Biomonitoring data
Health data
FAIR principles
GDPR

ABSTRACT

Management of datasets that include health information and other sensitive personal information of European study participants has to be compliant with the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679). Within scientific research, the widely subscribed'FAIR' data principles should apply, meaning that research data should be findable, accessible, interoperable and re-usable. Balancing the aim of open science driven FAIR data management with GDPR compliant personal data protection safeguards is now a common challenge for many research projects dealing with (sensitive) personal data.

In December 2020 a workshop was held with representatives of several large EU research consortia and of the European Commission to reflect on how to apply the FAIR data principles for environment and health research (E&H). Several recent data intensive EU funded E&H research projects face this challenge and work intensively towards developing solutions to access, exchange, store, handle, share, process and use such sensitive personal data, with the aim to support European and transnational collaborations. As a result, several recommendations, opportunities and current limitations were formulated.

New technical developments such as federated data management and analysis systems, machine learning together with advanced search software, harmonized ontologies and data quality standards should in principle facilitate the FAIRification of data. To address ethical, legal, political and financial obstacles to the wider re-use of data for research purposes, both specific expertise and underpinning infrastructure are needed. There is a need for the E&H research data to find their place in the European Open Science Cloud. Communities using health and population data, environmental data and other publicly available data have to interconnect and synergize. To maximize the use and re-use of environment and health data, a dedicated supporting European infrastructure effort, such as the EIRENE research infrastructure within the ESFRI roadmap 2021, is needed that would interact with existing infrastructures.

* Corresponding author at: VITO Health, Flemish Institute for Technological Research (VITO), Boeretang 200, 2400 Mol, Belgium.
E-mail address: eva.govarts@vito.be (G. Eva).

https://doi.org/10.1016/j.envint.2022.107334
Received 22 December 2021; Received in revised form 30 May 2022; Accepted 1 June 2022
Available online 6 June 2022

1. Introduction

1.1. Changing research and innovation landscape

In recent years there has been growing attention for the open science movement, which advocates a new open way of data governance. Practicing open science requires four fundamental concepts: open access, open data, open source, and open standards. It's an attempt to offer unfettered dissemination of scientific discourse and return. This should enable reproducible science giving full access to the major components of scientific research (Schroeder 2013). Each one of the open science concepts is necessary to realize the core aims of sharing results and stimulating scientific progress. Open data provides the opportunity for verification, analysis and subsequent publication of scientific results in new forms. Open source embodies scientific methods, so that new computational processes can be independently examined, reviewed and reused. Open access facilitates the review and validation of research processes and results. Finally, open standards, while not absolutely essential to open science, simplify the process of exchanging data, methods and publications thereby accelerating the research process (Schroeder 2013).

On a European level, the open science movement is supported by the European Data Strategy (European Commission 2020) published in February 2020 and more specifically for research data by the European Open Science Cloud initiative (EOSC) (European Commission 2018a). EOSC provides "a trusted environment for sharing and analyzing data from all publicly funded research" (European Commission 2018b). The ultimate goal of EOSC is to achieve a fundamental transformation of the whole research lifecycle and to make it more credible with increased integrity, more efficient, collaborative, transparent and more responsive to societal challenges. A fundamental prerequisite of the EOSC is that data shall be FAIR (Findable, Accessible, Interoperable and Re-usable). The FAIR principles provide a set of 15 guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets (Wilkinson et al. 2016). The FAIR principles imply that access to data should be technically possible and legally permitted, hence not necessarily open by default. One of the principles states that (meta)data are to be released with a clear and accessible data usage license, but it does not dictate any type of license nor who can reuse the data, under which conditions and controlled by whom.

Besides increased efforts towards open science data, protection of individuals' fundamental rights in the digital age has also gained priority. On May 25, 2018 the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) was put into effect by which the European Parliament, the European Council and the European Commission intend to strengthen and unify the protection of personal data within the European Union (EU). Personal data is any information that relates to an individual who can directly or indirectly be identified. If someone processes personal data, there are the seven data protection and accountability principles outlined in Article 5.1–2: lawfulness, fairness and transparence, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability. It's required to handle data securely by implementing appropriate technical and organizational measures. Moreover, there are strict rules about what constitutes consent from a data subject to process their information. As such data subjects have the following privacy rights: the right to be informed, the right of access, the right of rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object and rights in relation to automated decision making and profiling. This is just a summary of the main points of the GDPR, the regulation itself (not including the accompanying directives) is 88 pages (GDPR, Regulation (EU) 2016/679).Aiming for FAIR data and open science, while respecting the GDPR, is a huge challenge for many research projects dealing with sensitive personal data. Issues arise not only from legal, ethical, political and intellectual property perspectives, but also from the lack of necessary resources including technical enablers (like automated/semi-automated access control tools) needed to tackle these issues. EU projects, working with personal data, including sensitive personal data defined as special categories of personal data in Art. 9 and 10 of the GDPR, all face these challenges, and several had to reconsider and redesign their procedures to implement the FAIR data and open science principles whilst respecting the GDPR. Different research fields demonstrate diverse levels of maturity in sharing health data transnationally. Some areas have historically paved the way i.e. genomic data in the areas of rare diseases or cancer research, such as the European research project Solving the unsolved Rare Diseases (Solve-RD) (Zurek et al. 2021) and the International Cancer Genome Consortium (ICGC) (Zhang et al. 2019). Noteworthy, initiatives like the Global Alliance for Genomics and Health (GA4GH) aim to enable responsible sharing of genomic data for the benefit of human health, developing policy recommendations and standards (Rehm et al. 2021). In Europe, one of the main goals of ELIXIR (Electronic LIbrary eXchange for Information Resources) is to coordinate, integrate and sustain European bioinformatics resources to easily find, share and analyze data (https://elixir-europe.org/).

On December 7, 2020, a virtual workshop with representatives of DG Research & Innovation, DG Environment and the Joint Research Centre, representing the European Commission Information Platform for Chemical Monitoring (IPCHEM), reflected on how to apply the FAIR data principles in the context of open science and GDPR with representatives of 6 large EU research consortia: the European Human Biomonitoring Initiative (HBM4EU), the European Child Cohort Network (LifeCycle), the federated FAIR platform for cohort data connecting Europe and Canada (EUCAN-Connect), the Biobanking and BioMolecular Resources Research Infrastructure — European Research Infrastructure Consortium (BBMRI-ERIC), the European Joint Program on Rare Diseases (EJP-RD) and the European Human Exposome Network (EHEN). From their experiences and lessons learnt, we aim to reach a common understanding and formulate opportunities, current limitations and barriers, and recommendations how personal data should be dealt with in future research initiatives or projects to promote maximal use of research data across borders and across disciplines in support of EU policy making and international cooperation.

1.2. Personal data in E&H research

Research projects in the environment & health (E&H) area and health area collect comprehensive personal data, containing information about chemical exposure levels, lifestyle, food consumption, behavior, socio-demographics, residence, occupation, health status, etc. This includes sensitive personal data that is considered as special categories of personal data according to the GDPR, like data on ethnic origin and health-related data. Several techniques exist to anonymize personal data such as data masking, data swapping, generalization, data perturbation, data aggregation and toolkits for anonymization of data are readily available (Corporate Finance Institute 2020). However, anonymous data, to the highest standards without any residual risk for re-identification, will lose much of their value for nuanced research (Dwork 2014; van Veen 2018). It can destroy the value that data holds as, for example, data aggregation lowers the resolution of the data, which in turn reduces the analysis potential of the data, or the noise is modulated to minimize risk of inferring information about data subjects. Although much can be learned from aggregated data (anonymous in most cases) some research questions require more in-depth analysis of individual level personal data e.g., studying exposure-effect associations, studying exposure to chemical mixtures at individuals' level, linking exposure data to individual health outcomes etc.

Furthermore, anonymization can deprive a data subject of its rights as stipulated in the GDPR, e.g., the right of access to data by the data subject and the right to data portability, which can no longer be realized. Implementing a policy for handling incidental findings is virtually impossible for anonymized data. Anonymization also breaks the

provenance link needed for full traceability of medically relevant research. Moreover, anonymization is not always possible if human samples containing DNA for example are biobanked. Furthermore, in longitudinal cohorts in which participants are followed up over the years, the identification keys should be retained to allow follow up, thus making anonymization impossible. However, anonymization is not only a matter of destroying the key. The resulting anonymized data need to be resilient to other attacks beyond just singling out persons, such as linking two records (within the same or different datasets), or inferring personal information not present in the dataset. So simply dropping the keys may or may not be sufficient. The process of anonymizing data is a trade-off between utility (usefulness of the data) and privacy protection.

As a result, pseudonymization of personal data in E&H research is often preferred over anonymization. Pseudonymization means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific person without the use of additional information. Such additional information must be kept carefully separate from the personal data. Processing of pseudonymous personal data is subject to the GDPR, which implies a legal basis for processing the data subject's personal data should be in place. In E&H research, this legal basis could be covered by the informed consent (IC) form, in which the explicit consent of the data subject is requested to use and share data on EU level. Note, however, that informed consent, as defined in the Declaration of Helsinki, and "explicit consent" as defined as a legal basis under the GDPR, are different concepts that are often not compatible. One of the reasons being, that explicit consent needs to be freely given, meaning it must be given on a voluntary basis, which often cannot be guaranteed (Art. 7 of the GDPR). Consent is only one of six bases mentioned in the GDPR. The others are contract, legal obligations, vital interests of the data subject, public interest and legitimate interest as stated in Art. 6(1) of the GDPR. Hence, in E&H research, another legal basis than consent – such as public interest – is often more appropriate (note that "public interest" also needs to match specific requirements). This is also the case for data coming from older research projects where explicit consent is lacking. However, there is also a legal basis for secondary use of health data for scientific research as specified in Art. 9(2) of the GDPR (European Commission 2021).

Even pseudonymization has an impact on utility of the data, e.g., linking the data across different data sources becomes much more difficult and possible only if specific measures are taken upfront when pseudonyms are generated. Hence any applications of privacy enhancing technologies, including anonymization and pseudonymization, need to be considered by design as a part of FAIR implementation strategy as guided by FAIR-Health (Holub et al. 2018).

1.3. Data management and analysis of E&H data

Over the last 15 years different models were used for data sharing in European E&H projects (Table 1), evolving from central analysis with physical exchange of the data between multiple parties (model 1) or local analysis and central *meta*-analysis (model 2) to centralized systems where data is stored centrally at hosting facility (model 3) towards federated systems (model 4) or hybrid systems combining federated data access and centralization of the data (model 5). Federated data management and analysis allows analyzing the data while not physically exchanging them. Currently, there is a rapid expansion of algorithms that are generated allowing to perform the statistical analyses on a virtually pooled database using federated analysis systems (Kholod et al. 2020; Li et al. 2020; Zerka et al. 2020). Several statistical models/techniques are implemented in these systems, such as generalized linear models, etc. However, there is still progress to be made to develop algorithms implementing more advanced statistical techniques, like data mining, mixture analyses techniques (e.g. elastic net, Bayesian Adaptive Sampling) and machine learning. For this type of data analysis, centralized systems in which the data are brought together in one place, still offer more flexibility. Furthermore, there should be a balance between resources required and incentives to implement federated management & analysis systems. Servers have to be installed at all institutes, and very powerful computers are needed for federated analysis. This comes along with financial and technical expertise needs. So, for research projects having the capacity to invest in such a system but also requiring more complex analyses, a hybrid system (model 5), combining a centralized and a decentralized approach seems most apt to analyze the data fit for purpose.

Moreover, linking personal data with environmental data systems would be of high interest in the E&H research domain. This could be done by mapping human exposure data of a certain geospatial area to environmental data of the same area using for example Nomenclature of Territorial Units for Statistics (NUTS) information. However, it would be interesting to do this at even higher resolution level, using the individual geospatial coordinates of the subjects. GDPR does not allow, however, to share this information with external partners without a proper legal basis. Solutions are currently being developed within the framework of the Personal Health Train where linkages between data sources, including environmental data, can be done without exchanging privacy information (Beyan et al. 2020).

There is a recent shift in the data management landscape towards a situation where the citizens are in control of their own personal data (Mirchev et al. 2020). For example, the decentralized SOLID (https://solidproject.org/) technology gives individuals a personal vault to store rich personal data along with the ability to control access permission to this data, resulting in true personal data control as well as improved privacy. Combining this with dynamic informed consent (Dankar et al. 2020) offers the opportunity to continuously inform participants about research protocols and support the participant's autonomy and decision-making power. However, this puts a burden on citizens to get familiar and well informed about the research demands. Making people their own data custodian, may however result in problems, such as the available data not reflecting the total population, under-representing vulnerable groups such as lower socioeconomic status, minorities or elderly.

2. FAIR principles

2.1. Findable

The first step in facilitating (re-)use of data, is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRificaton process.

Data collections should be easy to identify and locate. This is also fully compatible with the open science concept of open data. This can be done through the open publication of metadata (a set of descriptive elements providing information on the container of the data) in a searchable resource. Both the data and the metadata should be assigned a unique and persistent identifier. The metadata information should be detailed enough, standardized to the extent possible, quality controlled and kept up to date to exploit its full potential.

Different dedicated libraries or platforms publish available metadata such as i) IPCHEM (https://ipchem.jrc.ec.europa.eu/), run by JRC and hosting, among others, human biomonitoring (HBM) metadata in collaboration with HBM4EU and the European Environment Agency (EEA), ii) BBMRI-ERIC-directory and locator (https://directory.bbmri-eric.eu/), iii) Birthcohorts.net (https://www.birthcohorts.net/), iv) HealthyCloud (https://healthycloud.eu/). In the rare disease field, EJP-RD is developing a Virtual Platform to allow queries across resources with different data types (e.g. patient registries, biobanks, genomics, mouse models, cell lines, etc.) under the same metadata model (https://resourcemap.ejprarediseases.org/#/). Each platform has its focus, presenting metadata in the chemical & environment area, the health area or a combination of both. A connection between platforms to

Table 1
Overview of models for data sharing

	Data hosting location	Access Authorization	Data harmonization procedure	Data Quality Control procedure	(re)-Use	Scalability
Model 1 Central analysis • ENRIECO (Casas et al. 2015) • CHICOS (Birks et al. 2016; Sonnenschein-van der Voort et al. 2014) • HBM4EU data analyses based on existing studies	**Data is stored at different locations, physical exchange of data copies.** ⊖ Data stored by data custodian but also by different users, difficult to keep track of all locations where copies of the data are kept. ⊖ Many physical exchanges of data, GDPR issues ⊕ Most easy way forward in case no trust established yet for a centralized hosting facility.	Bilateral data exchange after signing data transfer agreement	Harmonization done by data custodian on demand of data user. ⊖ Each analysis/paper requires own harmonization (not efficient)	Done by data custodian. No external QC system. ⊖ Additional effort for data user to check and possibly correct the data	Pooling of data needs to be done by central data analyst. ⊖ Re-use of data is mostly not covered (not included in the agreements) ⊕ Data user has access to individual data → flexibility in data analysis	Sending data sets bilateral with multiple parties is not a sustainable solution for the future. Difficult to trace for data custodian if for example a participant wants to withdraw its data.
Model 2 Local analysis & central meta-analysis • PACE (Felix et al. 2018) • ENRIECO (Govarts et al. 2012)	**Data is kept local with data custodians** ⊕ No actual transfer of individual data required, GDPR compliant	Summary statistics are provided (anonymous data) → no GDPR issues, open access possible	Meta-analysis based on published results: no harmonization required. Meta-analysis based on requested results: harmonization done by data custodian on demand of data user. ⊖ Each analysis/paper requires own harmonization (not efficient)	Done by data custodian ⊖ Not possible for data user to check the data	⊖ Limited to meta-analysis on summary statistics. ⊖ Meta-analysis might not be an optimal way of analyzing the data.	Requires resources and experience from data custodian to analyze the data themselves.
Model 3 Centralized system • COPHES (Den Hond et al. 2015) • HELIX (Maitre et al. 2018) • ESCAPE (Beelen et al. 2014) • ELAPSE (Strak et al. 2021) • HBM4EU co-funded studies (Gilles et al. 2021) • EXPANSE (Vlaanderen et al. 2021) • EJP-RD (genome-phenome data resources) • IPCHEM	**Data is stored centrally at hosting facility** ⊖ Physical data transfer required, GDPR issues ⊖ Hosting facility needs to become data processor ⊕ One central master version of the dataset	Data transfer agreements needed. Different national legislations may impede centralized solutions	Harmonization done by data custodian ⊕ Harmonization can be done for different analyses/papers → more efficient than model 1 and 2	Can be done by hosting facility. ⊕ Techniques to automatically identify anomalies in the data.	Pooling of data done by hosting facility. ⊕ Data user has access to individual data → flexibility in data analysis	Need of trust by data custodians in central deposit. Resources needed for hosting and maintenance. Entire dynamic cohort datasets are better maintained by the cohorts themselves. Centralized system suitable for specific analyses in a specific project at a specific time point.
Model 4 Federated System • LifeCycle (Jaddoe et al. 2020) • European Human Exposome Network (EHEN)	**Data is kept local with data custodian** ⊖ Requires substantial initial investment in infrastructure, know-how, time. ⊕ No physical data transfer required, GDPR compliant ⊕ Data custodian keeps control of the dataset	Approval by cohort (DAA = data access agreement) Machine readable access rights can be defined	Harmonization done by data custodian following agreed protocols. ⊕ Efficient harmonization is done once.	Can be organized on project level. ⊕ Techniques to automatically identify anomalies in the data.	⊖ Limited to current possibilities with federated analysis ⊕ Rapidly evolving	Suitable for large dynamic datasets such as longitudinal cohorts. Resources needed for infrastructure costs, as capacity and know-how for all involved data custodians and data users.
Model 5 Hybrid system/ Future visions	**Combination of data kept local with data custodian and accessed via federated systems and data stored centrally at hosting**	Automated authorization and access control based on metadata info.	Harmonization done by data custodian following agreed protocols. ⊕ Efficient harmonization is done	⊕ Techniques to automatically identify anomalies in the data.	⊕ Central data analysis for more complex analysis not (yet) possible with federated analysis systems.	Resources needed for infrastructure costs in case of federated data sharing, but also possibility for data custodians to share data

(*continued on next page*)

Table 1 (*continued*)

	Data hosting location	Access Authorization	Data harmonization procedure	Data Quality Control procedure	(re)-Use	Scalability
• ATHLETE (Vrijheid et al. 2021)	**facility** ⊖ Hosting facility needs to become data processor ⊖ Partly physical data transfer, GDPR issues ⊕ Partners with no capacity/know-how can entrust their data to a central hosting facility ⊕ Data users get access to the data via federated system		once.		⊕ Federated data analysis for standard statistical analyses.	physically if no capacity/ know-how.

Advantages/positive aspects visualized by ⊕ symbol, disadvantages/limitations visualized by ⊖ symbol.
Abbreviations: GDPR, General Data Protection Regulation; QC, Quality Control

retrieve all relevant information from specific data collections would be needed. This requires that metadata should be searchable. Open source search tools such as Molgenis (https://www.molgenis.org) (van der Velde et al. 2019) or Maelstrom (https://www.maelstrom-research.org/maelstrom-catalogue) (Bergeron et al. 2018) can be used to create searchable platforms such as described above (e.g. BBMRI-ERIC Directory) (Holub et al. 2016), to query networks, studies, and variables and to explore harmonization potential across studies. Further examples are YODA which is the platform developed at Utrecht University, The Netherlands (https://www.uu.nl/en/research/yoda). This is an open source data management platform allowing for local storage of data under GDPR and to have metadata linked. Also, the European LifeCycle project created a metadata catalogue building on the Molgenis software platform (Jaddoe et al. 2020). The catalogue shows which variables are harmonized and how it is done, well documented per cohort. In HBM4EU, IPCHEM was used to make the HBM data findable. The sharing of genomics data and the ELIXIR model on -omics data apply a federated data model, using the Beacon standard for data discovery (https://beacon-network.org/#/). The personal data remain at the local cohort database and the metadata is shared. Federated searching, also known as meta searching, broadcast searching or cross-searching, is the ability to search multiple information resources from a single interface and return an integrated set of results.

2.2. Accessible

Once a user has found the required data, it should be clear if and how data can be accessed, possibly including authentication and authorization.

Preferentially, metadata should indicate the data access rights and conditions of a dataset, ideally in a machine-readable representation such as Data Use Ontology (https://www.ebi.ac.uk/ols/ontologies/duo) (Lawson et al. 2021). If regulation changes those access rights and conditions may change as well. So, it is never a static system as the ethical, legal, societal issues (ELSI) may change over time. The use and exchange of personal sensitive data needs to be GDPR compliant. Until now, multiple data access agreements need to be established between each data custodian and data user, which results in a huge administrative process, certainly in big European projects like HBM4EU. In Table 1 an overview is given of models used for data sharing and the accompanying access authorization in different EU E&H projects. In recent years, some projects have been working towards machine readable access rights, e.g. in EHEN and ATHLETE (Table 1). By automated access control, this administrative burden could be reduced.

It is the data custodian that sets the FAIR implementation framework (FIF) and FAIR digital object framework (FDOF) (Bonino da Silva Santos 2021). Access rights are based on the ICs. Often the responsibility to provide access to the study data are laid down at the level of the data custodian but must be in line with the ICs. In BBMRI-ERIC each of the biobanks have their own consents and share and communicate their best practices for future development of the IC. To get access to data, one should go back to the owners of the biobanks. Access rights may however evolve over time. ICs and hence access rights can also be differentiated e.g., separate for clinical care and for other data etc.

In all cases only the data that are needed for statistical analysis should be made accessible as data minimization principles need to be applied to maximally protect an individual's privacy. Besides the permission to access the pseudonymized data, also technical solutions are needed to provide access. This requires several considerations by data custodians such as secure data storage and data exchange (authentication, authorization). In the last years, a lot of effort has gone into the development of decentralized systems in which the data are kept at the local level but can be accessed remotely to develop a federated or virtual database as expressed by "bringing the analysis to the data instead of the data to the analysis". Federated access systems require that each of the local data hosting institutes install specific web server applications such as DataSHIELD (Gaye et al. 2014) to make the data accessible while they stay at the data host's institute. Federated database infrastructures offer an alternative to the physical exchange of dataset copies with data users (no downloads). This enables a data custodian to better monitor what happens with their dataset as the dataset never leaves the hosting institutes. For reproducibility of results, version control should be embedded throughout.

2.3. Interoperable

The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.

Within E&H research, the data needs to be integrated or pooled with other datasets for maximizing their use, for example to visualize them for different users and stakeholders, or through joint statistical analysis at EU level. Over the years, the need to access and use different datasets has been growing. The advantages are obvious as shown by research on rare disease or to find genetic disease markers (Lochmüller et al. 2018; Matalonga et al. 2021; Zurek et al. 2021). Bringing data together, and thereby increasing sample sizes and statistical power, has allowed big progress in these fields. The same is true for the E&H research field, where more robust results can be found by combining datasets. Different models and approaches are possible with consequences for dealing with privacy issues and data analysis (Table 1). An evolution has taken place, from projects applying central data analysis after bilateral data exchange (model 1) or not exchanging individual data and performing local analysis (model 2), towards centralized (model 3) or federated systems (model 4). Some of them having applied a combination of the latter two the so-called hybrid systems (model 5). To facilitate the

exchange of data, data harmonization is key. This includes, but is not limited to, the creation of common vocabularies, ontologies, standards and formats, considering the specific requirements of the research questions in the respective field (Kush et al. 2020; Wey et al. 2021). Currently, such established ontologies are lacking for environment and health data. Projects such as LifeCycle and HBM4EU developed scripts that document harmonization in detail also including the quality control steps (Gilles et al. 2021; Jaddoe et al. 2020). This requires significant upfront effort but will make interoperability and hence re-use of the data easier. Cataloging systems as mentioned earlier (Molgenis or Maelstrom) allow to search for information not only in metadata but also beyond. Beneficial use of interoperable data requires not only stringent quality control of the data itself, but also of the previous parts of the chain such as data provenance, data preparation but also sample acquisition. This should be documented in an accessible way to prevent misinterpretation and misuse of the data.

Being able to connect datasets and connect to the large data infrastructures at EU level should be reflected on when writing a proposal, e.g. by including a data stewardship plan.

2.4. Re-usable

The ultimate goal of FAIR data is to increase data re-use. To be re-usable, metadata and data should be well described so that they can be replicated and/or combined in different settings.

This information should be presented in the searchable metadata. At least the metadata and the variable coding should be made findable, allowing codes to be updated over time if needed. Furthermore, open source publications are needed that report on the data and studies. Artificial Intelligence (AI) tools can be used to search for information in these publications and eventually make further use of the data (Moore et al. 2019; Rugard et al. 2020). Machine accessible description of the whole history of the data (data provenance) is an important asset. Several projects require that newly derived information or data generated by data users, is provided back to the data custodians. For BBMRI-ERIC this information can be stored centrally with BBMRI-ERIC acting as data controller. Depending on the data model used in a project, this has consequences as well on the data re-use (Table 1). As part of the European Open Science Data Cloud (EOSC), European Infrastructures are needed to build and maintain solid systems for high quality data storage and data analysis according to the FAIR principles. EOSC-Life brings together 13 Life Science research infrastructures to create an open, digital and collaborative space for biological and medical research. It addresses the data policies needed for human research data under GDPR (https://www.eosc-life.eu/). ENVRI-FAIR is the connection of the Cluster of 13 Environmental Research Infrastructures (ENVRI) to the EOSC. (https://envri.eu/home-envri-fair/).

Another important aspect of data reusability is its trustworthiness and fitness for purpose, which links to the formal definition of data quality as a compliance to a set of requirements (ISO8000) defined by the purpose. It links to open standards in the open science concept and also reacts to the demand for reproducibility in life sciences research in the last two decades (Freedman et al. 2015). These aspects can be covered to large extent by systematic generation of provenance information for data and linking it into a chain that goes back to the original data sources, of course subject to authorization for sensitive data provenance. In order to develop interoperable machine-actional provenance standards for the biotechnology and life sciences domains including provenance information management requirements, the International Standardization Organization (ISO) developed ISO standard 23,494 as a part of the Technical Committee 276 on biotechnology.

To promote and facilitate re-use of data, user communities should be formed. A good example is the HealthyCloud that intends to create a healthy ecosystem of public health research and health data. Stakeholders are involved in the activities underpinning data sharing. Involvement of international and local stakeholders ensure interconnectivity and mutual understanding of needs and potential barriers. To achieve re-use of the data, incentives could be generated by creating win–win situations for all involved actors by creating more potential when data could be shared and re-used. For example, by providing collaboration with stakeholders and policy makers that will directly refer and use the data for policies on environment and health. Or by establishing a user community with external stakeholders that may provide funding for common initiatives in which they are interested.

3. Barriers to implementation of FAIR data in E&H research

The issues to be overcome in implementing FAIR data and open science, while respecting GDPR, in the E&H research area are not primarily of technical nature but rather concern legal, ethical and political barriers, as well as lack of resources and good incentives for data custodians to embark on sharing their data.

Implementing the FAIR and open science principles in a GDPR-compliant manner respecting all legal and ethical constraints remains a challenge. The GDPR specifies that there should be a legal basis for processing the data subject's personal data. An explicit consent as legal basis can often not be guaranteed (Art. 7 of the GDPR). Public interest or secondary use of health data for scientific research, as specified in Art. 9 (2) of the GDPR, could be used as legal basis, but here also specific requirements should be met (European Commission 2021). To overcome ethical barriers, dynamic informed consents (Dankar et al. 2020) can be used to enable on-going engagement and communication between individuals and the custodians and the users of their data, while respecting the GDPR. The GDPR principles apply to (sensitive) personal data such as obtained in E&H research but are a barrier for open data and accessibility. However, data could be made accessible under controlled access protocols.

The political barrier is not much of an issue in the area of genetics and rare diseases, as for the rare diseases the scarcity of the data required collaboration and sharing of data to make advancements, and in the field of genetics the need for rapid replication of results mitigated this barrier. In the E&H research context, we are currently in a phase where the transition is driven by "a stick and not the carrots" i.e. actors are not yet motivated by a clear win–win situation. As mentioned in the section on re-usability, a win–win situation could be obtained by establishing a user community with external stakeholders who could use the data and may provide funding for common initiatives in which they are interested.

From a technical point of view, new enabling technologies are available, and they are rapidly evolving. Virtual research environments can be established in which data can be analyzed without the data having to leave the custodian servers. Federated data analysis systems such as DataSHIELD (Gaye et al. 2014) offer solutions to make data remotely accessible and combinable with other data sources without the need for transferring the data. However, the current application possibilities of this technology do not yet offer a conclusive solution for all types of analyses. Moreover, these federated systems come along with the need of financial and technical expertise. Project funding should include the necessary resources needed to implement the FAIR data and open science principles. Resources and strategies for implementation should be budgeted and foreseen in the study design phase of new projects.

4. Conclusion and recommendations

In the December 2020 workshop representatives of several large EU research consortia and of the European Commission reflected on how to apply the FAIR data and open science principles in European E&H research projects. The European E&H research scene has now moved towards hybrid data sharing systems, implementing a fit-for-purpose data sharing model by centralized and/or federated systems. This should allow making optimal use and re-use of generated data in the EU

projects, as a sustainable solution. The FAIR principles and open science concepts are widely endorsed by the environment and health researchers, however there are still some steps to take to allow full implementation.

New technical developments such as federated data management and analysis systems, machine learning, advanced search software, **common ontologies and data quality standards and further harmonization** will facilitate that data are findable, accessible, interoperable and re-usable. New investments in infrastructure and expertise are needed as well as more knowledge to remove further ethical, legal and political obstacles. Big consortia are building expertise along these lines. E&H research data, including HBM data, should find its place in the **European Open Science Cloud**. Communities on health and population data, environmental data and other publicly available data have to **interconnect and synergize**. Moreover, HBM data and environmental data should be connectable to the Health data cloud. At an early phase, when writing a project proposal and starting a new project a **Data Stewardship Plan** should be developed on how to connect to the large European infrastructures. The latter should prevent data ending up in a black hole, and as such enabling sustainable solutions.

As data sharing in the E&H field will concern sensitive personal data including health data, the management of GDPR compliant access control should be an important feature of supporting research infrastructure, in addition to management of data (standards, quality, interoperability) and technology support for federated data analysis. A European research infrastructure for environment and health data would be one of the FAIR data infrastructures that could be linked to other initiatives like BBMRI-ERIC, ELIXIR, HealthyCloud and others. The recent development of the research infrastructure for EnvIRonmental Exposure assessmeNt in Europe (EIRENE) (https://www.eirene-ri.eu/) within the ESFRI Roadmap 2021 could serve this purpose. The EIRENE RI mission is to establish a sustainable research infrastructure enabling the advancement of environment and health research in Europe by bringing together complementary capacities available in the member states, harmonizing and upgrading them to address current scientific and societal challenges in the areas of chemical exposures and population health.

As a result of the workshop, the following recommendations are formulated:

1. **Invest in/support new technical developments**.

New technologies such as federated data management and analysis systems should be further developed and find their application into E&H research. Advanced search software and machine learning will facilitate that data are findable, accessible, interoperable and re-usable.

2. **Further implement common ontologies and data harmonization**.

Implementation of existing ontologies and development of new common (harmonized) ontologies, and data quality standards, that are currently lacking are essential for the implementation of federated data management and analysis systems and will promote re-use of data. Resources and strategies for implementation should be included in the design phase of new projects.

3. **Integrate E&H data in EOSC**.

Integration of E&H research data should be implemented in the development and deployment of the EOSC, taking into account this type of personal research data.

4. **Interconnect and synergize different domains**.

For optimal use of available data, bridges must be built between the different research domains and related data infrastructures to promote and facilitate collaboration.

5. **New projects should develop a Data Stewardship Plan**.

Writing a data stewardship plan at an early stage when project proposals are developed raises awareness among researchers and forces active planning on how to connect their project data to European infrastructures. This should prevent data ending up in a black hole, and not being re-used any further after the project.

6. **Remit of an EU level E&H data research infrastructure** should comprise both technology support for data management and analysis, and enable GDPR compliant access to sensitive human data for research across EU countries and research studies. This should result in a sustainable way of maintaining all data generated within EU projects.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the policy officers at the European Commission who contributed to the workshop and the following discussion: Sofie Norager, Peter Korytar, Andreas Holtel, Jana Makedonska, Christina Kyriakopoulou.

HBM4EU is co-financed under Horizon 2020 (grant agreement No 733032). EUCAN-Connect is funded by the European Commission within the call topic SC1-BHC-05-2018: “International flagship collaboration with Canada for human data storage, integration and sharing to enable personalized medicine approaches” (Grant Agreement No 824989). The Canadian project partners have been funded by the Canadian Institutes of Health Research (CIHR) and the Fonds de recherche du Québec – Santé. The LifeCycle project received funding from the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 733206 LifeCycle). The HELIX project received funding from the European Community’s Seventh Framework Programme (grant agreement no 308333 HELIX). ISGlobal acknowledges support from the Spanish Ministry of Science and Innovation through the “Centro de Excelencia Severo Ochoa 2019-2023” Program (CEX2018-000806-S), and support from the Generalitat de Catalunya through the CERCA Program. Roel Vermeulen is supported by EXPOSOME-NL and EXPANSE. EXPOSOME-NL is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.017). EXPANSE has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 874627. EOSC-Life project supported by EU Horizon 2020, grant agreement no. 824087. EJP-RD and Solve-RD projects have been financed under Horizon 2000 (grant numbers H2020 779257 and H2020 825575).

References

Beelen, R., Stafoggia, M., Raaschou-Nielsen, O., Andersen, Z.J., Xun, W.W., Katsouyanni, K., Dimakopoulou, K., Brunekreef, B., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Houthuijs, D., Nieuwenhuijsen, M., Oudin, A., Forsberg, B., Olsson, D., Salomaa, V., Lanki, T., Yli-Tuomi, T., Oftedal, B., Aamodt, G., Nafstad, P., De Faire, U., Pedersen, N.L., Östenson, C.-G., Fratiglioni, L., Penell, J., Korek, M., Pyko, A., Eriksen, K.T., Tjønneland, A., Becker, T., Eeftens, M., Bots, M., Meliefste, K., Wang, M., Bueno-de-Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., de Hoogh, K., Key, T., Peters, A., Cyrys, J., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Dratva, J., Ducret-Stich, R., Vilier, A., Clavel-Chapelon, F., Stempfelet, M., Grioni, S., Krogh, V., Tsai, M.-Y., Marcon, A.,

Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Vineis, P., Hoek, G., 2014. Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 european cohorts. Epidemiology 25 (3), 368–378.

Bergeron, J., Doiron, D., Marcon, Y., Ferretti, V., Fortier, I., Beiki, O., 2018. Fostering population-based cohort data discovery: the maelstrom research cataloguing toolkit. PLoS ONE 13 (7), e0200926.

Beyan, O., Choudhury, A., van Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., Karim, M.R., Dumontier, M., Decker, S., da Silva Santos, L.O.B., Dekker, A., 2020. Distributed analytics on sensitive medical data: the personal health train. Data Intelligence 2 (1-2), 96–107.

Birks, L., Casas, M., Garcia, A.M., Alexander, J., Barros, H., Bergström, A., Bonde, J.P., Burdorf, A., Costet, N., Danileviciute, A., Eggesbø, M., Fernández, M.F., González-Galarzo, M.C., Hanke, W., Jaddoe, V., Kogevinas, M., Kull, I., Lertxundi, A., Melaki, V., Andersen, A.-M., Olea, N., Polanska, K., Rusconi, F., Santa-Marina, L., Santos, A.C., Vrijkotte, T., Zugna, D., Nieuwenhuijsen, M., Cordier, S., Vrijheid, M., 2016. Occupational exposure to endocrine-disrupting chemicals and birth weight and length of gestation: a European meta-analysis. Environ. Health Perspect. 124 (11), 1785–1793.

Bonino da Silva Santos LO. 2021. Fair digital object framework documentation. Available: https://fairdigitalobjectframework.org/[accessed March, 4 2021].

Casas, M., Nieuwenhuijsen, M., Martínez, D., Ballester, F., Basagaña, X., Basterrechea, M., Chatzi, L., Chevrier, C., Eggesbø, M., Fernandez, M.F., Govarts, E., Guxens, M., Grimalt, J.O., Hertz-Picciotto, I., Iszatt, N., Kasper-Sonnenberg, M., Kiviranta, H., Kogevinas, M., Palkovicova, L., Ranft, U., Schoeters, G., Patelarou, E., Petersen, M.S., Torrent, M., Trnovec, T., Valvi, D., Toft, G.V., Weihe, P., Weisglas-Kuperus, N., Wilhelm, M., Wittsiepe, J., Vrijheid, M., Bonde, J.P., 2015. Prenatal exposure to pcb-153, p, p'-dde and birth outcomes in 9000 mother-child pairs: Exposure-response relationship and effect modifiers. Environ. Int. 74, 23–31.

Corporate Finance Institute. 2020. Data anonymization. Available: https://corporatefinanceinstitute.com/resources/knowledge/other/data-anonymization/ [accessed March, 4 2021].

Dankar, F.K., Gergely, M., Malin, B., Badji, R., Dankar, S.K., Shuaib, K., 2020. Dynamic-informed consent: a potential solution for ethical dilemmas in population sequencing initiatives. Comput. Struct. Biotechnol. J. 18, 913–921.

Den Hond, E., Govarts, E., Willems, H., Smolders, R., Casteleyn, L., Kolossa-Gehring, M., Schwedler, G., Seiwert, M., Fiddicke, U., Castaño, A., Esteban, M., Angerer, J., Koch, H.M., Schindler, B.K., Sepai, O., Exley, K., Bloemen, L., Horvat, M., Knudsen, L.E., Joas, A., Joas, R., Biot, P., Aerts, D., Koppen, G., Katsonouri, A., Hadjipanayis, A., Krskova, A., Maly, M., Mørck, T.A., Rudnai, P., Kozepesy, S., Mulcahy, M., Mannion, R., Gutleb, A.C., Fischer, M.E., Ligocka, D., Jakubowski, M., Reis, M.F., Namorado, S., Gurzau, A.E., Lupsa, I.-R., Halzlova, K., Jajcaj, M., Mazej, D., Tratnik, J.S., López, A., Lopez, E., Berglund, M., Larsson, K., Lehmann, A., Crettaz, P., Schoeters, G., 2015. First steps toward harmonized human biomonitoring in Europe: demonstration project to perform human biomonitoring on a European scale. Environ. Health Perspect. 123 (3), 255–263.

Dwork, C.R., 2014. The algorithmic foundations of differential privacy. Foundations Trends Theor. Computer Sci. 211–407.

European Commission. 2018a. Final report and action plan from the European commission expert group on fair data. Turning fair into reality. Available: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf [accessed March, 4 2021].

European Commission. 2018b. Final report and recommendations of the commission 2nd high level expert group on the european open science cloud (eosc): Prompting an eosc in practice. Available: https://ec.europa.eu/info/sites/default/files/prompting_an_eosc_in_practice.pdf [accessed March, 4 2021].

European Commission. 2018b. Final report and recommendations of the commission 2nd high level expert group on the european open science cloud (eosc): Prompting an eosc in practice. Available: https://ec.europa.eu/info/sites/default/files/prompting_an_eosc_in_practice.pdf [accessed March, 4 2021].

European Commission. 2018b. Final report and recommendations of the commission 2nd high level expert group on the european open science cloud (eosc): Prompting an eosc in practice. Available: https://ec.europa.eu/info/sites/default/files/prompting_an_eosc_in_practice.pdf [accessed March, 4 2021].

Felix, J.F., Joubert, B.R., Baccarelli, A.A., Sharp, G.C., Almqvist, C., Annesi-Maesano, I., et al., 2018. Cohort profile: Pregnancy and childhood epigenetics (pace) consortium. Int. J. Epidemiol. 47, 22–23u.

Freedman, L.P., Cockburn, I.M., Simcoe, T.S., 2015. The economics of reproducibility in preclinical research. PLoS Biol. 13, e1002165.

Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E.M., Minion, J., Boyd, A.W., Newby, C.J., Nuotio, M.-L., Wilson, R., Butters, O., Murtagh, B., Demir, I., Doiron, D., Giepmans, L., Wallace, S.E., Budin-Ljøsne, I., Oliver Schmidt, C., Boffetta, P., Boniol, M., Bota, M., Carter, K.W., deKlerk, N., Dibben, C., Francis, R.W., Hiekkalinna, T., Hveem, K., Kvaløy, K., Millar, S., Perry, I.J., Peters, A., Phillips, C.M., Popham, F., Raab, G., Reischl, E., Sheehan, N., Waldenberger, M., Perola, M., van den Heuvel, E., Macleod, J., Knoppers, B.M., Stolk, R.P., Fortier, I., Harris, J.R., Woffenbuttel, B.HR., Murtagh, M.J., Ferretti, V., Burton, P.R., 2014. Datashield: Taking the analysis to the data, not the data to the analysis. Int. J. Epidemiol. 43 (6), 1929–1944.

Gilles, L., Govarts, E., Rambaud, L., Vogel, N., Castaño, A., Esteban López, M., et al., 2021. Hbm4eu combines and harmonises human biomonitoring data across the eu, building on existing capacity - the hbm4eu survey. Int. J. Hyg. Environ. Health 237, 113809.

Govarts, E., Nieuwenhuijsen, M., Schoeters, G., Ballester, F., Bloemen, K., de Boer, M., Chevrier, C., Eggesbø, M., Guxens, M., Krämer, U., Legler, J., Martínez, D., Palkovicova, L., Patelarou, E., Ranft, U., Rautio, A., Petersen, M.S., Slama, R., Stigum, H., Toft, G., Trnovec, T., Vandentorren, S., Weihe, P., Kuperus, N.W., Wilhelm, M., Wittsiepe, J., Bonde, J.P., 2012. Birth weight and prenatal exposure to polychlorinated biphenyls (pcbs) and dichlorodiphenyldichloroethylene (dde): A meta-analysis within 12 european birth cohorts. Environ. Health Perspect. 120 (2), 162–170.

Holub, P., Swertz, M., Reihs, R., van Enckevort, D., Müller, H., Litton, J.-E., 2016. Bbmri-eric directory: 515 biobanks with over 60 million biological samples. Biopreserv Biobank 14 (6), 559–562.

Holub, P., Kohlmayer, F., Prasser, F., Mayrhofer, M.T., Schlünder, I., Martin, G.M., Casati, S., Koumakis, L., Wutte, A., Kozera, Ł., Strapagiel, D., Anton, G., Zanetti, G., Sezerman, O.U., Mendy, M., Valík, D., Lavitrano, M., Dagher, G., Zatloukal, K., van Ommen, G.B., Litton, J.-E., 2018. Enhancing reuse of data and biological material in medical research: from fair to fair-health. Biopreserv Biobank 16 (2), 97–105.

Jaddoe, V.W.V., Felix, J.F., Andersen, A.-M., Charles, M.-A., Chatzi, L., Corpeleijn, E., Donner, N., Elhakeem, A., Eriksson, J.G., Foong, R., Grote, V., Haakma, S., Hanson, M., Harris, J.R., Heude, B., Huang, R.-C., Inskip, H., Järvelin, M.-R., Koletzko, B., Lawlor, D.A., Lindeboom, M., McEachan, R.R.C., Mikkola, T.M., Nader, J.L.T., de Moira, A.P., Pizzi, C., Richiardi, L., Sebert, S., Schwalber, A., Sunyer, J., Swertz, M.A., Vafeiadi, M., Vrijheid, M., Wright, J., Duijts, L., Jaddoe, V. W.V., Felix, J.F., Duijts, L., El Marroun, H., Gaillard, R., Santos, S., Geurtsen, M.L., Kooijman, M.N., Mensink-Bout, S.M., Vehmeijer, F.O.L., Voerman, E., Vrijheid, M., Sunyer, J., Nieuwenhuijsen, M., Basagaña, X., Bustamante, M., Casas, M., de Castro, M., Cirugeda, L.E., Fernández-Barrés, S., Fossati, S., Garcia, R., Júlvez, J., Lertxundi, A.C., Lertxundi, N., Llop, S., López-Vicente, M., Lopez-Espinosa, M.-J., Maitre, L., Murcia, M., Lea, J., Urquiza, H., Warembourg, C., Richiardi, L., Pizzi, C., Zugna, D., Popovic, M., Isaevska, E., Maule, M., Moccia, C., Moirano, G., Rasella, D., Hanson, M.A., Inskip, H.M., Jacob, C.M., Salika, T., Lawlor, D.A., Elhakeem, A., Cadman, T., Andersen, A.-M., de Moira, A.P., Strandberg-Larsen, K.M., Pedersen, M., Vinther, J.L., Wright, J., McEachan, R.R.C., Wilson, P., Mason, D., Yang, T.C., Swertz, M.A., Corpeleijn, E., Haakma, S., Cardol, M., van Enckevoort, E., Hyde, E., Scholtens, S., Snieder, H., Thio, C.H.L., Vafeiadi, M., Chatzi, L., Margetaki, K.C.A., Roumeliotaki, T., Harris, J.R., Nader, J.L., Knudsen, G.P., Magnus, P., Charles, M.-A., Heude, B., Panico, L., Ichou, M., de Lauzon-Guillain, B., Dargent-Molina, P., Cornet, M., Florian, S.M., Harrar, F., Lepeule, J., Lioret, S., Melchior, M., Plancoulaine, S., Järvelin, M.-R., Sebert, S., Männikkö, M., Parmar, P., Rautio, N., Ronkainen, J., Tolvanen, M., Eriksson, J.G., Mikkola, T., Koletzko, B., Grote, V., Aumüller, N., Closa-Monasterolo, R., Escribano, J., Ferré, N., Gruszfeld, D., Gürlich, K., Langhendries, J.-P., Luque, V., Riva, E., Schwarzfischer, P., Totzauer, M., Verduci, E., Xhonneux, A., Zaragoza-Jordana, M., Lindeboom, M., Schwalber, A., Donner, N., Huang, R.-C., Foong, R.E., Hall, G.L., Lin, A., Carson, J., Melton, P., Rauschert, S., 2020. The lifecycle project-eu child cohort network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. Eur. J. Epidemiol. 35 (7), 709–724.

Kholod, I., Yanaki, E., Fomichev, D., Shalugin, E., Novikova, E., Filippov, E., Nordlund, M., 2020. Open-source federated learning frameworks for iot: a comparative review and analysis. Sensors 21 (1), 167. https://doi.org/10.3390/s21010167.

Kush, R.D., Warzel, D., Kush, M.A., Sherman, A., Navarro, E.A., Fitzmartin, R., Pétavy, F., Galvez, J., Becnel, L.B., Zhou, F.L., Harmon, N., Jauregui, B., Jackson, T., Hudson, L., 2020. Fair data sharing: the roles of common data elements and harmonization. J. Biomed. Inform. 107, 103421. https://doi.org/10.1016/j.jbi.2020.103421.

Lawson, J., Cabili, M.N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S. R., Boyles, R.R., Brookes, A.J., Brush, M., Burdett, T., Clissold, H., Donnelly, S., Dyke, S.O.M., Freeberg, M.A., Haendel, M.A., Hata, C., Holub, P., Jeanson, F., Jene, A., Kawashima, M., Kawashima, S., Konopko, M., Kyomugisha, I., Li, H., Linden, M., Rodriguez, L.L., Morita, M., Mulder, N., Muller, J., Nagaie, S., Nasir, J., Ogishima, S., Ota Wang, V., Paglione, L.D., Pandya, R.N., Parkinson, H., Philippakis, A.A., Prasser, F., Rambla, J., Reinold, K., Rushton, G.A., Saltzman, A., Saunders, G., Sofia, H.J., Spalding, J.D., Swertz, M.A., Tulchinsky, I., van Enckevort, E.J., Varma, S., Voisin, C., Yamamoto, N., Yamasaki, C., Zass, L., Guidry Auvil, J.M., Nyrönen, T.H., Courtot, M., 2021. The data use ontology to streamline responsible access to human biomedical datasets. Cell Genom 1 (2), 100028. https://doi.org/10.1016/j.xgen.2021.100028.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: abide results. Med. Image Anal. 65, 101765. https://doi.org/10.1016/j.media.2020.101765.

Lochmüller, H., Badowska, D.M., Thompson, R., Knoers, N.V., Aartsma-Rus, A., Gut, I., Wood, L., Harmuth, T., Durudas, A., Graessner, H., Schaefer, F., Riess, O., 2018. Rd-connect, neuromics and eurenomics: collaborative european initiative for rare diseases. Eur. J. Hum. Genet. 26 (6), 778–785.

Maitre, L., de Bont, J., Casas, M., Robinson, O., Aasvang, G.M., Agier, L., Andrušaitytė, S., Ballester, F., Basagaña, X., Borràs, E., Brochot, C., Bustamante, M., Carracedo, A., de Castro, M., Dedele, A., Donaire-Gonzalez, D., Estivill, X., Evandt, J., Fossati, S., Giorgis-Allemand, L., R Gonzalez, J., Granum, B., Grazuleviciene, R., Bjerve Gützkow, K., Småstuen Haug, L., Hernandez-Ferrer, C., Heude, B., Ibarluzea, J., Julvez, J., Karachaliou, M., Keun, H.C., Hjertager Krog, N., Lau, C.-H., Leventakou, V., Lyon-Caen, S., Manzano, C., Mason, D., McEachan, R., Meltzer, H. M., Petraviciene, I., Quentin, J., Roumeliotaki, T., Sabido, E., Saulnier, P.-J., Siskos, A.P., Siroux, V., Sunyer, J., Tamayo, I., Urquiza, J., Vafeiadi, M., van Gent, D., Vives-Usano, M., Waiblinger, D., Warembourg, C., Chatzi, L., Coen, M., van den Hazel, P., Nieuwenhuijsen, M.J., Slama, R., Thomsen, C., Wright, J., Vrijheid, M., 2018. Human early life exposome (helix) study: a European population-based exposome cohort. BMJ Open 8 (9), e021311. https://doi.org/10.1136/bmjopen-2017-021311.

Matalonga, L., Hernández-Ferrer, C., Piscia, D., Cohen, E., Cuesta, I., Danis, D., Denommé-Pichon, A.-S., Duffourd, Y., Gilissen, C., Johari, M., Laurie, S., Li, S., Matalonga, L., Nelson, I., Peters, S., Paramonov, I., Prasanth, S., Robinson, P., Sablauskas, K., Savarese, M., Steyaert, W., van der Velde, J.K., Vitobello, A., Schüle, R., Synofzik, M., Töpf, A., Vissers, L.E.L.M., de Voer, R., Aretz, S., Capella, G., de Voer, R.M., Evans, G., Pelaez, J.G., Holinski-Feder, E., Hoogerbrugge, N., Laner, A., Oliveira, C., Rump, A., Schröck, E., Sommer, A.K., Steinke-Lange, V., Paske, I.t., Tischkowitz, M., Valle, L., Banka, S., Benetti, E., Casari, G., Ciolfi, A., Clayton-Smith, J., Dallapiccola, B., de Boer, E., Denommé-Pichon, A.-S., Ellwanger, K., Faivre, L., Graessner, H., Haack, T.B., Hammarsjö, A., Havlovicova, M., Hoischen, A., Hugon, A., Jackson, A., Kleefstra, T., Lindstrand, A., López-Martín, E., Macek, M., Morleo, M., Nigro, V., Nordgren, A., Pettersson, M., Pinelli, M., Pizzi, S., Posada, M., Radio, F.C., Renieri, A., Rooryck, C., Ryba, L., Schwarz, M., Tartaglia, M., Thauvin, C., Torella, A., Trimouille, A., Verloes, A., Vissers, L., Vitobello, A., Votypka, P., Vyshka, K., Zurek, B., Baets, J., Beijer, D., Bonne, G., Cohen, E., Cossins, J., Evangelista, T., Ferlini, A., Hackman, P., Hanna, M. G., Horvath, R., Houlden, H., Johari, M., Lau, J., Lochmüller, H., Macken, W.L., Musacchia, F., Nascimento, A., Natera-de Benito, D., Nigro, V., Piluso, G., Pini, V., Pitceathly, R.D.S., Polavarapu, K., Cruz, P.M.R., Sarkozy, A., Savarese, M., Selvatici, R., Thompson, R., Udd, B., Van de Vondel, L., Vandrovcova, J., Zaharieva, I., Baets, J., Balicza, P., Chinnery, P., Dürr, A., Haack, T., Hengel, H., Houlden, H., Kamsteeg, E.-J., Kamsteeg, C., Lohmann, K., Macaya, A., Marcé-Grau, A., Maver, A., Molnar, J., Münchau, A., Peterlin, B., Riess, O., Schöls, L., Schüle-Freyer, R., Stevanin, G., Synofzik, M., Timmerman, V., van de Warrenburg, B., van Os, N., Wayand, M., Wilke, C., Tonda, R., Laurie, S., Fernandez-Callejo, M., Picó, D., Garcia-Linares, C., Papakonstantinou, A., Corvó, A., Joshi, R., Diez, H., Gut, I., Hoischen, A., Graessner, H., Beltran, S., Haack, T.B., Graessner, H., Zurek, B., Ellwanger, K., Ossowski, S., Demidov, G., Sturm, M., Schulze-Hentrich, J. M., Schüle, R., Kessler, C., Wayand, M., Schöls, L., Hengel, H., Heutink, P., Brunner, H., Scheffer, H., Hoogerbrugge, N., 't Hoen, P.A.C., Steyaert, W., Sablauskas, K., Kamsteeg, E.-J., van de Warrenburg, B., te Paske, I., Janssen, E., Steehouwer, M., Yaldiz, B., Brookes, A.J., Veal, C., Gibson, S., Wadsley, M., Mehtarizadeh, M., Riaz, U., Warren, G., Dizjikan, F.Y., Shorter, T., Straub, V., Bettolo, C.M., Specht, S., Clayton-Smith, J., Banka, S., Alexander, E., Jackson, A., Faivre, L., Thauvin, C., Duffourd, Y., Tisserant, E., Bruel, A.-L., Peyron, C., Pélissier, A., Beltran, S., Gut, I.G., Laurie, S., Piscia, D., Matalonga, L., Papakonstantinou, A., Bullich, G., Corvo, A., Garcia, C., Fernandez-Callejo, M., Hernández, C., Picó, D., Paramonov, I., Lochmüller, H., Gumus, G., Bros-Facer, V., Rath, A., Hanauer, M., Olry, A., Lagorce, D., Havrylenko, S., Izem, K., Rigour, F., Durr, A., Davoine, C.-S., Guillot-Noel, L., Heinzmann, A., Coarelli, G., Bonne, G., Evangelista, T., Allamand, V., Nelson, I., Yaou, R.B., Metay, C., Eymard, B., Cohen, E., Atalaia, A., Stojkovic, T., Macek, M., Turnovec, M., Thomasová, D., Kremliková, R.P., Franková, V., Havlovicová, M., Kremlik, V., Parkinson, H., Keane, T., Spalding, D., Senf, A., Danis, D., Robert, G., Costa, A., Patch, C., Hanna, M., Houlden, H., Reilly, M., Vandrovcova, J., Muntoni, F., Sarkozy, A., Timmerman, V., Baets, J., Van de Vondel, L., Beijer, D., de Jonghe, P., Banfi, S., Torella, A., Ferlini, A., Selvatici, R., Rossi, R., Neri, M., Aretz, S., Spier, I., Peters, S., Oliveira, C., Pelaez, J.G., Matos, A.R., José, C.S., Ferreira, M., Gullo, I., Fernandes, S., Garrido, L., Ferreira, P., Carneiro, F., Swertz, M.A., Johansson, L., van der Vries, G., Neerincx, P.B., Roelofs-Prins, D., Köhler, S., Metcalfe, A., Rooryck, C., Trimouille, A., Castello, R., Morleo, M., Varavallo, A., De la Paz, M.P., Sánchez, E.B., Martín, E.L., Delgado, B.M., de la Rosa, F.J.A.G., Radio, F.C., Tartaglia, M., Renieri, A., Benetti, E., Balicza, P., Molnar, M.J., Maver, A., Peterlin, B., Münchau, A., Lohmann, K., Herzog, R., Pauly, M., Macaya, A., Marcé-Grau, A., Osorio, A.N., de Benito, D.N., Lochmüller, H., Thompson, R., Polavarapu, K., Beeson, D., Cossins, J., Cruz, P.M.R., Hackman, P., Johari, M., Savarese, M., Udd, B., Horvath, R., Capella, G., Valle, L., Holinski-Feder, E., Laner, A., Steinke-Lange, V., Schröck, E., Rump, A., 2021. Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. Eur. J. Hum. Genet. 29 (9), 1337–1347.

Mirchev, M., Mircheva, I., Kerekovska, A., 2020. The academic viewpoint on patient data ownership in the context of big data: scoping review. J. Med. Internet Res. 22 (8), e22214. https://doi.org/10.2196/22214.

Moore, J.H., Boland, M.R., Camara, P.G., Chervitz, H., Gonzalez, G., Himes, B.E., Kim, D., Mowery, D.L., Ritchie, M.D., Shen, L.i., Urbanowicz, R.J., Holmes, J.H., 2019. Preparing next-generation scientists for biomedical big data: artificial intelligence approaches. Per Med. 16 (3), 247–257.

Rehm, H.L., Page, A.J.H., Smith, L., Adams, J.B., Alterovitz, G., Babb, L.J., et al., 2021. Ga4gh: International policies and standards for data sharing across genomic research and healthcare. Cell Genomics 1, 100029.

Rugard, M., Coumoul, X., Carvaillo, J.-C., Barouki, R., Audouze, K., 2020. Deciphering adverse outcome pathway network linked to bisphenol f using text mining and systems toxicology approaches. Toxicol. Sci. 173 (1), 32–40.

Schroeder W. 2013. Practicing open science.

Sonnenschein-van der Voort, A.M., Arends, L.R., de Jongste, J.C., Annesi-Maesano, I., Arshad, S.H., Barros, H., et al., 2014. Preterm birth, infant weight gain, and childhood asthma risk: A meta-analysis of 147,000 european children. J. Allergy Clin. Immunol. 133, 1317–1329.

Strak, M., Weinmayr, G., Rodopoulou, S., Chen, J., de Hoogh, K., Andersen, Z.J., et al., 2021. Long term exposure to low level air pollution and mortality in eight european cohorts within the elapse project: Pooled analysis. BMJ 374, n1904.

van der Velde, K.J., Imhann, F., Charbon, B., Pang, C., van Enckevort, D., Slofstra, M., Barbieri, R., Alberts, R., Hendriksen, D., Kelpin, F., de Haan, M., de Boer, T., Haakma, S., Stroomberg, C., Scholtens, S., van de Geijn, G.-J., Festen, E.A.M., Weersma, R.K., Swertz, M.A., Wren, J., 2019. Molgenis research: Advanced bioinformatics data software for non-bioinformaticians. Bioinformatics 35 (6), 1076–1078.

van Veen, E.-B., 2018. Observational health research in europe: Understanding the general data protection regulation and underlying debate. Eur. J. Cancer 104, 70–80.

Vlaanderen, J., de Hoogh, K., Hoek, G., Peters, A., Probst-Hensch, N., Scalbert, A., Melén, E., Tonne, C., de Wit, G.A., Chadeau-Hyam, M., Katsouyanni, K., Esko, T., Jongsma, K.R., Vermeulen, R., 2021. Developing the building blocks to elucidate the impact of the urban exposome on cardiometabolic-pulmonary disease: The eu expanse project. Environ Epidemiol 5 (4), e162. https://doi.org/10.1097/EE9.0000000000000162.

Vrijheid, M., Basagaña, X., Gonzalez, J.R., Jaddoe, V.W.V., Jensen, G., Keun, H.C., McEachan, R.R.C., Porcel, J., Siroux, V., Swertz, M.A., Thomsen, C., Aasvang, G.M., Andrušaitytė, S., Angeli, K., Avraam, D., Ballester, F., Burton, P., Bustamante, M., Casas, M., Chatzi, L., Chevrier, C., Cingotti, N., Conti, D., Crépet, A., Dadvand, P., Duijts, L., van Enckevort, E., Esplugues, A., Fossati, S., Garlantezec, R., Gómez Roig, M.D., Grazuleviciene, R., Gützkow, K.B., Guxens, M., Haakma, S., Hessel, E.V. S., Hoyles, L., Hyde, E., Klanova, J., van Klaveren, J.D., Kortenkamp, A., Le Brusquet, L., Leenen, I., Lertxundi, A., Lertxundi, N., Lionis, C., Llop, S., Lopez-Espinosa, M.-J., Lyon-Caen, S., Maitre, L., Mason, D., Mathy, S., Mazarico, E., Nawrot, T., Nieuwenhuijsen, M., Ortiz, R., Pedersen, M., Perelló, J., Pérez-Cruz, M., Philippat, C., Piler, P., Pizzi, C., Quentin, J., Richiardi, L., Rodriguez, A., Roumeliotaki, T., Sabin Capote, J.M., Santiago, L., Santos, S., Siskos, A.P., Strandberg-Larsen, K., Stratakis, N., Sunyer, J., Tenenhaus, A., Vafeiadi, M., Wilson, R.C., Wright, J., Yang, T., Slama, R., 2021. Advancing tools for human early lifecourse exposome research and translation (athlete): project overview. Environ. Epidemiol. 5 (5), e166. https://doi.org/10.1097/EE9.0000000000000166.

Wey, T.W., Doiron, D., Wissa, R., Fabre, G., Motoc, I., Noordzij, J.M., Ruiz, M., Timmermans, E., van Lenthe, F.J., Bobak, M., Chaix, B., Krokstad, S., Raina, P., Sund, E.R., Beenackers, M.A., Fortier, I., 2021. Overview of retrospective data harmonisation in the mindmap project: Process and results. J. Epidemiol. Community Health 75 (5), 433–441.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al., 2016. The fair guiding principles for scientific data management and stewardship. Sci. Data 3, 160018.

Zerka, F., Barakat, S., Walsh, S., Bogowicz, M., Leijenaar, R.T.H., Jochems, A., Miraglio, B., Townend, D., Lambin, P., 2020. Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO Clin Cancer Inform (4), 184–200. https://doi.org/10.1200/CCI.19.00047.

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D., Ferretti, V., 2019. The international cancer genome consortium data portal. Nat. Biotechnol. 37 (4), 367–369.

Zurek, B., Ellwanger, K., Vissers, L.E.L.M., Schüle, R., Synofzik, M., Töpf, A., de Voer, R. M., Laurie, S., Matalonga, L., Gilissen, C., Ossowski, S., 't Hoen, P.A.C., Vitobello, A., Schulze-Hentrich, J.M., Riess, O., Brunner, H.G., Brookes, A.J., Rath, A., Bonne, G., Gumus, G., Verloes, A., Hoogerbrugge, N., Evangelista, T., Harmuth, T., Swertz, M., Spalding, D., Hoischen, A., Beltran, S., Graessner, H., Haack, T.B., Zurek, B., Ellwanger, K., Demidov, G., Sturm, M., Kessler, C., Wayand, M., Wilke, C., Traschütz, A., Schöls, L., Hengel, H., Heutink, P., Brunner, H., Scheffer, H., Steyaert, W., Sablauskas, K., de Voer, R.M., Kamsteeg, E.-J., van de Warrenburg, B., van Os, N., te Paske, I., Janssen, E., de Boer, E., Steehouwer, M., Yaldiz, B., Kleefstra, T., Veal, C., Gibson, S., Wadsley, M., Mehtarizadeh, M., Riaz, U., Warren, G., Dizjikan, F.Y., Shorter, T., Straub, V., Bettolo, C.M., Specht, S., Clayton-Smith, J., Banka, S., Alexander, E., Jackson, A., Faivre, L., Thauvin, C., Vitobello, A., Denommé-Pichon, A.-S., Duffourd, Y., Tisserant, E., Bruel, A.-L., Peyron, C., Pélissier, A., Beltran, S., Gut, I.G., Laurie, S., Piscia, D., Matalonga, L., Papakonstantinou, A., Bullich, G., Corvo, A., Garcia, C., Fernandez-Callejo, M., Hernández, C., Picó, D., Paramonov, I., Lochmüller, H., Gumus, G., Bros-Facer, V., Hanauer, M., Olry, A., Lagorce, D., Havrylenko, S., Izem, K., Rigour, F., Stevanin, G., Durr, A., Davoine, C.-S., Guillot-Noel, L., Heinzmann, A., Coarelli, G., Allamand, V., Nelson, I., Yaou, R.B., Metay, C., Eymard, B., Cohen, E., Atalaia, A., Stojkovic, T., Macek, M., Turnovec, M., Thomasová, D., Kremliková, R.P., Franková, V., Havlovicová, M., Kremlik, V., Parkinson, H., Keane, T., Senf, A., Robinson, P., Danis, D., Robert, G., Costa, A., Patch, C., Hanna, M., Houlden, H., Reilly, M., Vandrovcova, J., Muntoni, F., Zaharieva, I., Sarkozy, A., Timmerman, V., Baets, J., Van de Vondel, L., Beijer, D., de Jonghe, P., Nigro, V., Banfi, S., Torella, A., Musacchia, F., Piluso, G., Ferlini, A., Selvatici, R., Rossi, R., Neri, M., Aretz, S., Spier, I., Sommer, A.K., Peters, S., Oliveira, C., Pelaez, J.G., Matos, A.R., José, C.S., Ferreira, M., Gullo, I., Fernandes, S., Garrido, L., Ferreira, P., Carneiro, F., Swertz, M. A., Johansson, L., van der Velde, J.K., van der Vries, G., Neerincx, P.B., Roelofs-Prins, D., Köhler, S., Metcalfe, A., Verloes, A., Drunat, S., Rooryck, C., Trimouille, A., Castello, R., Morleo, M., Pinelli, M., Varavallo, A., De la Paz, M.P., Sánchez, E.B., Martín, E.L., Delgado, B.M., de la Rosa, F.J.A.G., Ciolfi, A., Dallapiccola, B., Pizzi, S., Radio, F.C., Tartaglia, M., Renieri, A., Benetti, E., Balicza, P., Molnar, M.J., Maver, A., Peterlin, B., Münchau, A., Lohmann, K., Herzog, R., Pauly, M., Macaya, A., Marcé-Grau, A., Osorio, A.N., de Benito, D.N., Lochmüller, H., Thompson, R., Polavarapu, K., Beeson, D., Cossins, J., Cruz, P.M.R., Hackman, P., Johari, M., Savarese, M., Udd, B., Horvath, R., Capella, G., Valle, L., Holinski-Feder, E., Laner, A., Steinke-Lange, V., Schröck, E., Rump, A., 2021. Solve-rd: Systematic pan-european data sharing and collaborative analysis to solve rare diseases. Eur. J. Hum. Genet. 29 (9), 1325–1331.