Devriendt, T., Shabani, M., & Borry, P. (2022). Policies to regulate data sharing of cohorts via data infrastructures: An interview study with funding agencies. *INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS*, *168*, 7 pages. doi:10.1016/j.ijmedinf.2022.104900

Policies to Regulate Data Sharing of Cohorts via Data Infrastructures: An Interview Study with Funding Agencies

Thijs Devriendt[a,*], Mahsa Shabani[b], Pascal Borry[a]

*Corresponding author: thijs.devriendt@kuleuven.be

mahsa.shabani@ugent.be

pascal.borry@kuleuven.be

[a]Centre for Biomedical Ethics and Law, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

[b]METAMEDICA, Faculty of Law and Criminology, UGent, Ghent, Belgium

Devriendt, T., Shabani, M., & Borry, P. (2022). Policies to regulate data sharing of cohorts via data infrastructures: An interview study with funding agencies. *INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS*, *168*, 7 pages. doi:10.1016/j.ijmedinf.2022.104900

Abstract

Background: Platforms are being constructed to stimulate cohort data sharing. Nevertheless, many policy barriers impede data sharing. Various interventions have been proposed to address these barriers, including incentive creation and data sharing mandates.

Aim: To understand funding agencies perspectives on policy interventions to encourage data sharing.

Methods: Members of funding agencies were recruited to participate in semi-structured interviews. Most funding agencies were located in Europe. Interview transcripts were analyzed through inductive content analysis.

Results: Mandates for cohort data sharing are not supported by agencies due to data protection and the preconditions for issuing mandates. Recommendation of platforms may be based on certification schemes, such repository accreditation. Monitoring mechanisms for cohort data sharing are currently absent, which complicates assessing compliance of researchers with funding agencies' policies and evidence production. Funding agencies are not imposing conditions on data access committees.

Conclusions: Policy measures that, in some ways, restrict the decision-making authority of researchers in terms of data sharing are not generally supported. Concrete steps are proposed to enable evidence-based policy making. Incentive design is paramount if funding agencies do not wish to impose restrictions on the decision-making authority of researchers.

Keywords: science policy, data infrastructure, data sharing, incentives, cohort data, mandates

1. Introduction

Researchers may outline their intention to share data in data sharing statements (DSSs) in published articles. In many cases, they describe that data can be made available upon reasonable request. For disease- or population-based cohorts, these requests are often circulated to the professional email address of cohort holders or formal data requests systems (e.g., institutional websites). These access requests are processed by data access committees (DACs). These DACs exist in various forms, including members affiliated with the research study (e.g., the principal investigator) or formal boards of institution-affiliated researchers.

This mode of working is applicable for individual-level data. Aggregate data may be made accessible through other process, depending on their level of privacy-sensitivity (e.g., open or registered access). Despite the existence of these routes to acquire data access, recent empirical research indicates that researchers' data sharing practices differ substantially from their DSSs (1). This lack of data sharing can be attributed to many factors, including ethical-legal barriers and lack of suitable infrastructures, resources and incentives for data sharing (1–3). The lack of data infrastructures is currently being addressed as there are many infrastructures (hereafter "platforms") under construction. These platforms can operate on the national or international level and may be disease-specific or disease-agnostic. Furthermore, they may have divergent data governance arrangements. Examples include Health-RI in the Netherlands, the Swiss Personal Health Network (SPHN) in Switzerland, Dementias Platform UK (DPUK) and platforms of the EUCAN-projects (http://www.eucanshare.eu/related-projects/). Nevertheless, the functionality of some of these platforms, particularly international ones, may be impeded by complex policy challenges (3–5). From this perspective, the lack of data sharing is a policy problem rather than a (purely) technical problem. To fundamentally tackle issues around incentives for data sharing, the science policy framework must be adjusted to ensure that data sharing and reuse through platforms is properly stimulated.

One way to stimulate data sharing is through the creation and installation of incentives for research teams and institutions. These incentives can take various forms. Reputational incentives may require altering attribution and evaluation systems. The former addresses how persons are linked to outputs through their labor (e.g., authorship, contributorship…). The latter addresses how outputs are weighted when deciding upon the prioritization of resources. Financial incentives

may involve changing funding distribution dynamics to research teams and institutions or installing cost-recovery mechanisms. The use of incentives for data sharing does, per definition, not entail strict compulsion of researchers to take certain actions. Other approaches that may stimulate data sharing, which generally require action by research funding agencies, are possible. These include the issuing of data sharing mandates or data retention policies, the direction of researchers towards established platforms upon acquiring funding and the prescription of rules for the organization of DACs. Some academics have argued in favor of these latter approaches to stimulate data sharing. For example, Danchev *et al*. suggest a comprehensive set of measures including that *"journals and funders should work towards incentivizing data sharing via funding mechanisms and data authorship, (…) discourage ambivalent wording in DSSs and possibly mandate data sharing. They can promote the use of pointers to data set location in repositories (…) [and] standardized choices for embargo periods, access requirements and conditions for data reuse"* (6). Research funding agencies are well-place to impose conditions upon data derived from research projects they fund. These agencies may operate on the national or international level, sponsor research in all or select research areas and derive their funding from governments (i.e., publicly-funded agencies) or donations (i.e., foundations). Some funding agencies, such as the European Commission, may in the future also require that trusted or certified repositories and infrastructures are used within some Horizon Europe work programs (7). International organizations may also help in creating guidelines or standards that funding agencies can recommend or enforce. For instance the Global Alliance for Genomics and Health (GA4GH) has developed guidelines and procedural standards for the composition and procedures of formal DACs and the due processing of data access requests (8). Contrary to installing incentives, the enactment of these policy

measures by funding agencies can impinge on the decision-making authority of cohort holders. While they would still be able to share data as they please, they might not be able to prevent other researchers from accessing their data. These measures are best understood as complementary to incentive design.

Our previous work has been oriented at understanding researchers' perspectives on incentives for data sharing, conceptualizing the use of incentives for cohort data sharing and how they fit within the science policy framework, and considering the use of policy measures to stimulate data sharing. Nevertheless, these exercises carry little weight if any outputs are not compatible with funding agencies' views on modifications to the science policy framework. In other words, any particular proposal intended to stimulate data sharing (e.g., data metrics) must fit within broader policy orientations (e.g., introduction narrative CVs). The implementation of measures depends both on their acceptability to research communities and the willingness of funding agencies to integrate them into the science policy framework. For this reason, we considered that engaging with funding agencies to understand their position on incentives and policy measures was essential. The goals of this interview study were to investigate the perspectives of funding agencies on potential alterations to recognition systems in academia; incentives to enhance cohort data sharing; data sharing policies and the governance of cohort data; and potential interactions between science policy and data sharing platforms for cohorts. This article reports on the results related to data sharing policies and potential interactions between science policy and data infrastructures (e.g., recommending use of infrastructures). The results associated with recognition systems in academia and reputational incentives for researchers to engage in cohort data sharing will be reported elsewhere due to extensive nature of the findings.

2. Methods

Funding agencies were invited to select candidates for participation in semi-structured interviews. In total, members of 17 funding agencies were interviewed. Funding agencies that participated came primarily from Europe (15/17) while two came from North-America (2/17). The profile of funding agencies can be found in the Supplementary Data. The interview with one member was disqualified due to being of very short length, interrupted and few topics relevant to the objectives being discussed. Interviews were conducted via Microsoft Meetings or Zoom. All interviews were audio-recorded and transcribed verbatim. Interview transcripts were analyzed through inductive content analysis, in which codes emerge from data rather than being predetermined (9). The study was approved by the Social and Societal Ethics Committee (SMEC) at KU Leuven (G-2021-3823). The full methodological details can be found in the Supplementary Data.

2. Results

The codes that emerged from analyzing interview transcripts clustered around various themes. With regard to data sharing policies, five themes emerged: (a) data management plans; (b) monitoring mechanisms for data sharing; (c) recommending or mandating the use of platforms; (d) mandating cohort data sharing; and (e) the organization of data access committees. Quotes that support many of the statements can be found in annex in Supplementary Data.

a. Data management plans

Nearly all funding agencies that participated in this study have adopted data management plans (DMPs). The purpose was DMPs is to encourage researchers to

reflect on their data management practices. DMPs are submitted at the beginning of the research project and can be regularly updated throughout the funding period. DMPs were often reported to be based on the template provided by *Science Europe*. Most funding agencies only required DMPs to be submitted in recent years (~2016-2022). Therefore, some agencies declared they had not yet evaluated these DMPs at the end of the grant period, for instance, for comprehensiveness of documentation or legibility. Nevertheless, one funding agency reported having performed an evaluation of the quality of DMPs, which indicated that data outputs were being insufficiently described. For this agency, it was difficult to assess whether outputs were shared in practice at the end of the funding period.

> *The quality of the data management plans was so poor that there is no way that you could tie the two [DMP and downstream sharing] together. When you ask researchers about their research outputs, they are mainly talking about publications. So even though our grant form asks for software and databases, we were getting not as much information submitted to us. (Interviewee 3 – foundation/specialist/international)*

b. Monitoring mechanisms for cohort data sharing

Monitoring mechanisms

Many funding agencies reported that they either did not possess means to monitor whether data have been made available and shared. Others reported that some monitoring mechanisms are available, in principle, yet they were not used for large-scale monitoring. Several software packages were anticipated to be used for monitoring purposes, such as Dimensions, Researchfish, CRIS-systems and tools provided by the FAIRware project. These mechanisms either rely on (a) tracking persistent identifiers (PIDs) of datasets available in data repositories or within publications; or (b) researchers self-reporting availability statements on their research data through infrastructures. These monitoring mechanisms would not

allow funding agencies to monitor cohort data sharing. Tracking PIDs is not applicable as access to cohorts must be applied for each individual research project. Furthermore, self-reported data availability statements would also not provide information on cohort data sharing (e.g., here, researchers report that their cohort data falls under "restricted access").

Funding agencies generally expressed an interest in having being able to monitor cohort data sharing. However, interviewees did often not foresee specific uses for the information acquired from monitoring mechanisms (e.g., financial penalties). Several interviewees stated that they want a record for each data access request to be created, which includes reasons for DACs rejecting requests. This was anticipated to allow parties to see whether DACs are declining access for inappropriate reasons.

Financial penalties for non-compliance

Monitoring mechanisms enable observing whether researchers are in compliance with data sharing policies of funding agencies. Without such mechanisms, non-compliance will not have consequences. Some funding agencies had reflected on whether financial penalties for not sharing data should be adopted, such as partially withholding the last tranche of project funding or withholding reimbursement for article processing charges (APCs). One interviewee expressed clear reluctance towards impose sanctions due to not being able to resolve the lack of data sharing at late stages in the project. Therefore, they instead supported interventions at earlier stages.

> *We could […] come down stringently at the end of grants and say: "If this does not happen, we are not going to give you the final ten percent of the grant". But changing the publication status of an article is far simpler than*

> *putting data out there. With many datatypes, if the first time you are thinking about it is at the end of a grant, it is way too late. (Interviewee 3 - foundation/specialist/international)*

Key performance indicators based on monitoring

Monitoring mechanisms would allow gathering quantitative data on research institutions' data sharing practices. In this context, several interviewees brought up the development of key performance indicators (KPIs) on data sharing. In general, KPIs are associated with goals that are valuable to organizations (e.g., number of articles for research institutions). KPIs help organizations assess whether their actions are having the desired impacts. In this sense, KPIs are an intricate part of evidence-based policy making. These KPIs were mentioned to be essential in establishing norms and monitoring growth trajectories for data sharing. This would allow assessing whether the rate of data sharing is increasing at different research institutions over time. Another potential use of KPIs was explained to be their implementation into performance-based funding distribution mechanisms for research institutions. In general, these mechanisms have the purpose of allocating financial resources to research institutions depending on certain performance indicators (e.g., total number of articles, impact factor). The integration of KPIs on data sharing into these mechanisms was foreseen to make data sharing itself more financially advantageous for research institutions.

> *We have a system where universities get funding based on articles. […] So, it is a race to get articles in the [higher] journals [because] those institutions will get more funding from the government. […] There is one concern if we are changing that system [with data sharing KPIs]. There might be some areas that are not that data-intensive and they will get less recognition. […] If we start to […] reward data sharing [and] you do not have the infrastructure, it is impossible for you and you will not get the points. (Interviewee 14 – public agency/generalist/national-regional)*

Nevertheless, these interviewees also described the complexities of installing these mechanisms in practice. For instance, KPIs would need to be field-specific.

Furthermore, attempts to integrate novel KPIs into financial distribution mechanisms could lead to arduous discussions with the management of research institutions. One interviewee stated that research institutions might fear that novel KPIs would ultimately restrict researchers' behavior. This argument has also been used to criticize current reward and funding distribution systems in academia: Because some behaviors are disproportionally rewarded over others, researchers cannot freely decide how to spend their time (if they wish to be successful). Another interviewee mentioned that KPIs on data sharing could lead to unfair competition due to some institutions having more research data than others.

c. Recommending or mandating the use of platforms

The emergence of platforms for data sharing raises the question of whether funding agencies will *obligate* or *recommend* researchers to use them, or whether they will not take action. Most funding agencies stated that they cannot mandate the use of *particular* data infrastructures or platforms. Various arguments were raised to support this position. Interviewees described difficulties with being able to easily assess the breadth of studies acceptable to particular data platforms (i.e., scope of data catalogues). Some platforms would be more suitable for sharing specific types of data than others. Therefore, researchers and data stewards were argued to be best placed to determine what is the most suitable infrastructure for their research data. Interviewees also raised that selecting and listing repositories or platforms for every research field cannot be resourced by their funding agencies. Furthermore, it was argued that these lists could not be updated quickly enough to keep up with the various platforms under construction. One interviewee explained that an obligation by funding agencies to share through particular platforms might imply having to sustain those platforms in the long-term. Lastly,

some funding agencies argued that recommending or mandating use of platforms should be based on certification schemes that are developed by external parties (e.g., international groups).

> *It is such a range [of platforms]. We would have to look at so many disease-specific platforms and to try to keep ourselves up to date would be really difficult. I think it would be difficult for us to mandate. We might as well publicize information about different platforms and say: "Here are some platforms that you might want to look at." (Interviewee 7A – public agency/specialist/national-regional)*

Certification schemes developed by external parties might be used to decide upon recommendation of infrastructures. Funding agencies may provide lists of trusted repositories based on formal accreditation schemes, such as CoreTrustSeal, compliance with the concept of "trusted repository" by *Science Europe* or recommendations by international groups, such as subgroups of the Research Data Alliance (RDA).

> *It is a strong advice to use a trusted repository. What we consider a trusted repository is according to the Science Europe guidelines. […] But we do not as an agency make recommendations. Because we fund research in all of the research in all academic areas. So, making recommendations for repositories would be a huge list. (Interviewee 10 – public agency/generalist /national-regional)*

Additional conditions for endorsing the use of platforms were raised less explicitly in discussions about suitable criteria for certification schemes and the functionality of platforms. These conditions pertained mainly to transparency over data access conditions and data governance. These include regulations on DAC membership, transparency over decision-making by DACs, the scope of accepted research data, transparency of data use conditions of all data sets, and the existence of appeal procedures when data access is declined.

> *Each of [the platforms] has to have a governance system and has to be transparent in terms of how they make decisions in terms of what data is*

> *accepted in, what data is provided and so on. […] There [should be] robust conditions around access to this data. […] These are the standards we expect. Ideally, there should be [transparency about] the access committee [and] their names. (Interviewee 13 – public agency/specialist/national-regional)*

d. Mandating cohort data sharing

In general, interviewees stated they did not believe their funding agency would be willing to mandate cohort data sharing. One fundamental assumption within the arguments provided by interviewees was that mandating data sharing implies having to prescribe *where* (meta)data ought to be made available. Therefore, similar arguments as those under the previous section were raised. Furthermore, some interviewees reported that their funding agencies would be hesitant to mandate the sharing of personal data, which are privacy-sensitive. In the European context, these data are subject to the provisions of the General Data Protection Regulation (GDPR) and dealing with legal complexities falls within the purview of research institutions. Interestingly, one interviewee explained that their funding agency deems that the degree to which cohort data have been made FAIR cannot be duly assessed in practice. Therefore, they could only recommend making data FAIR.

> *Is there some kind of authoritative body somewhere that can tell you and I that this dataset is FAIR? […] How do you determine [that]? If we are to write into our grant agreement a legally binding clause that tells people that their data should be FAIR, we should also be prepared to answer the question "Is my data FAIR now, is this good enough?" […] If there was a set of established repositories, with established data formats and standards, then it is easy to say: "You have to submit data according to these standards in this database". (Interviewee 8 – foundation/specialist/international)*

e. The organization of data access committees

Most funding agencies reported that they do not have formal policies on the organization of DACs. However, the organization of DACs would sometimes be

reviewed *ad hoc* in the negotiations with grantees (e.g., independence of the committee and absence of conflict of interest). Interviewees explained that the lack of guidance on DACs stems from the open research data movement still being in its "early days". Furthermore, it was argued that there is little in-house expertise to assess the adequacy of the organization of DACs. Interviewees also reported having no monitoring mechanisms for the decision making of DACs (i.e., whether data access requests are accepted or declined).

> *We leave [that] up to the researchers. We say, "Show us your terms of use" but we do not have the opportunity nor the expertise to assess that thoroughly. We rely on the universities that, we hope, organize those [DACs] properly. (Interview 6 – public agency/specialist/national-regional)*

Several interviewees described that the future goal for the organization of these DACs is an independent body, where conflict of interest is absent, with transparent data access conditions and committee membership. Furthermore, an appeal procedure that could allow researchers to bring forward instances of non-compliance with data access conditions was considered to be of importance. One interviewee described that there is a trade-off between independence of DACs on the one hand and expertise about data on the other.

> *We would like to see […] data access committees that have less conflict of interest […] Obviously, there is a certain trade-off at some point. I think to just have a completely centralized [committee] completely detached from the projects might not work as well. […] If needed to include the researchers that generated the data but to [move] further away from that. (Interviewee 1 – public agency/generalist/national-regional)*

3. Discussion

Our results illustrate the multi-faceted nature of policy making around data sharing platforms: DMPs, monitoring mechanisms, key performance indicators, data sharing mandates, certification schemes for platforms and the organization of

DACs. Funding agencies might shy away from forceful approaches to stimulate data sharing (e.g., data sharing mandates, obligatory usage of platforms with centralized DACs, enforce particular configurations of DACs…). Furthermore, there are currently no monitoring mechanisms or broad support to tie consequences to non-compliance. The lack of suitable monitoring mechanisms for data sharing also impedes gathering quantitative evidence on data sharing. In turn, this inhibits evidence-based policy making: Why would funding agencies attempt to address problems when they cannot realistically observe the effects of their interventions to resolve problems? In the following sections, the consequences of positions on certain science policy measures will be explored. Subsequently, it will be explained how platforms could address shortcomings of the science policy framework through the combined action of informaticians and funding agencies.

3.1 Science policy measures

3.1.1 Data sharing mandates

Our results indicate that mandates for data sharing are not widely supported by funding agencies for various reasons. Furthermore, they do not possess suitable monitoring mechanisms to identify non-compliance with data sharing policies. Funding agencies might refrain from coupling financial penalties to mandates. These factors could explain why mandates for clinical data sharing have so far not been commonly issued by funding agencies. Nevertheless, many academics have argued that mandates are necessary to compel researchers to share data (10–13). While some place this responsibility on funding agencies, others look at scientific journals as being responsible for imposing stricter conditions for data sharing (6,13). One example would be that journals could require researchers to describe detailed DSSs in their publications. Nevertheless, recent empirical research

indicates that DSSs in publications do not reflect actual data sharing practices (6). Some researchers have advocated for strict enforcement of policies and sanctioning when researchers are not compliant with data sharing policies (13,14). For instance, Couture *et al.* suggest that funding agencies should "dedicat[e] more resources to enforcing compliance with data requirements, provid[e] data-sharing tools and technical support to awardees, and administer stricter consequences for those who ignore data sharing preconditions" (14). Scientific journals may be powerless in stimulating cohort data sharing, as they cannot check the truth of DSSs and have few (realistic) means to sanction lack of sharing. In conclusion, data sharing mandates are rather unlikely to be widely used due to lack of support for implementation. Furthermore, they would not guarantee to ensure data sharing due to lack of monitoring and sanctioning. Anger *et al*. came to similar conclusions in their recent interview study with funding agencies (15).

3.1.2 Recommending or obligating the use of platforms

Based on our results, it seems improbable that funding agencies will recommend researchers to use *specific* platforms for data sharing. Agencies might instead rely on certification schemes for platforms that are developed by external parties (e.g., RDA groups, *Science Europe…*). These schemes set forward the minimal technical and governance conditions that platforms should comply with. In practice, there are no suitable certification schemes for data sharing platforms. The scheme for "trusted repositories" by *Science Europe* does not address any dimension of proper data access governance. Data access governance includes machine-readable data use conditions for cohorts, composition and decision-making procedures of DACs, oversight mechanisms and conditions for cohorts to enter platforms. CoreTrustSeals' accreditation scheme does also not propose suitable criteria for

this "type" of platform either. If conditions on data governance are not integrated into certification schemes, it is possible that researchers will be directed to submit (meta)data into "closed" rather than "open" platforms. For instance, the tracking of data contribution and usage patterns may illustrate that cohort holders are only sharing with select groups of people within platforms (i.e., clique-sharing) rather than larger groups (5). In that case, the functioning of the platform would not very be different from "closed" cohort consortia. Another example is that cohort holders that have data within platforms could prevent others from submitting their data (e.g., via a committee that decide on who can enter the platform). This could be advantageous for cohort holders within platforms in the short-term as they maximize visibility for their cohorts and hold the monopoly over useful tools within platforms (e.g., privacy-preserving data analysis tools). In conclusion, it is unlikely that funding agencies will recommend or obligate researchers to use platforms, if no suitable certification schemes for platforms are developed.

3.1.3 Imposing conditions on data access committees

Our results illustrated that funding agencies do not formally impose conditions on DACs that are setup by researchers. In practice, this means that the principal investigators often decide on data sharing, which contributes to restrictive data sharing practices. This situation may be avoided by organizing DACs at departmental, institutional, consortium, infrastructure or funder level (i.e., "centralization" of DACs). Another complementary way of avoiding conflicts of interest is by making all data access conditions, DAC composition and procedures transparent (i.e., "formalization" of DACs). Centralization and formalization of DACs are currently not common practice. Platforms, especially international ones, cannot easily influence individual research teams to change their DAC composition

or procedures. This means that even though cohorts have (meta)data within platforms, there is no guarantee that these data will be accessible when data access requests are submitted. In conclusion, it is currently not likely that funding agencies will impose certain conditions for DACs, which would end up promoting data sharing.

3.2 How can platforms help funding agencies take science policy measures?

The aforementioned sections demonstrate that policy measures that restrict the decision-making authority of researchers on data sharing will not be imposed. How can cohort data sharing then be stimulated? We suggest that platforms can help funding agencies *enact* certain policies if they are designed for this purpose. From this perspective, infrastructures can stimulate data sharing through both technical means (e.g., data catalogues, virtual research environments…) and policy means (e.g., enabling science policies measures). In the following paragraphs, the actions that should be taken collectively by informaticians and funding agencies are described. These actions are also outlined in Table 1.

Step 1: Build appropriate monitoring mechanisms for cohort data sharing

Two dimensions of monitoring data sharing should be addressed. First, the truth of statements in DMPs should be verifiable with little effort. If so, funding agencies could check whether data was shared through an infrastructure if researchers have written this in their DMPs. Second, it should be possible to check whether DACs are actually sharing data when they are requested. Funding agencies need to be able to detect if DACs are either systematically declining data access requests for unjustifiable reasons or are not reviewing requests at all. Monitoring mechanisms

based on data citations or self-reported data availability statements will not capture these dimensions at all for cohorts.

The first dimension can be addressed by attaching traceable persistent identifiers (PIDs) to cohorts within platforms. These PIDs should, in principle, enable funding agencies to check whether cohorts have been contributed to platforms. For instance, funding agencies can demand these DOIs are uploaded in DMPs. Cohorts should preferably be tagged with standardized data use conditions, including informed consent codes (e.g., DUO codes, ADA-M) (16,17). This information should be included in metadata maturity assessments for cohorts: Are their data access procedures and DAC membership sufficiently transparent? Are study metadata sufficiently detailed? Formal assessment of these dimensions would allow funding agencies to more easily check veracity of statements in DMPs.

The second dimension can be addressed by integrating formal data management systems for DACs into platforms. These systems could be designed to allow data applicants to tag their data access requests with (semi-)standardized information. Furthermore, responses of DACs can be tracked, including their reasons for declining the data access requests. This information should be bundled in data access records, which can be made available to (meta)researchers. This would allow funding agencies to check whether access requests were declined for legitimate reasons or not. We have described the concept of this system and its advantages in-detail in prior work (5). We strongly recommend that funding agencies obligate researchers to use these systems when they are mature. If not, researchers can always avoid being monitored by either joining platforms that refuse to adopt monitoring mechanisms or not joining platforms at all.

Step 2: Use monitoring mechanisms to formulate KPIs on data sharing

It is impossible to develop KPIs on cohort data sharing without monitoring mechanisms. As explained earlier, funding agencies cannot monitor whether statements in DMPs are true or whether DACs are accepting data access requests. Furthermore, the use of PIDs for cohorts is uncommon. For cohort studies, one potential proxy for academic data reuse is the number of citations on their associated Cohort Profile(s). Cohort Profiles were initially introduced by the International Journal of Epidemiology in 2004 to allow cohort holders to share the experiences of running their studies and to stimulate collaboration between cohort holders (18). Researchers can cite these profiles when they reuse data. However, they do not seem to be cited consistently by external researchers. They also appear to be cited by the cohort holders themselves and not every cohort study has its own profile. Further limitations include not being able to capture data sharing instances that did not (e.g., negative results) or do not (e.g., data sharing with commercial companies) result in publication, or where data access requests were refused by DACs. Therefore, KPIs based on Cohort Profiles would not reflect openness (e.g., granted/total data access requests). Because of these limitations, we suggest that KPIs should be based on the monitoring system described in the previous section (5). Currently, the collection of indicators, such as data usage and deposition patterns, remains non-standardized and fragmented across platforms (13).

Step 3: Set up research programs for evidence-based policy making to identify and evaluate appropriate interventions

These KPIs for cohort data sharing enable conducting zero-measurement and follow up on cohort data sharing by funding agencies. This also facilitates using traditional designs for scientific research studies (e.g., pre-post study designs,

comparative studies…). Policy makers that provide advice to funding agencies can use these methods to detect problems and identify the most effective policy measures. This method of collecting information on data sharing would, in our view, be generally superior to the iterative circulation of surveys on data sharing for multiple reasons (5).

Step 4: Consider integrating field-specific KPIs for data sharing into performance-based funding distribution instruments for research institutions.

These instruments have the purpose of distributing resources to research institutions based on certain performance indicators. The integration of KPIs makes institutions *compete* on the basis of some dimension of cohort data sharing. In turn, this incentivizes institutions to invest into data management and infrastructure units that span across medical faculties. These units should then be made formally part of the data lifecycle to help research teams set up efficient pipelines for data sharing. Institutions that outperform others would receive more financial resources, while those that do not act would lose resources. Note that introducing this dynamic requires careful conceptualization of indicators and assessment of their all limitations, such as technical shortcomings and perverse incentives.

Step 5: Create a certification scheme for platforms

A certification scheme that includes technical and data access governance conditions should be created so funding agencies can actively recommend use of platforms. This scheme should be sufficiently flexible to accommodate diversity in technical and data governance configurations of platforms.

4. Conclusion

Mandates for cohort data sharing are not generally supported by funding agencies. Recommendation of platforms by agencies can only be based on formal certification schemes, which are currently non-existent. Funding agencies do not impose conditions for the organization of DACs. Monitoring mechanisms for cohort data sharing are absent, which complicates the design of key performance indicators for data sharing, verifying compliance with data sharing policies and evidence production on the impact of science policy measures. We suggest that funding agencies and informaticians consider creating monitoring mechanisms for cohort data sharing within platforms. Furthermore, incentive design is paramount if funding agencies do not wish to impose restrictions on the decision-making authority of researchers.

Author's contributions

TD, MS and PB conceptualized the work and methodology. TD performed data collection, analysis, interpretation and drafting of manuscript. MS and PB critically reviewed the manuscript.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program [grant 825903]. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of interest

None.

Limitations

Devriendt, T., Shabani, M., & Borry, P. (2022). Policies to regulate data sharing of cohorts via data infrastructures: An interview study with funding agencies. *INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS*, *168*, 7 pages. doi:10.1016/j.ijmedinf.2022.104900

Terminology

Definitions

Certification scheme: A list of technical (e.g., DOIs) and governance (e.g., transparency over data use conditions) requirements that platforms should comply with

Cohorts: Observational research studies that contain clinical data on groups of patients or members of the general public (e.g., longitudinal cohorts, case-control studies, cross-sectional)

Cohort holders: Those persons that are responsible for managing research cohorts and, often, have contributed to data collection processes

Cohort Profile: Article that explains the study rationale, set-up and suitability for certain types of analyses

Data access committee: Committee that processes data access requests by external researchers

Data sharing platforms: Research infrastructures that contain multiple technical components (e.g., data catalogues, data management system, virtual research environments) to facilitate cohort data sharing and analysis

Data sharing mandate: Funding agency requiring researchers to share or make FAIR data that were generated under projects they funded

Science policy: General term used to refer to the entire policy framework: encompasses data sharing policies (e.g., mandates, DACs, recommendation, embargos…) and attribution, evaluation and funding distribution systems (e.g., grant evaluation procedures).

Devriendt, T., Shabani, M., & Borry, P. (2022). Policies to regulate data sharing of cohorts via data infrastructures: An interview study with funding agencies. *INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS*, *168*, 7 pages. doi:10.1016/j.ijmedinf.2022.104900

Abbreviations

DMP: Data management plan

DAC: Data access committee

DSS: Data sharing statement

FAIR: Findable, Accessible, Interoperable and Reusable

GA4GH: Global Alliance for Genomics and Health

GDPR: General Data Protection Regulation

KPI: Key performance indicator

Table 1. Recommendations for funding agencies and medical informaticians

State of affairs	Recommendations for medical informaticians	Recommendations for funding agencies
Funding agencies are unwilling to enforce cohort data sharing top-down	Overall approach: - Build data sharing platforms that *enable* certain policy interventions to be taken. This requires integrating monitoring mechanisms for cohort data sharing into platforms.	Overall approach: - Mandate transparency over data sharing practices - Consider policy interventions enabled by monitoring mechanism - Create and install incentives for data sharing (for institutions and research teams)

Cohort data sharing cannot be monitored	- Develop monitoring mechanisms for cohort data sharing (e.g., formal data management systems where DACs are registered) (5) - Assess whether this monitoring mechanism allows formulating data sharing indicators (relative-absolute) - Avoid fragmentation of monitoring mechanisms over platforms - Build system to send information back to funding agencies	- Consider integrating KPIs into financial distribution instruments to research institutions - Provide KPIs to meta-researchers to engage in evidence-based policy making - Consider recognizing and integrating indicators on data sharing into grant evaluation processes
No control or transparency over decision-making and organization of DACs	- Integrate monitoring mechanism into data sharing platforms	- Issue advice to researchers on how to design their DACs - Obligate researchers to submit their DACs in data management systems, if mature - Mandate transparency over DAC membership, procedures and organization - Encourage research institutions to create DACs

Platform use cannot be recommended	- Create guidelines that could support recommending platform use - Formulate KPIs that may be integrated into these certification schemes (5)	- Utilize certification schemes to recommend use of cohort data platforms

Summary points

What is already known on the topic?

- The rate of data sharing of DACs after requests to share is low
- The lack of data sharing is influenced by the science policy framework
- Data infrastructures are being constructed to facilitate cohort data sharing

What this study added to our knowledge:

- Funding agencies do not generally endorse data sharing mandates for various reasons
- Funding agencies lack monitoring mechanisms for cohort data sharing
- Funding agencies are not imposing conditions on DACs
- Certification schemes to recommend contributions to platforms are lacking

Limitations

There are various limitations to our study. The vast majority of funding agencies were from Europe, which means that results may not be generalizable to other regions. Furthermore, our discussions on science policy orientations around platforms were, at times, somewhat theoretical and anticipatory. Several interviews mentioned that there were still internal discussions ongoing on policy approaches. It is entirely possible that funding agencies will adopt different positions. There was only one data analyst involved in the coding process, which may mean that the coding process was insufficiently rigorous. We have made the underlying transcript data, coding scheme, and supporting codes available.

References

1. Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: mixed-methods study. J Clin

Epidemiol. 2022 May. doi: 10.1016/j.jclinepi.2022.05.019

2. Devriendt T, Borry P, Shabani M. Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers. PLoS One. 2021;16(7 July):1–14. doi: 10.1371/journal.pone.0254202

3. Devriendt T, Borry P, Shabani M. Credit and Recognition for Contributions to Data-Sharing Platforms Among Cohort Holders and Platform Developers in Europe: Interview Study. J Med Internet Res. 2022;24(1):e25983. doi: 10.2196/25983

4. Devriendt T, Ammann C, Asselbergs FW, Bernier A, Costas R, Friedrich MG, et al. An agenda-setting paper on data sharing platforms: euCanSHare workshop. Open Res Eur. 2021;1:80. doi: 10.12688/openreseurope.13860.2

5. Devriendt T, Shabani M, Lekadir K, Borry P. Data sharing platforms : instruments to inform and shape science policy on data sharing ? Scientometrics. 2022;(0123456789). doi: 10.1007/s11192-022-04361-2

6. Danchev V, Min Y, Borghi J, Baiocchi M, Ioannidis JPA. Evaluation of Data Sharing after Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement. JAMA Netw Open. 2021;4(1):1–12. doi: 10.1001/jamanetworkopen.2020.33972

7. Burgelman J-C, Pascu C, Szkuta K, Von Schomberg R, Karalopoulos A, Repanas K, et al. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. Front Big Data. 2019;2(December):1–6. doi: 10.3389/fdata.2019.00043

8. Data Access Committee Guiding Principles and Procedural Standards Policy. Global Alliance for Genomics and Health (GA4GH). 2021;1–11. Available at: www.ga4gh.org/wp-content/uploads/GA4GH-Data-Access-Committee-Guiding-Principles-and-Procedural-Standards-Policy-Final-version.pdf

9. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs. 2008;62(1):107–15. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2007.04569.x

10. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu YL, Snyders LN, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PLoS One. 2018;13(5):1–12. doi: 10.1371/journal.pone.0194768

11. Sim I, Stebbins M, Bierer BE, Butte AJ, Drazen J, Dzau V, et al. Time for NIH to lead on data sharing. Science. 2020 Mar;367(6484):1308–9. doi: 10.1126/science.aba4456

12. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020;21(10):615–29. doi: 10.1038/s41576-020-0257-5

13. Naudet F, Siebert M, Pellen C, Gaba J, Axfors C, Cristea I, et al. Medical journal requirements for clinical trial data sharing: Ripe for improvement. PLoS Med. 2021;18(10):1–8. doi: 10.1371/journal.pmed.1003844

14. Couture JL, Blake RE, McDonald G, Ward CL. A funder-imposed data publication requirement seldom inspired data sharing. PLoS One. 2018;13(7):1–13. doi: 10.1371/journal.pone.0199789

15. Dyke SOM, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, et al. Consent Codes: Upholding Standard Data Use Conditions. PLoS Genet. 2016;12(1):e1005772--e1005772. doi: 10.1371/journal.pgen.1005772

16. Bernier A, Knoppers BM. Longitudinal Health Studies: Secondary Uses Serving the Future. Biopreserv Biobank. 2021;00(00):1–10. doi: 10.1089/bio.2020.0171

17. Ebrahim S. Cohorts, infants and children. Int J Epidemiol. 2004;33(6):1165–6. doi: 10.1093/ije/dyh368

18. Ebrahim S. Cohort Profiles: what are they good for? Int J Epidemiol. 2021;50(2):367–70. doi: 10.1093/ije/dyab054