Murtagh, M. J., Turner, A., Minion, J. T., Fay, M., & Burton, P. R. (2016). International Data Sharing in Practice: New Technologies Meet Old Governance. Biopreservation and Biobanking, 14(3), 231-240. DOI: 10.1089/bio.2016.0002

International Data sharing in practice: new technologies meet old governance

Madeleine J Murtagh[1*], Andrew Turner[1], Joel T Minion[1], Michaela Fay[1], Paul R Burton[1]

Author affiliations: [1] Data to Knowledge Research Group, School of Social and Community Medicine, University of Bristol UK

* Corresponding author: Madeleine J Murtagh

Abstract

The social structures that govern data/sample release aim to safeguard the confidentiality and privacy of cohort research participants (without whom there would be no data or samples) and enable the realisation of societal benefit through optimising the scientific use of those cohorts. Within collaborations involving multiple cohorts and biobanks, however, the local, national and supra-national institutional and legal guidelines for research (which produce a multiplicity of data access governance structures and guidelines) risk impeding the very science which is the *raison d'etre* of these consortia.

We present an ethnographic study which examined the epistemic and non-epistemic values driving decisions about data access and their consequences in the context of the pilot of an integrated approach to co-analysis of data; we demonstrate how the potential analytic flexibility offered by this approach was lost under contemporary data access governance. We identify three dominant values: protecting the research participant, protecting the study, and protecting the researcher. These values were both supported by and juxtaposed against a 'public good' argument and each was used as a rationale to both promote or to inhibit sharing of data. While protection of the research participants was central to access permissions, decisions were also attentive to the desire of researchers to see their efforts in building population biobanks and cohorts realised in the form of scientific outputs.

We conclude that systems for governing and enabling data access in large consortia need to: (1) protect disclosure of research participant information or identity; (2) ensure the specific expectations of research participants are met; (3) embody systems of review that are transparent and not compromised by the specific interests of one particular group of stakeholders; and, (4) facilitate data access procedures that are timely and efficient. Practical solutions are urgently needed. New approaches to data access governance should be trialled (and formally evaluated) with input from and discussion with stakeholders.

Key words: DataSHIELD, data sharing, data access, epistemic values, non-epistemic values, governance, biobank, cohort

Introduction

The requirement for large samples sizes and pooled analyses of population-based, phenotypically and genotypically rich data and sample repositories is axiomatic in contemporary bioscience(1, 2) . Major funders internationally - notably the NIH and Wellcome Trust(3) – encourage, and where necessary, impose an imperative to share the data produced from publicly funded biomedical resources, biobanks and longitudinal cohort studies (hereafter, 'study(ies)'[i])(4). This imperative to share positions science as a 'public good' (i.e. publicly funded science should benefit the public)(4). It also references a pragmatic concern about value for money, namely that science and its outputs (in this case data and samples) should be used widely to ensure public benefit.

The scientific (or epistemic) rationale for sharing is founded on a set of four quintessential precepts (or values). First, that no single cohort study can provide the depth and breadth of data and samples required to achieve sufficient power for valid (and robust) analysis of often small, biological effects and associations.

Second, that no single cohort can provide the heterogeneity required to produce findings that will robustly hold for diverse populations with very different cultural, social and environmental contexts (e.g. epigenetic influences).

Third, a fundamental requirement for achieving robust analyses requires that data from phenotypic variables or derived from samples (hereafter 'data') must be harmonised retrospectively – or, more rarely, standardised sufficiently well prospectively – to allow direct comparison.

Fourth, to enable sharing, data must be accessible; that is, they must be both small enough for transport under current IT systems and have the appropriate ethical and/or legal permissions in place to allow for their release. Increasingly, data are too large for easy *and* secure transport. And, in more practical terms, it would simply not be feasible for the study investigators who generated a dataset to answer all the potential research questions that could be asked of those data.

From a societal perspective, the social structures (legislation, guidelines, ethical review and data access controls) must meet the very reasonable expectations of safeguarding the confidentiality and privacy of cohort research participants (without whom there would be no data or samples) and their equally reasonable expectation of societal benefit through optimising the scientific use of those studies (5-7).

In response to the widely held value commitments directing science to the achievement of societal benefit – and in particular to allow valid privacy-protecting analysis of data from multiple cohorts – an integrated approach to secure co-analysis of data has been developed combining the DataSHaPER (8, 9) harmonisation method, the Opal (10) and Mica (10) databasing and publishing tools, and DataSHIELD (11) analysis software to offer an open-source mechanism for the secure analysis of harmonised research participant-level data without data ever needing to leave their host site (12) – See Figure 1.

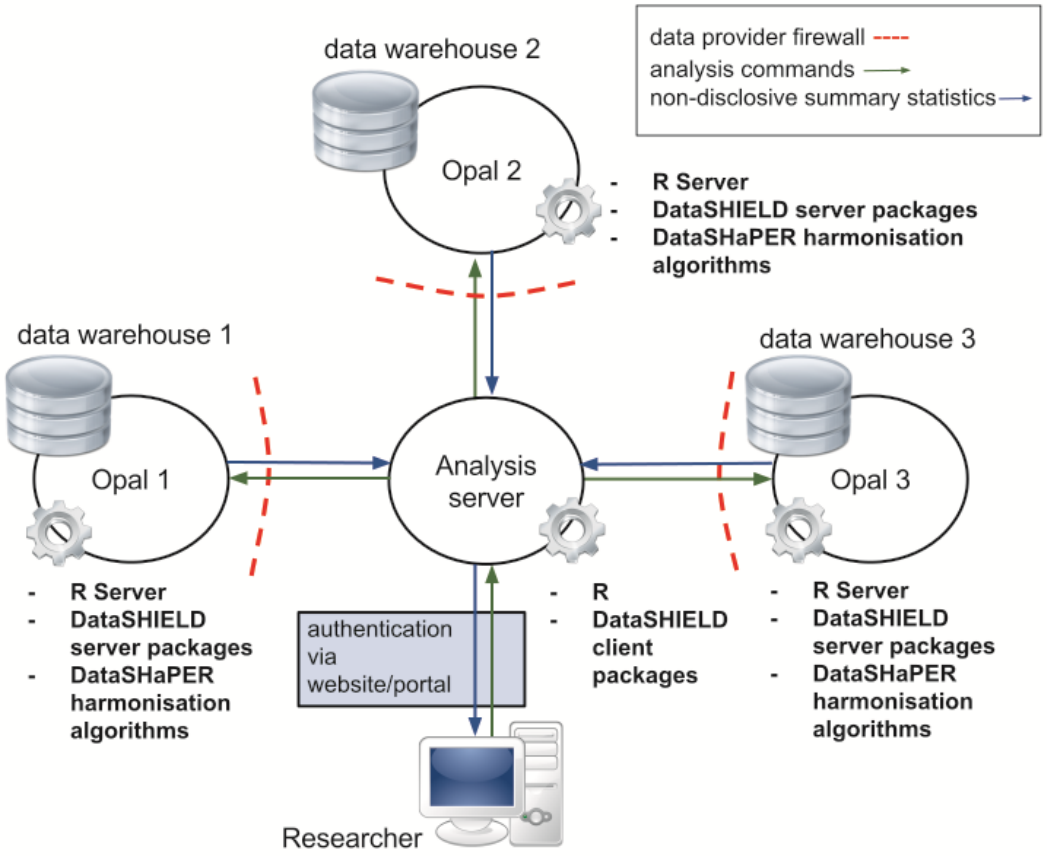
DataSHaPER is a structured method for data harmonisation involving a series of steps progressing from initial definition of a research question through to the generation and validation of a harmonised dataset. Fundamental elements within DataSHaPER include creation of a 'DataSchema' to identify and define key variables in a harmonised dataset as well as creation of a library of algorithms for transforming selected variables from each individual study to produce the DataSchema variables required.

Opal, Mica and DataSHIELD are open source software applications designed respectively to store, explore and analyse study data. Both Mica and DataSHIELD are integrated into Opal, which provides

the core infrastructure needed to store, describe and manipulate data. Mica creates websites that provide details about data, metadata and other material from multiple studies (e.g. searchable study catalogues; descriptions of datasets; electronic management of data access). DataSHIELD permits the secure co-analysis of study data and offers researchers a secure, privacy-protecting interface for analysing data held in Opal through a specially tailored R statistical environment.

Importantly, in combination with Mica and Opal, DataSHIELD denies users access to the research participant level data themselves while providing full and flexible access to the information held in those data (11). Collectively, this integrated suite of infrastructural and methodological tools provides a new technological solution to key challenges implicit in creating, accessing and securely co-analysing the extensive high quality harmonized data that underpin so much of modern biomedical science: hereafter, 'the integrated approach to secure co-analysis of data'.

Figure 1[ii]



These tools (comprising the integrated approach to secure co-analysis of data) were developed (DataSHIELD) or extended (Opal, DataSHaPER technology) as part of the five year Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BioSHaRE-EU) project(13). The BioSHaRE-EU project ultimately involved 15 European population-based cohort studies (https://www.bioshare.eu/studies) and successfully aimed to facilitate data harmonisation and standardisation for pooled analysis to investigate several common complex diseases and traits of substantial Public Health relevance (10-12) (also see page 4 - "Project Context" at website@ https://www.bioshare.eu/content/final-publishable-summary. The tools provide a general platform for many data storage and analysis scenarios: the software tools are open-source and continue to be actively developed beyond BioSHaRE-EU as they are rolled out in new initiatives.

Here it is perhaps important to recognise that there are a number of ways in which data can be made available to users. None of these is 'right' or 'wrong' in all circumstances (14). On the contrary, the crucial issue is to choose amongst the possible options that are appropriate for the particular context that has been encountered (14). In this regard, it is sometimes implied that there is a stark choice between investing research resources in creating large centralized data storage sites from which users can obtain all of their data (a centralized warehousing model) or a distributed model in which studies manage their own data and provide access for remote analysis - or joint co-analysis of several studies – by external users. This, latter, is a federated model. But, in reality, both approaches have strengths and weaknesses and there is no sense in which one is 'better' than the other.

For example, if highly specialist data demanding substantive informatics expertise (*e.g.* DNA sequence data) are to be held and made available locally at many studies (the federated model), the extensive knowledge, documentation, enabling hardware and governance expertise will necessarily be replicated many times over. This would be highly inefficient and strongly argues the case for specialist genomic data to be made available under a centralized model – as is done, for example, at the EGA (European Genome Phenome Archive, Europe) and at dbGaP (the database of Genotypes and Phenotypes, US).

In contrast, in relation to the rich array of general phenotypes (often covering tens of thousands of variables) generated over time by a major longitudinal study, a federated model may well be much more efficient. This is because the scientists and other professionals who know a study best (in almost all regards) are directly responsible for managing its data.

To consider just a few benefits, this greatly facilitates: the identification and correction of data errors; the addition of new data (particularly when this represents updated assessment of data items collected earlier) to an appropriately structured database; the identification and management of data associated with an individual who has withdrawn from the study (the "right to be forgotten" now having been enshrined in European legislation), and provision of access to a much greater depth of knowledge about the data that may be sought by a user who needs to fully understand the context under which they were originally collected.

Although the BioSHaRE-eu project placed a particular emphasis on federated analysis – because that is the area in which new tools are particularly needed – it was fully recognised that both approaches have their appropriate places, and the tools that were developed can, in fact, be applied to data held in a single central repository as well as in a federated system.

Over the course of the BioSHaRE-EU project a multi-method ethnographic study (including a formal evaluation of the tools above) was undertaken (15-17). The ethnographic study captured the enthusiasm of users and developers alike and explored the significant challenges they encountered when using the integrated approach to secure co-analysis of data.

Challenges varied according to the experience of respondents (e.g. their level of computational literacy) and to the particular standpoint from which they spoke (e.g. developer, implementer, user, study representative). Some difficulties when developing and implementing a new approach are unsurprising; indeed, it could be argued that challenges encountered in early iterations are a necessary part of advancing the usability and applicability of any new technology or tool.

In this instance, the integrated approach to secure co-analysis of data was experienced as being demanding for individual research users. In particular, effective use of the novel multi-faceted infrastructure – and associated methods – that enabled and facilitated a secure federated approach to data access and co-analysis saw the emergence of a new type of scientist: one as familiar with

computational systems and algorithmic syntax as with the rudiments of rigorous statistical analysis and interpretation.

A key feature of these *usage* challenges was their potential for amelioration or accommodation within the ongoing development of the tools comprising the integrated approach to secure co-analysis of data; for example, building a more 'user friendly' interface for DataSHIELD became part of the development strategy. In contrast, the other major challenge which was identified during the implementation pilot was found to lie outside the scope of remediation by the tool development process: namely, the complex array of governance mechanisms, pragmatic and political considerations that directly impact on the provision of access to particular data.

Under a typical contemporary governance model for data access, each new use of data held by a study requires a separate application for data access specifying which data are required, the research question(s) to be addressed (i.e. the uses to which those data will be put), explicit identification of the scientists who require access to the data and - if data are to be physically transferred how those data will ultimately be stored. All BioSHaRE-EU cohort studies required application to a dedicated study-specific data access committee, additionally one also required application for ethical approval for the research question.

Contemporary data access is typically based on the physical transfer of research participant-level data (often called *microdata*) using systems enabling secure transmission or else on the provision of secure physical sites which users can visit to carry out an appropriately secured analysis of the data. Under both of these approaches research participant-level data are accessed directly during analysis. It is, in part, because these research participant-level data are potentially disclosive of research participant identity and/or sensitive information that formal governance systems are ubiquitous.

However, an alternative to this model is used in meta-analyses (e.g. GWAS) wherein multiple studies each contribute to an overarching joint analysis by conducting in-house analysis of their own research participant-level data and then transmitting the resultant study-level estimates (and standard errors) to a designated analysis centre that combines the results, often using random effects meta-analysis. Crucially, this means that it is only study-level statistics – not research participant-level data – that must physically be shared, and study-level data are generally understood to be non-disclosive[iii].

The integrated approach to secure co-analysis of data using DataSHIELD for federated analysis of multiple studies represents a novel approach that may be viewed as combining the best features of both standard approaches: research participant-level data are never directly visualised or physically accessed but the analysis is typically able to use all of the information held in the data. Furthermore, analytic control lies primarily with the analysis centre – avoiding the need to request or direct individual studies to undertake analysis themselves – while the control of data security remains with the original studies and the data custodians within those studies.

Despite its potential benefits, this new integrated approach to data co-analysis had a clear potential to generate important new questions relating to study governance. The cohorts involved in the pilot of the integrated approach to secure co-analysis of data therefore needed to carefully assess their criteria for access based on an approach in which physical access to data was not required for analysis of those data. In doing that, however, it was recognised that some new challenges might be difficult to anticipate. In the event, the seeking of access permission comprised two components:

(1) permission for physical access to harmonise the cohorts' data using DataSHaPER – this would be led by the cohorts themselves and each cohort could decide what role, if any, they wanted the

central harmonization team to play and it is that which would determine whether or not the central team needed permission to work with the study's research participant-level data;

(2) access permission to enable full research participant-level data analysis using DataSHIELD though, given that the harmonisation had then been completed, no physical access to the data was required at any study.

This pilot of the integrated approach to secure co-analysis of data thereby offered a unique opportunity to identify the foundational epistemic (scientific) and non-epistemic (social, cultural, moral, economic, etc.) values upon which the studies had made decisions about provision of access to their data. Such values usually remain tacit, embedded within official policy or criteria for decision making. But these values – as do all values – have significant consequences for action or practice. Indeed, we rely on them to do so. However, as a partial driver for this ethnographic research, it was recognised that there may be actions or practices consequent on such values that are in practice counterproductive and that this may only become evident when new modes of analysis render some governance protections no longer applicable or unnecessarily prohibitive: governance of data is always a balance between constraint (protection) and openness (use).

Drawing upon the BioSHaRE-EU ethnographic study, this paper therefore examines the epistemic and non-epistemic values driving decision making by studies about sharing data, the intersection of such practices with the requirements of a new mode of co-analysis of data, and their consequences.

Methods

We employed an ethnographic methodology(19, 20) to understand data sharing and data access in the setting in which these occurred: in this case, the implementation pilot of an integrated approach to secure co-analysis of data. The ethnographic study of BioSHaRE-EU comprised participant observation and individual and group interviews undertaken over the length of the project between December 2011 and November 2015. Here we present findings from data generated from interviews with the developers, implementers and users of the integrated approach and with cohort contacts from the 15 cohorts involved in BioSHaRE-EU.

Telephone interviews were conducted with 21 people between September 2014 and January 2015, lasting an average of 45 minutes each. The discussions were audio recorded, transcribed verbatim, and analysed thematically using the constant comparative method (21). A purposive sample for interview was generated to include clinicians, epidemiologists, statisticians, bioinformaticians, study co-ordinators and social scientists who had been actively involved in or associated with BioSHaRE-EU for much or all of its existence.

During the interviews respondents were asked about their experience in using, developing and/or implementing one or more of the four tools, and their thoughts on the usability and appropriateness of the tools for sharing and translating data from multiple cohort and biobank studies. Interview analysis was further informed by the broader understandings gained from participant observation undertaken during BioSHaRE-EU meetings. We used these ethnographic observations to orientate and inform our primary analysis and do not present them here as a second substantive source of evidence in their own right.

Analysis and theoretical orientation: Ethnographic analysis is often descriptive (e.g. thematic analyses of qualitative findings), aiming to provides in-depth understanding or 'thick description'(22) of the phenomenon under examination. It may also include, as we do here, a more interpretive form of analysis to better understand or explicate the patterns of behaviour, views or practices observed. Ethnographic analysis typically proceeds inductively, meaning that the narrative or analytical categories developed are closely-linked to the richness of the data, rather than being led by a priori

theory. This approach allows qualitative researchers to develop an insider's perspective and therefore better understand and explain social phenomena.

In this paper we focus upon the values that shape decision making about data sharing in general and in the context of the co-analysis approach used in the BioSHaRE-EU pilot specifically. Theoretically, we take the position that values expressed in the language used to describe them are, as claimed of language itself (23-26), performative. That is, values have effects: values do not simply describe principles or standards but actively shape the definition and consequences of such principles or standards.

If, for example, we hold that research participant privacy is an important value, then we act not only to maintain that value in certain ways but also to preclude certain other actions. Co-analysis of data from multiple cohorts potentially operates in the context of competing values and interests: those of the cohort research participants, and those of the researchers aiming to use the data derived from the involvement of such research participants.

To understand how values shaped the practice of data sharing we, therefore, examined the consequences of two forms of values, epistemic and non-epistemic, called upon by study representatives and others in their descriptions of and rationales for data sharing. The term 'epistemic' is used by philosophers, derived from the Greek epistēmē (knowledge), to mean "related to knowledge or its validation" (Oxford English Dictionary). Thus 'epistemic values' refer to knowledge-related principles or standards and non-epistemic values refers to socio-ethical, cultural, economic and other non-'knowledge related' values, assumptions and standpoints). While this distinction (epistemic and non-epistemic) is somewhat arbitrary it produces a useful foil for considering the range of values informing scientific and governance decisions in practice.

Qualitative analysis is necessarily interpretive and as such any interpretation made of the data is one of a number of possible interpretations. That being the case, it is standard practice to build an argument for that analysis using a representative selection of extracts along with detailed interpretation so that readers may assess that interpretation for themselves (c.f. (27)). We therefore present the findings of our analysis below using quotations from the interviews.

Respondents consented to the recording, transcription and analysis of their interview with the understanding that quotations would be included in academic publications. Interviews were conducted in English but, given the international nature of the research, for most respondents English was their second, third or even fourth language. In order to maintain the integrity of respondents talk we have not 'corrected' the language or grammar used but have lightly edited it for ease of reading where necessary. Where the words of both interviewer and respondent are included, these are differentiated by "INT" (interviewer) and "RES" (respondent).

Because the BioSHaRE-EU community is small and its members publicly identifiable, there is no indication given in quotations of the standpoint from which the respondent is speaking; unless this is necessary to explicate analysis. In fact, most speakers embodied multiple standpoints, for example, simultaneously being both developers and users of the tools as well as having some relationship with both BioSHaRE-EU and a specific cohort.

While the analysis used the concept of epistemic and non-epistemic values to understand how values shaped data sharing, the findings are presents an argument which first identifies respondents' underpinning values (e.g. protecting the research participant, study or researcher), the strategies used to achieve these values (e.g. varying modes of controlling access), critiques of these values and strategies and the consequences of these values in terms of their effects on undertaking research.

Ethical approval was granted by the University of Leicester, College of Medical and Biological Sciences Research Ethics Committee.

Results

Data sharing values

Why use the DataSHIELD approach?

As noted in the introduction, the process of using a novel integrated approach to secure co-analysis of data during their development and piloting within BioSHaRE-EU brought with it considerable challenges. Why then did the developers, implementers, users and studies persevere with a process that many regarded as onerous? In the words of one study representative, the security provided for sensitive data by this approach outweighed the "hassle" (currently) required to make it work.

> *Data security is a major issue when you bring data together from different sites or to simply protect the data from being circulated anywhere. This approach [federated analysis under DataSHIELD] is a very important tool to address those issues - to improve data safety when handling sensitive material. So that is why we decided to get involved with this: to set up the service and everything that's necessary to work with it. And that was the basic motivation, conceiving this as a useful tool which addresses some serious issues in epidemiology. […] It improves data safety. That's the main advantage. Because in the end you need a strong 'pro' argument otherwise you won't undergo the hassle of working with this kind of entry access with the data when you have other possibilities which are easier and more flexible.*

Maintaining data security was a key driver and underpinning non-epistemic value evidenced in the ethnographic data and was consistent across the studies involved in BioSHaRE-EU. Data access arrangements as currently configured in European biobanks, however, presented a significant impediment to shared data analysis. These are inconsistent across biobanks and can be extremely time consuming. One study contact (also a user) summarised the issues as follows:

> *It's very hard if you want to do analysis in the scientific world. The need for doing analysis with more than one cohort is increasing. For almost every analysis you need to pool data from more than one cohort. But often it's difficult to pool them physically, to get the data from other cohorts on to your computer and do the analysis. It's a long and a complicated way because every cohort has its own board; it has to decide whether they want to give the data away or not. So it's a complicated process. So if it's possible to do it in the way that DataSHIELD does – this federated analysis so you don't have to apply for the data but you can just get full data without seeing the individual data – then it's a big advantage. It's easier to get the permission to do it. But on the other hand, technically, [the DataSHIELD approach] is complicated.*

The challenge of multiple (and at times complex) procedures involved in gaining access to multiple studies is well known to anyone involved in meta-analyses or GWAS – in these instances, however, no research participant identifiable data need be released because primary analysis is done in house and study-level results then pooled. The DataSHIELD approach offered the possibility of bypassing problems of physical data access while still undertaking a fully efficient joint analysis of the research participant-level data in all studies at once.

Protecting research participants

Protecting the research participant has fundamentally shaped the evolution of ethics and data access (non-epistemic value). Research participants are asked to consent to specific projects or actions based on the principle that have the right to know what they are agreeing to. The Declaration of Helsinki (28) holds that the primary purpose of medical research is to generate new knowledge. Where such research involves human research participants the Declaration holds that the goal of knowledge generation cannot take precedence over the rights and interests of research participants and that national or international ethical, legal or regulatory requirements cannot undermine the protections under the Declaration.

The conditions under which research participants are involved in research should include that: research protects their dignity, rights to self-determination (ie. research participants are adequately informed prior to consent), privacy and confidentiality; and, research only be carried out by appropriately trained and qualified researchers. These precepts have carried through to access-governance through the principle of one application for each proposed research question with access restricted to named *bona fide* researchers.

> *If I want to do a kind of research then I have to submit a proposal to the involved cohorts, and if they decide that I can do that then I get permission to use the data. But they have to be sure also, the cohorts, that only the people who have the permission for this certain analysis have access to the data. I think that's an important point.*

For the studies, governance of data (or sample) use offers a critical point to ensure that the expectations and trust of their research participants is maintained (non-epistemic value). The other key governance apparatus is receipt of a positive ethical opinion for the research protocol and associated consent and research participant information package. These documents outline: the scope of research participant involvement; the uses to which their data or material may be put; definition of any restrictions on that usage (e.g. who is permitted to use those data/samples); the cohort's responsibilities regarding research participant confidentiality and privacy; and other specific policies (e.g. return of results).

Consents and research participant information provide the framework establishing research participant expectations of study involvement and therein guide decision making about data access for further research. The integrated approach included importing study data into an Opal instance which remained behind the study's firewall. Despite this, many study contacts needed to persuade their study's ethics, access or scientific boards that the integrated approach to secure co-analysis of data fulfilled existing commitments to research participants based on non-epistemic values.

> *From the ethical point of view, in the beginning I would say that people here were very, were concerned because probably they didn't immediately understand the fact that data uploaded was safe – because the novelty of all of this infrastructure – with DataSHIELD. So, the novelty is that the data are still in-house but there is the possibility to work on them without using or accessing individual data. So at the beginning it was a bit hard to make clear that there was no problem in ethics. But in the end it was clarified so we could go on.*

When asked why such concerns likely arose in the first place, this respondent added:

> *Probably there was some fear that some sensitive data could be disclosed and used in some way unethically. Or probably just some fear that somebody else*

could publish some research when this dataset is quite new. We have not published anything on this data yet because the collection is ongoing.

Elsewhere a study representative echoed these data access concerns as well as a further concern for many studies.

RES [The study's access committee] were afraid that all the data would be put into something that they couldn't sort of control who got access. And I had to make a specific point about you can't really take out individual data because that's more sensitive. That people who do the research on the data pool, they can only do it via DataSHIELD. And then just get some summary data back.

INT Were they happy with that?

RES They were much more happy with that. They thought that when we have data in Opal, who's going to control what projects are being done? What persons can get access and all that kind of thing? And also because in [name of study] you have to pay a certain amount of money for each paper. But for when you have data in a consortium they don't charge the full price.

While research participant protection and uncertainty about the new technologies (Opal, Mica, DataSHIELD) were primary concerns for the studies, these were not their only concerns. Control of who used their data had intellectual property and fiduciary implications (non-epistemic value). Often the originators of the cohorts involved in the BioSHaRE-EU pilot of the integrated approach to secure co-analysis of data (that is, the PI of the cohort itself) were themselves involved in using the data for scientific ends. In other cases, additional data was produced within the study by external researchers working with original samples (e.g. sequencing samples for particular research interests or disease categories). The resulting new data were viewed by some studies as the 'property' of those researchers who had produced it (non-epistemic value).

As a result the dynamic nature of these studies (i.e. that data collection and sample transformation is ongoing) necessitated upkeep. Funders increasingly expect studies to raise at least part of the funds required for such maintenance, most notably by charging for data access. Arguably, charging for data access constrains access (29) but, at the same time, it may contribute to long-term sustainability; potentially a balance between the number of users 'now' versus the number over the useful lifetime of the study. Control of data access also performs the function of protecting the cohort and the cohort originators or data producers (referred to in this paper as researchers).

Protecting the study and researchers

Studies used two particular means to control who used their data and to therein protect the study and its researchers: first, by ensuring the scientific rigour (apparently epistemic value) of proposed research through scientific review of data access applications; and, second, by controlling duplication of research (underway or upcoming) by study, or other external, researchers (non-epistemic value).

So researchers submit an application to a research project. We look at it from a scientific point of view: Is it feasible? Does it make sense? Are the researchers bona fide researchers? So we don't want any journalists or anybody who's not a scientist accessing a resource for wrong doing. So we have quite a strict registration process. And they can be from academia or from the commercial world, but they have to be a bona fide researchers so that we limit any old Tom, Dick and Harry accessing the database. And if there's people that are submitting

> *applications which strongly overlap then we suggest they collaborate: we put them in touch.*

The practice of reviewing the science was viewed as providing a means of protecting the study from reputational damage (non-epistemic value). Restricting overlapping research served to protect the interests of the producers of study data; both the PI/investigators of the original study and additional researchers who produced new data for the study via their later research. Operationalising protection of study PI or data originator's IP was typically achieved through three alternative responses: (1) duplication was simply not allowed; (2) data originators were given a limited time for exclusive use of the data; (3) researchers with similar ideas were brought together.

Not all studies imposed such restrictions. Practices varied widely, with some respondents declaring such scientific review as "old fashioned". One respondent, themselves a senior bioscientist, stated:

> *Bioscientists who are trying to control what research is done by other people and [who] believe that their way of doing it is better than what those other people are doing are being very arrogant and inappropriate. […] Basically, if you allow infrastructures to be available, then some of the time they will do fantastic science and some of the time they will do rubbish science. But the point is, is that the rubbish science ultimately won't get anywhere and therefore it is not up to us to decide which of those things get through because we can't predict necessarily which is really good stuff and which is the bad stuff.*

Others who also viewed this protectionist stance as "old fashioned" countered with a public good argument (non-epistemic value):

> *This is a kind of old fashioned way to think that "This is my project" and "This is my cohort" and "This is my biobank" - "I've started to collect this when I was in medical school and now when I retire I will give it as an inheritance to my best post doc." It's not yours. If you work in the public office, your salary comes from public salary, your grants come from public tax money: it's everyone's. It's your wife's, it's your children's, it's your neighbours' data. And they should have the benefits of it, not you. Of course, you have some kind of an ownership over that because you have collected it and you maybe know [it] best, but that's an ownership based on your knowledge and skills, not based on ownership [of] property. So I very strongly think that people should be more open about – in parentheses – 'their data'. It's not theirs. It's something they have assisted in collecting, but it's collected by a giant effort of scientists and tax payers.*

Consequences of the current model of data access

By way of further examining the consequences of study values around data sharing, we offer a description of the experience of BioSHaRE-EU researchers in accessing data for a pilot analysis using DataSHIELD. Ethnographic observation and interviews revealed that a key challenge BioSHaRE-EU researchers encountered was gaining access to the data they required to conduct analyses using the integrated approach. Users and developers were equally impeded by the multiple, repetitive and time consuming procedures required to gain permission to use data for analysis.

> *The whole process that starts [...] with applying for data to actually publishing your research paper takes a lot of time because so many actors are involved. This has been a major difficulty for me.*

The complex governance of data access was widely experienced as an encumbrance and major deterrent to research at every stage of the process, not simply at the outset. As research advanced, for instance, the need for additional data might arise. When this happened, BioSHaRE-EU researchers had to again undertake the time consuming process of requesting data and securing its release:

> *It's like writing a mini proposal every time you need to access another bit of data.*

> *I think the biggest challenge [...] has been trying to gather all of the data need[ed] from all of the different cohorts, you know, with different data request forms, different criteria, different speeds, different ethics requirements, you know, from five different biobanks and cohorts in, of course, four very different countries. That's been a huge undertaking, just obtaining the data.*

When asked to describe any specific difficulties in accessing data, one BioSHaRE-EU researcher closely involved in the implementation of the tools identified the crux of the problem:

> *I think, in general, time is always an issue with data access, be it within the BioSHaRE-EU project or outside the BioSHaRE-EU project. Often the delays to data access are quite significant and it delays then the analysis and everything else that comes after accessing the data. […] I think if we were to redo BioSHaRE-EU, it'd be great if we could have a project where a group of five or six cohorts – whoever wants to participate in the project – agrees that there's going to be a group of variables from each of the studies that will be used for multiple research questions rather than doing one off research data access requests for every research question that we have – which is the traditional way of doing it. You build up your research proposal. You send it [to] each cohort that you want to have access to. They review it and then they give you access to the variables that you requested for your project. … I think one thing within BioSHaRE-EU is that we wasted a lot of time preparing the data access requests for each and every one of the research questions and that delayed a lot of the research.*

This description of time delays in gaining access to data presents an account of data sharing *practice as usual*, wherein separate data access applications (and in some cases additional requests for ethical review) are required for "each and every" research question and to each and every cohort involved in the analysis. The impact of this "traditional" approach is to extend considerably the time needed to conduct analysis, particularly when new, unanticipated research questions emerge during analysis. As noted by the respondent, this may happen in any analysis. However it is precisely the analytic flexibility offered by DataSHIELD that is lost under contemporary data access governance.

The "traditional" approach to data access was contrasted with an approach based on permissions attached to use of specific datasets, where permission is sought from multiple cohorts such that access is variable-specific rather than research question-specific. In the case of variable-specific access permission any research question that fell within the bounds of the consents signed by study research participants would be possibly without having to seek access for each specific research question as it emerged, as is currently the case. Thereby flexibility would be made possible, for example, in harmonised datasets for which permission had been granted for each variable.

During the BioSHaRE-EU project, the values shaping data access were largely consistent across studies, though there was evidence of change as seen in the critiques of existing approaches to data access that emerged. Respondents reported that analysis under DataSHIELD could potentially

ameliorate some access concerns while still upholding the core non-epistemic value of cohorts to protect research participants. There remained, nevertheless, a sense that existing systems had become entrenched for historical and contextual reasons; that is, that practices that had developed over time within the specific cultures and practices of each study.

> *I'm afraid that there is no way to avoid that different biobanks are preparing their own protocols. Because I think that people are coming from different cultures and so they have their own idea about the proper way to do things.*

Many of the impediments to change were seen as human rather than scientific or technical challenges.

> *I think that the very big hurdles in science are human hurdles, human relationships. Problems that have nothing to do with science.*

Moreover, the challenges identified time and again by respondents, in these interviews and in the ethnographic observations, arose from the interplay and tension between "competition" and "collaboration". Conflicting values were seen to play out in terms of "scientific turf" and were further entrenched by controlling data access in ways that allowed researchers exclusive or time limited access to data for their own research questions.

While this may seem at odds with notions of open science, there were sound structural reasons for taking this position. In a contemporary science that privileges discovery science and its impacts – where the outputs (peer reviewed papers) of that science are the pathway to securing and retaining professional positions and achieving promotion and recognition (non-epistemic value) – protecting data access is a rational response. Respondents clearly identified a need for other mechanisms by which to recognise infrastructural science.

> *So the general view is moving in the right direction. But one of the things that's also been recognised is that given [individual and funder investment in large bioscience infrastructures], there has to be better ways of measuring people's output other than just simply their papers.*

Some of these mechanisms already exist (e.g. BRIF(29) and proposals for recognition of data resource generation in future national research quality assessments), but these are not yet widely used.

Discussion

What is the way forward for governance of data access?

Each cohort involved in BioSHaRE-EU was constrained by local, national and supra-national institutional and legal guidelines for research. These guidelines varied considerably not only across regions (e.g. EU) but even within single countries. The resultant data access guidelines also varied. Predictably bespoke data access procedures were developed by individual studies at different times as they began to collect and then release data and samples. Such practices then continued to change as local, national or international policy evolved and changed. The combined result has inevitably been a multiplicity of diverse data access governance requirements.

Not surprisingly, the pilot of the integrated approach for co-analysis under DataSHIELD found itself betwixt and between old models of governance and the new modes and technologies of co-analysis. Because DataSHIELD did not allow access to the research participant level data held by studies, it offered the potential for new models of data access governance, with lower thresholds for granting permission and much simpler – possibly multi-study - mechanisms for seeking that access. However,

without such a model already being in place all analysis, to date, using the DataSHIELD approach has still needed to employ existing processes for securing permission to use the data.

Paradoxically, this means that if the integrated approach to secure co-analysis of data is seen as being successful and becomes widely used – with governance mechanisms evolving in recognition of that change – its ease of use and relative benefit will markedly increase. However, in the meanwhile, those developing the approach recognise that they must ensure that the system offers enough benefits (including flexibility and ease of use) and acceptable costs (such as having to seek permission for a new approach that is not well understood by data access boards), that even without the new governance mechanisms being in place, the integrated approach to secure co-analysis of data is still realistically competitive with more traditional approaches.

Crucially, this setting – in which the costs and benefits of the new and traditional approaches are well balanced - provided us with a unique analytic counterpoint for examining the epistemic and non-epistemic values underpinning and shaping contemporary data access governance. Our analysis enabled us to ask the question, "Which values remain in need of protection when privacy-protecting data analysis is possible?" and in turn, "What forms of data access governance is necessary to protect epistemic and non-epistemic values and interests of cohort research participants and researchers?"

Three dominant values emerged: protecting the research participant, protecting the study, and protecting the researcher. These values were both supported by and juxtaposed against a 'public good' argument. Each was used as a rationale either to promote or to inhibit sharing of data. While DataSHIELD ameliorates these fears, studies will need to be persuaded that it is as effective as it claims to be before they will accept this new approach. Ultimately, DataSHIELD may prove especially protective of both research participants and studies since users cannot directly access research participant level data. But studies can still withhold data to protect ongoing work. In one analysis a particular cohort refused access to data which was already being used in an analysis being conducted by a research group unrelated to BioSHaRE-EU.

The originating epistemic driver of data sharing is as much pragmatic as it is principled: bigger infrastructures are needed to answer the questions scientists want to answer, but collaboration is needed to achieve this. An overriding non-epistemic value in the study setting remains the protection of the research participants who provide the data and samples. Another important value is reflected in the understandable desire of researchers to see their efforts in building population biobanks and longitudinal cohort studies realised in the form of scientific findings and outputs. Indeed, the one epistemic value identified (scientific rigor) acted to reproduce non-epistemic values (e.g. 'protecting the researcher').

Reviewing the epistemic and non-epistemic values shaping data sharing and practices of data access enabled us to identify those criteria that most effectively facilitate that access, optimise the value of studies while protecting research participants. On this evidence, systems for governing and enabling data access need to:

(1) protect disclosure of research participant information or identity;

(2) ensure the specific expectations of participants are met (e.g. non-commercialisation, exclusion of particular stakeholders such as tobacco companies);

(3) embody systems of review that are transparent and are not compromised by the specific interests of one particular group of stakeholders (though complete independence is unlikely because

governance of access must be administered by individuals sufficiently knowledgeable and experienced and the field is a small one); and

(4) facilitate data access procedures that are timely and efficient. Good practice is in place for access to individual cohorts but practices need to be adapted to the particular demands of access to multiple cohorts or biobanks.

In order to truly optimise the scientific value of the resources into which so much human and fiscal capital continues to be invested, supra-study and supra-national solutions for access governance are needed for collaborative projects such as BioSHaRE-EU to streamline data access and ameliorate key challenges.

Possibilities include centralised portals for data access applications or joint agreement for blanket approval for the use of particular harmonised variables or dataset of variables. Because roadblocks to data access involve embedded cultural and structural values that hamper data access multi-study, and preferably regional or international, approaches which are mindful of these non-epistemic values need to be explored and tested.

Examples already exist of data governance and access mechanisms that aim to provide harmonised, streamlined data access at national (e.g. METADAC in the UK, (31)) and international levels (e.g. the Population Public Project for Genomics and Society - International Policy interoperability and data Access Clearinghouse (P3GIPAC) (32) or Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) (33, 34), Global Alliance for Genomics and Health (GA4GH)).

However, these are not generally aimed at data access to multiple studies simultaneously for one common analysis. Nonetheless their principles and approaches might be adapted by multi-study consortia. At the same time, new solutions may challenge current funding models which, by default, promote multiple applications and the multiple opportunities for cost-recovery they present. Funders, too, need to be part of the discussion of ways forward.

New consortia of researchers and multiple cohorts and biobanks, like BioSHaRE-EU, continue to be established. We have learned, in the findings presented above, that the complexity of interfacing with multiple governance system in such consortia risks impeding the very science which is the *raison d'etre* of these consortia. Increasingly journals are requiring that open access be made available to the data upon which research papers are based (eg. PLOS journals, F1000, BMC Genomics). Practical solutions are urgently needed and should, ideally, form part of the integrated approach to secure co-analysis; development of a governance component within that approach as it evolves will, necessarily, be designed to fit the range of contexts into which it will be applied.

Any new approach for access to multiple studies must first avoid reproducing or exacerbating features of traditional governance mechanisms that are no longer needed, or have always been questionable. New approaches should be trialled (and formally evaluated) in negotiation with all stakeholders including: scientists and research participants associated with the cohorts and biobanks in these new consortia, potential data users, funders, ethical committees and governance boards with oversight responsibility, and the broader research community.

The need for shared data access will only grow as new and more data-intensive research questions are posed by the scientific community. In particular, there is increasing demand for the deep phenotype, genotype and other -omics data and samples that are offered by cohorts and biobanks but which are only available in large enough numbers when studies can be combined.

Acknowledgements

We would like to thank all participants in the BioSHaRE-EU ethnography and evaluation for their generosity in enabling us to work alongside them and for giving of their time in interviews.

The research and analysis leading to these results was supported by the Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BioSHaRE-EU) program which received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no 261433 and the Managing Ethico-social, Technical issues and Administration Data Access Committee (METADAC) which received funding from the Medical Research Council, Economic and Social Research Council and Wellcome Trust (MR/N01104X/1). The Data to Knowledge (D2K) Research Group is also supported by funding from: the European Union's Seventh Framework Programme BBMRI-LPC (313010) (Biobanking and Biomolecular Resources Research Infrastructure—Large Prospective Cohorts); Medical Research Council and Wellcome Trust, 58READIE project (G1001799/2) (Realizing Easy Access to Data and Infrastructural Enhancement for the 1958 Birth Cohort Biomedical Resource) and ALSPAC project (102215/Z/13/Z), and the Welsh and Scottish Farr Institutes, MRC funded E-Health Informatics Research Centres (EHIRCs) (MR/K006525/1; MR/K007017/1). .

References

1. Murtagh M, Thorisson G, Wallace S, Kaye J, Demir I, Fortier I, et al. Navigating the perfect [data] storm. Norsk epidemiologi. 2012;21:203-209.
2. Murtagh MJ, Demir I, Harris JR, Burton PR. Realizing the promise of population biobanks: a new model for translation. Human Genetics. 2011;130:333-345.
3. Walport M, Brest P. Sharing research data to improve public health. Lancet. 2011;377:537-539.
4. Carr D, Littler K. Sharing Research Data to Improve Public Health A Funder Perspective. Journal of Empirical Research on Human Research Ethics. 2015;10:314-316.
5. Trinidad S, Fullerton S, Bares J, Jarvik G, Larson E, Burke W. Genomic research and wide data sharing: views of prospective participants. Genetics in Medicine. 2010;12:486-495.
6. Rahm A, Wrenn M, Carroll N, Feigelson H. Biobanking for research: a survey of patient population attitudes and understanding. Journal of community genetics. 2013;4:445-450.
7. Knoppers B, Harris J, Tassé A, Budin-Ljøsne I, Kaye J, Deschênes M, et al. Towards a data sharing Code of Conduct for international genomic research. Genome Med. 2011;3:46.
8. Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. International journal of epidemiology. 2010;39:1383-1393.
9. Fortier I, Doiron D, Little J, Ferretti V, L'Heureux F, Stolk R, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. International journal of epidemiology. 2011;40:1314-1328.
10. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel B, Perola M, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerging themes in epidemiology. 2013;10:12.
11. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones E, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. International journal of epidemiology. 2014;43:1929-1944.
12. van Vliet-Ostaptchouk J, Nuotio M, Slagter S, Doiron D, Fischer K, Foco L, et al. The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. BMC endocrine disorders. 2014;14:9.
13. BioSHaRE-EU. BioSHare-EU website. www.bioshare.eu. 2015. [accessed November 6, 2015]

14. Burton PR, Murtagh MJ, Boyd A, Williams JB, Dove ES, Wallace SE, et al. Data Safe Havens in health research and healthcare. Bioinformatics. 2015:btv279.
15. Murtagh M, Turner A. Report: Deliverable D3.7 – Report on the social and epistemic implications of the DataSHIELD methodology https://www.bioshare.eu/sites/default/files/1/D3.7%20Report%20on%20the%20social%20and%20epistemic%20implications%20of%20the%20DataSHIELD%20methodology%20-%20incl%20date.pdf. 2015. [accessed 8 January 2016]
16. Murtagh M. Report: Deliverable D3.8 - Report on Social and epistemic implications of biobank standardisation and harmonisation https://www.bioshare.eu/sites/default/files/1/D3%208%20Report%20on%20Social%20and%20epistemic%20implications%20of%20biobank%20standardisation%20and%20harmonisation.pdf. 2015. [accessed May 11, 2016]
17. Murtagh M, Minion J, Fay M. Report of WP3 -Objective 3: To formally evaluate the application and utility of new harmonization and standardization tools developed and rolled out under the BioSHaRE-EU project. https://www.bioshare.eu/sites/default/files/1/EnMESHD-FinalReport%20incl%20date%20and%20TOC.pdf. 2015. [accessed May 11, 2016]
18. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4(8):e1000167.
19. Denzin N, Lincoln Y. The SAGE handbook of qualitative research: Sage, London; 2011.
20. Hammersley M, Atkinson P. Ethnography: Principles in practice: Routledge, London; 2007.
21. Glaser BG. The constant comparative method of qualitative analysis. Social problems. 1965;12:436-445.
22. Geertz C. The interpretation of cultures: Selected essays: Basic books, New York; 1973.
23. Austin J. How to do things with words: Oxford university press, Oxford; 1975.
24. Searle J. Speech acts: An essay in the philosophy of language: Cambridge university press, Cambridge; 1969.
25. Foucault M. The order of things: An archaeology of the human sciences: Routledge, London; 2002.
26. Butler J. Gender trouble: Routledge, London; 2002.
27. Potter J, Wetherell M. Discourse and social psychology: Beyond attitudes and behaviour: Sage, London; 1987.
28. World MAGA. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Journal international de bioéthique= International journal of bioethics. 2004;15:124.
29. Gilbert R, Goldstein H, Hemingway H. The market in healthcare data. BMJ. 2015;351.
30. Cambon-Thomsen A, Thorisson G, Mabile L, Andrieu S, Bertier G, Boeckhout M, et al. The role of a bioresource research impact factor as an incentive to share human bioresources. 2011.
31. Murtagh MJ. METADAC Website. www.metadac.ac.uk. 2015. [accessed November 11, 2015]
32. P^{3}G. International Policy interoperability and data Access Clearinghouse (IPAC) http://www.p3g.org/ipac. 2016. [accessed May 11, 2016]
33. Litton, JE. BBMRI-ERIC website. www.bbmri-eric.eu. 2015. [accessed November 6, 2015]
34. Perola, M. BBMRI-LPC website. www.bbmri-lpc.org. 2015. [accessed November 6, 2015]

[i] The terms 'cohort' and 'biobank' were used interchangeably by participants in the ethnographic research presented in this paper. For clarity, the term 'study' – also used by participants – refers here to both population biobanks and longitudinal cohorts. The term cohort is used to denote the 15 cohort studies involved in the BioSHaRE-EU project.

[ii] This figure shows the infrastructure setup used in the BioSHaRE-EU project. However, the infrastructure is quite flexible and this represents only one possible configuration. For example, the analysis

server could be removed altogether and instead researchers would each install the R client and DataSHIELD client packages locally.

[iii] This understanding has been questioned following Homer *et al* (18)*,* though to identify an individual under Homer would require a reference sample or data from the individual concerned and therefore is more a matter of forensic application than public disclosure, which has been the focus of concerns about privacy.