Public Health Genomics

Public Health Genomics 2012;15:243–253
DOI: 10.1159/000336673

Published online: June 20, 2012

Securing the Data Economy: Translating Privacy and Enacting Security in the Development of DataSHIELD

M.J. Murtagh[a] I. Demir[a] K.N. Jenkings[a] S.E. Wallace[a] B. Murtagh[a]
M. Boniol[b] M. Bota[b] P. Laflamme[c] P. Boffetta[b, e] V. Ferretti[d] P.R. Burton[a]

[a]Data to Knowledge for Practice, University of Leicester, Leicester, UK; [b]International Prevention Research Institute (iPRI), Lyon, France; [c]Public Population Project in Genomics (P^3G), Montreal, Que., and [d]Ontario Institute for Cancer Research, Toronto, Ont., Canada; [e]Mount Sinai School of Medicine, New York, N.Y., USA

Key Words
Data economy · DataSHIELD · Privacy · Transdisciplinary · Video ethnography

Abstract
Contemporary bioscience is seeing the emergence of a new data economy: with data as its fundamental unit of exchange. While sharing data within this new 'economy' provides many potential advantages, the sharing of individual data raises important social and ethical concerns. We examine ongoing development of one technology, DataSHIELD, which appears to elide privacy concerns about sharing data by enabling shared analysis while not *actually* sharing any individual-level data. We combine presentation of the development of DataSHIELD with presentation of an ethnographic study of a workshop to test the technology. DataSHIELD produced an application of the norm of privacy that was practical, flexible and operationalizable in researchers' everyday activities, and one which fulfilled the requirements of ethics committees. We demonstrated that an analysis run via DataSHIELD could precisely replicate results produced by a standard analysis where all data are physically pooled and analyzed together. In developing DataSHIELD, the ethical concept of privacy was transformed into an issue of security. Development of DataSHIELD was based on social practices as well as scientific and ethical motivations. Therefore, the 'success' of DataSHIELD would, likewise, be dependent on more than just the mathematics and the security of the technology.
Copyright © 2012 S. Karger AG, Basel

Data Sharing in the New Data Economy

Contemporary bioscience is seeing the emergence of a new data economy [1]: that is, the advent of new forms of social, ethical, institutional, academic, epistemic, national, and international structure and governance, with data as its fundamental unit of exchange. Increasingly, exploration of biomedical and social determinants of health and disease require identification and quantification of the relatively weak effect of one or more factors of primary relevance (e.g. a number of specified genes and environmental determinants) that are shrouded behind a smoke screen of other factors that are causally important but not of substantive interest (i.e. all of the other determinants that influence the trait of interest). In the biomedical setting, this means that very large numbers (10's or 100's of thousands) of participants are often needed [2].

KARGER

E-Mail karger@karger.ch
www.karger.com/phg

© 2012 S. Karger AG, Basel
1662–4246/12/0155–0243$38.00/0

Karger Open access

This is an Open Access article licensed under the terms of the Creative Commons Attribution- NonCommercial-NoDerivs 3.0 License (www.karger.com/OA-license), applicable to the online version of the article only. Distribution for non-commercial purposes only.

Madeleine Murtagh
Data to Knowledge for Practice, University of Leicester
Room 209, Adrian Building, University Road
Leicester, LE1 7RH (UK)
Tel. +44 116 252 2926, E-Mail mm399@le.ac.uk

At the same time, studies must invest in measurements that are defined as being of high quality [3], and this can be very expensive. Moreover, this places a pragmatic limit on the total number of participants that any single study can enroll, and many research questions of undoubted scientific interest simply cannot be answered by using data from one study alone. Instead, research groups and consortia are increasingly combining data from more than one study to carrying out large pooled meta-analyses [2, 4]. Although sharing data within this new 'economy' provides many potential advantages, the sharing of individual data raises important social and ethical concerns among the public, researcher, and funding and governance communities [1, 5–7].

Many studies and governmental data repositories have strict embargoes on the release of individual-level data [8]. These may be framed in the wording of consent forms or information leaflets, the concerns of ethical committees, or legal restrictions placed by governments or other controlling bodies on national data access and release. Sometimes the prohibition is all encompassing (no data can be passed to *any* third party), but it is often targeted (e.g. data cannot be passed across national boundaries; social data but not biomedical data can be shared). Where access *is* permitted, governance procedures may well stipulate that researchers must seek formal permission from one or more oversight bodies and/or ethics committees, which likely involves procedures that are onerous and time-consuming. This presents an important challenge. From a scientific perspective, methods that work directly with individual-level data are markedly more flexible, and often more efficient, than other approaches to pooled analysis. Governance restrictions reflect consideration of privacy, confidentiality, and the ownership and exploitation of scientific data and intellectual property. Arguably of great public concern is privacy. Indeed, the desire to advance bioscientific understanding of health and disease may, prima facie, seem incommensurable with concerns for individual privacy. But this all-too-common dualism pitting science against society provides little leverage for understanding the development of technologies that might address both scientific *and* social or ethical concerns.

Taking, for the purposes of this special issue of *Public Health Genomics*, privacy as the locus of our concerns, we examine the ongoing development of one technology, DataSHIELD, which appears to elide privacy concerns about sharing data by enabling shared analysis while not actually sharing any individual-level data. Combining the presentation of the development of DataSHIELD, including the outcomes of a workshop to test the technology, with a presentation of an ethnographic study of that workshop, we explore the coproduction of privacy and security with DataSHIELD. First, we describe DataSHIELD and outline the ethicolegal perspective on privacy, for this is the benchmark to which DataSHIELD is required to adhere. We then describe and discuss the workshop and the ethnographic study and their implications for the development of DataSHIELD.

Rationale for Developing DataSHIELD

Methods for performing pooled analyses that do not breach ethicolegal and governance stipulations already exist. Much of the recent progress in understanding the genetic variants that cause a range of common chronic diseases (e.g. cancer, coronary artery disease, diabetes, asthma, and arthritis [4]) has arisen from pooled analyses where the analysis is based on SLMA (study-level meta-analysis) [9] or aggregate-level analysis [10]. Here the association between each gene and the disease of interest is first estimated in each study separately. The study-specific measures of the magnitude of each association (e.g. odds ratios) are then taken and combined to calculate an overall odds ratio for all studies together. This is in effect a weighted average of the odds ratios in each of the separate studies that takes appropriate account of the amount of information in each study (e.g. large studies count more heavily). This approach is very effective. Furthermore, in most settings the overall odds ratio obtained from the SLMA is very similar – both in size and uncertainty (e.g. its 95% confidence interval) – to the result that would have been obtained if the analyst had actually been able to take all of the individual-level data from each participant in all of the studies combined and pooled them all together in one large data file and then carried out a single global analysis. A single large analysis of this latter nature typically requires that appropriate allowance is made for systematic differences (heterogeneity) between studies (often called adjustment for the effect of center), and it may then be called a full ILMA (individual-level meta-analysis) [9, 10].

Given the potential to use either SLMA or ILMA, why does pooled data analysis still present unresolved challenges? SLMA avoids the need to pass *any* individual level data from the collaborating studies to the analysis center. In consequence, many ethical committees, study oversight committees and scientific advisory boards have concluded that even if ethicolegal restrictions prohibit

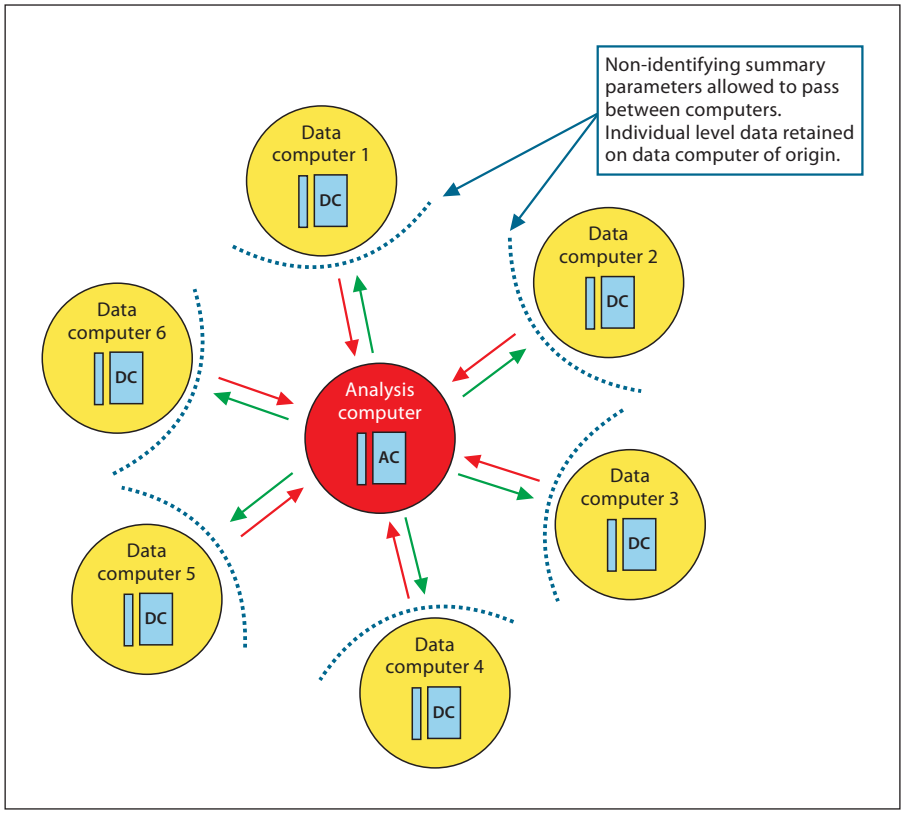
the transmission of individual-level data to third parties (thereby preventing conventional ILMA), it is entirely acceptable for study investigators to analyze their own data and then to pass the summary statistics generated by those analyses to an analysis center to provide a foundation for a pooled SLMA. This indicates an implied principle that – provided they carry no sensitive information about individual participants and no information about their identity – mathematical summaries *can* be passed to a third party for the purpose of pooled analysis even if the transfer of individual level data is prohibited. However, despite its many benefits, conventional SLMA is inflexible in one critical regard: a given pooled analysis can only be undertaken if it meets 3 conditions: it has been defined ahead of time, the investigators in every study have carried out the required analysis, and they have transmitted the appropriate summary statistics to the analysis center. Imagine, for example, 3 studies that have each estimated the 500,000 associations between 500,000 genetic variants and a disease of interest, and have each passed the summary statistics reflecting these associations to the analysis center. SLMA can easily be used to estimate the overall associations (pooled across all 3 studies) for each of the 500,000 variants. But whereas a simple – indeed, obvious – next step in the analysis of a single study would be to determine whether there are any sex-specific effects among those variants that exhibited a significant association, such an analysis would be fundamentally impossible under SLMA unless each study had also been asked *ahead of time* to provide summary statistics for the 500,000 sex-gene interactions reflecting each of these sex-specific differences.

Researchers are increasingly faced with analyses that involve assessments on a wide variety of important demographic factors, vast numbers of genetic variants, numerous biomarkers, and many other measures of the physical and social environments. Furthermore, the range of different models that may be fitted is, for practical purposes, almost limitless (e.g. different measures of disease, alternative combinations of potential causative factors). It is therefore clear that a rational approach to analysis will usually involve a substantial element of exploration. Unfortunately, this exploration will be prohibitively inefficient if it is undertaken as part of a conventional SLMA: where progress is halted after every analysis as the investigators in all studies are asked to produce the new summary statistics that are required to move the analysis forward. Ideally, therefore, an approach to pooled analysis is needed that is as flexible and efficient as ILMA and yet is as fundamentally secure as SLMA; the latter security being guaranteed by avoiding the need for any individual-level data to be transmitted to the center undertaking analysis. DataSHIELD exhibits these characteristics.

How Does DataSHIELD Work?

Figure 1 illustrates the basic IT infrastructure that underpins DataSHIELD. It describes a hypothetical implementation based on a pooled analysis involving data from 6 studies. Most crucially, all of the individual-level data that provide the basis of the analysis remain securely on data computers (DCs) at their home bases. An additional computer is identified as the analysis computer (AC). This is the computer on which the statistician will type the commands to enact and control the pooled analysis. In actuality, the analysis computer may be based at the same center as one of the data computers, or it may be one of the data computers – but it is simpler to envisage if it is assumed (as in the figure) that the AC is independent of all of the DCs.

Given the IT configuration implied in figure 1, a pooled analysis based on a conventional ILMA would require that the data from each of the DCs was first transmitted to the AC. The AC would therefore host the data for analysis as well as provide the point of entry for the instructions of the statistician. Under DataSHIELD, in contrast, the data remain on the DCs, and the AC serves primarily as an entry port for the analytic instructions and as a mathematical platform for integrating the analytic output. This is rendered possible because the analysis itself is parallelized. That is, rather than all of the data being analyzed at once – as in a conventional analysis – the data from each study are analyzed separately but contemporaneously. Such an analysis typically starts with the AC making a guess as to the 'true' results that will ultimately be generated from the pooled analysis and transmits this guess to all of the DCs (arrows pointing from the AC to the DCs in fig. 1). Crucially, the quality of this first guess does not impact the final result. The analysis then proceeds in steps (iterations). At the start of each step, the AC transmits a set of analysis instructions that tell each DC to run one step of the analysis (starting from the current best guess at the ultimate result of the analysis). At the end of the step, each DC returns summary statistics to the AC, and these characterize the current state of the analysis in that particular DC (arrows pointing from the DCs to the AC in fig. 1). When the AC has received the summary statistics from all DCs, the AC



Color version available online

Fig. 1. IT infrastructure underpinning DataSHIELD.

combines them to produce a series of indicator values that collectively tell the AC how to optimally modify the current best guess at the ultimate result or results to produce a new guess that is closer to the true results. The AC then relays this new best guess to each of the DCs, followed by the instructions to carry out the next step. Ultimately, after a small number of steps (typically 4–5), the best guess at the ultimate result or results of the analysis remains unchanged from step to step. The analytic process is then said to have converged, and the best guess is taken to represent the final answer.

For a broad class of mathematical models that are collectively known as generalized linear models, the results obtained using the approach described in the preceding paragraph are mathematically identical to those that would have been obtained if the required data from all of the collaborating studies had been placed into one large data file, and that file had then been analyzed directly – i.e. a full, conventional ILMA. However, unlike a conventional individual-level meta-analysis (see above), the information that has to be transmitted back and forth between the DCs and the AC consists only of analytic instructions and summary statistics, and these carry no sensitive information and are nonidentifying. Thus, DataSHIELD enables a pooled data analysis to be carried out as if one had full access to all of the individual-level data, although these data are actually held securely behind IT shields (short, dashed lines in fig. 1). This approach is very useful because many of the quantitative analyses that are undertaken most commonly in biomedical and social research can easily be framed as generalized linear models: these include basic contingency table analysis and many foundation methods for quantitative variables (such as t-testing and basic analysis of variance). It also includes many of the most widely used classes of regression analysis (e.g. linear, logistic and Poisson regression) and several types of survival analysis.

Ethicolegal Perspectives on Privacy

There is no agreed legal definition of privacy in either UK or US law. In 1890, Warren and Brandeis [11] described privacy as 'the right to be let alone,' in response to the new invasive technology of photography and its use

by newspapers to 'invad[e] the sacred precincts of private and domestic life'. This concept of privacy is but one of many perspectives; legal and other scholars have long attempted to derive a definitive characterization of privacy and the right to privacy, with little success. Gormley [12] notes that such attempts have been made by a range of scholars, from those who believe privacy was 'an expression of one's personality or personhood' to those who argued it was 'the moral freedom of an individual to engage in his or her own thoughts,' through to the stricter stance that privacy was about '[a] citizen's ability to regulate information about themselves, and thus control their relationships with other human beings'.

A further problem in medical ethics is the common conflation of privacy and confidentiality. While related, they are different – the state of privacy (being left alone) can be enabled and protected by keeping something confidential (enabling 'privacy' for someone else) – but they are often used interchangeably. Considerations of the protection of privacy in medical law tend to focus on breaches of confidentiality in the doctor-patient relationship. It is to this understanding that many implicitly refer when considering privacy in the context of medical research. Medical professionals have a duty of care to their patients to maintain confidentiality by keeping personal information private [13]. This is reflected in UK case law. For example, von Egdell [14] explored whether disclosure of personal information in the public interest could outweigh the duty of confidentiality. Thus, duty of care is carried into research ethics through institutionalized ethics and governance procedures and constraints.

In the US, Gormley [12] argued that legal privacy 'is heavily driven by the events of history'. That argument can also be made in relation to law in the UK. As a result of several high-profile cases and the enactment of the Human Rights Act of 1998, the courts are developing a new tort of breach of privacy [15]. For example, in Campbell versus Mirror Group Newspaper Ltd. [16, p. 995], the Lords agreed that information disclosed about model Naomi Campbell and her attendance at Narcotics Anonymous meetings was private, not public, and compared such details to medical records data: 'No distinction was to be drawn between the details of the claimant's therapy from Narcotics Anonymous and detail of a medical condition or its treatment. Those details were private information which imported a duty of confidence.' However, in his dissenting opinion, Lord Nicholls noted that whereas previously an initial confidential relationship between parties was needed for there to be a duty of confidence, this now was changing:

> Now the law imposes a 'duty of confidence' whenever a person receives information he knows or ought to know is fairly and reasonably to be regarded as confidential … The continuing use of the phrase 'duty of confidence' and the description of the information as 'confidential' is not altogether comfortable. Information about an individual's private life would not, in ordinary usage, be called 'confidential'. The more natural description today is that such information is private. The essence of the tort is better encapsulated now as misuse of private information [16, p. 1002, 1003].

This tort of privacy may be pertinent for DataSHIELD. An inappropriate transfer of private information could be seen as misuse, which would place it under a new tort of privacy. However, there is a difference between researchers using data from research participants and disclosure of medical information to medical professionals in a doctor-patient relationship. While the latter owe a duty (of care and) of confidence, it is not clear whether this also holds for the researcher, as they have a very different relationship with the participant. While researchers may not owe a duty of confidence, per se, to their participants, they would probably recognize their duty to not misuse private information. However, as Laurie [17] cautions us, we cannot rely solely upon the law as a remedy for ethics and governance. Additional solutions are required.

The Workshop and Ethnography

Drawing on Nowotny's and others' [18–21] concepts of mode 2 knowledge and transdisciplinary science, we conducted a deliberatively transdisciplinary study that included epidemiological and informatics development of DataSHIELD in and through an integration workshop and an ethnographic study of that workshop. Both studies are part of an ongoing program of work that aims to inform the development of DataSHIELD, gain an understanding of transdisciplinary science in genomics, and contribute a more socially and theoretically informed development of that science.

The DataSHIELD Implementation Workshop was designed to take key BioSHaRE-eu researchers, and 4 international research groups considering making use of DataSHIELD approach, through the theory and practicalities of setting up and running a DataSHIELD analysis. Participants were welcomed to the workshop on behalf of BioSHaRE-eu, P^3G and the DataSHIELD Project by Paul Burton (P.B.), and to the International Prevention Research Institute (iPRI) by its chief operating officer, Markus Pasterk. P.B. outlined the aims and objectives of the workshop. This was followed by round-table intro-

ductions of everyone present, with participants indicating their reasons for interest in DataSHIELD and in the workshop. P.B. gave a brief introduction to the ideas and principles underpinning DataSHIELD and to some of the current thinking about its potential uses and the opportunities and challenges that it presents. This was followed by an in-depth description, and live demonstration, of the DataSHIELD implementation in OPAL (see http://www.obiba.org), an open source database software application designed for epidemiological studies that can run DataSHIELD analyses on data computers via R-Opal, a specialized R client interface that is installed on the analysis computers. Participants worked individually, or in pairs, on a series of practical exercises that took them through the process of: (1) setting up a DataSHIELD session via R-Opal, (2) simple secure (nonidentifying) descriptive analysis of the data on the remote servers using the aggregating functions in DataSHIELD, and (3) secure multivariate analysis using generalized linear models via the glm.datashield function. The practical was based on 2 simulated datasets representing a hypothetical 2-center study aimed at exploring a candidate gene and body mass index as putative causes of early-onset coronary heart disease as reflected in a myocardial infarction before the age of 55 years. The simulation deliberately incorporated study-study heterogeneity in the sampling fraction of controls (influencing the regression intercept) and in the impact of body mass index on the risk of myocardial infarction. The required analysis was motivated as a conventional logistic regression model. Vincent Ferretti (V.F.) and Philippe LaFlamme (P.L.) then led a practical session that took workshop participants through the processes of: preparing real data in OPAL on several servers, linking R at a central analysis computer to those instances of OPAL, preparing the local data for analysis in R, and running a DataSHIELD analysis. This work was based on a real multicenter dataset involving data from 5 different studies exploring the relationship between sunbed use and melanoma [22]. iPRI had ethicolegal and scientific permission to collectively analyze these data in-house and had published the results of that analysis [22]. For the purposes of the DataSHIELD analysis, the data from the 5 studies were distributed across 5 servers in Canada and at iPRI.

The workshop was videorecorded and observed by the social science ethnographer, Madeleine Murtagh (M.M.), who was also an active participant in the workshop. The videoethnography component of the project had several aims: (1) to permit us to understand the social and professional challenges and opportunities arising in the context of developing a new technologically sophisticated tool in the highly transdisciplinary setting of contemporary bioscience in general, and biobanking in particular; (2) to help steer the DataSHIELD project to address such challenges and opportunities as and when they arise, and (3) to generate a number of targeted video films describing the aims, objectives, methods, and challenges associated with DataSHIELD. Rather than there being an objective and disengaged study of the sociotechnological development of DataSHIELD, the lead (social science) researcher M.M. was integrated into the development process [23, 24]. The study methodology adopted was an ethnography (participant observation) utilizing the research methods of ethnographic observation and note taking through participation (in the workshop, in this instance). A visual anthropologist, Barnaby Murtagh, was engaged to collect audiovisual data [25] at the workshop, which was to be used for both ethnographic video analysis [26] while also furthering the development of a short film about DataSHIELD. Following the workshop, the video data were edited into chronological sections of the various stages of the workshop. These we then posted to a Cloud for access and initial analysis and familiarization by the ethnographer and visual anthropologist, plus 3 other researchers: 2 sociologists Ipek Demir and Neil Jenkings and an ethicolegal scholar, Susan Wallace. After the familiarization period, a 5-day video analytic workshop was hosted where the researchers systematically examined the data, facilitated by the ethnographic notes and observations of the participant observer. Investigation of the workshop as a social process in the development of DataSHIELD focused on the locally emergent and situated practices and discussions, both implicit and explicit, of the issue of privacy. In the discussion that follows, workshop presenters (and authors of this paper) are identified by name when they are the source of a particular quotation; workshop participants are not identified by name, per their research consent agreement.

Transforming Privacy and Enacting Security

As we have stated, there is no legal definition of privacy. A fundamental epistemological obstacle to defining and enacting privacy lies in its transformation from an impact on an individual to an abstract and generalizable phenomenon embedded in ethics and cognate disciplines. Problems arise when professionals and institutions, including researchers, ethics committees and funders, attempt to translate these abstract concepts back

into the real world of socially embedded practice. The law is able only to identify which breaches of privacy will afford its intervention, this, as above, in terms of a duty of care or a duty of confidentiality.

Privacy, much deployed in arguments about ethics, rights and human freedoms, is a slippery concept [27]. Solove [28] identifies 6 general conceptions of privacy: the right to be let alone (as in [11]), limited (unwanted) access to the self, secrecy, control over personal information, personhood, and intimacy. Solove notes the overlaps between these. Privacy is supremely contextual; it changes over time, across settings and cultures such that multiple, sometimes contradictory, understandings of privacy and what is rightly private have coexisted and continue to coexist in society. In relation to the role of privacy in the development of personhood, for example, there are coherent arguments both for the protection of privacy (specifically surveillance) as inherent to the development of personhood [29] and, reflecting Goffman's work [30, 31] on the presentation of self, to advocate publicness as central to development of the self [32, 33].

Feminists point to the rhetorical use of privacy in the history of legal sanctions against domestic violence [34, 35], citing the ongoing use of a discourse of privacy to warrant nonprosecution of 'wife beaters' for most of the 19th century following the outlawing of domestic violence. One judgment (in favor of the defendant) in State v. Rhodes (1868 WL 1278, N.C. 1868) determined 'not [to] inflict upon society the greater evil of raising the curtain upon domestic privacy, to punish the lesser evil of trifling violence.' The judgment carefully avoids sanctioning 'wife beating' per se, but the effect of its privileging of domestic privacy is to do just that. We call this function of language 'performative' [36, 37], that is, the deployment of certain discourses, rationales or arguments to warrant a point of view of set of actions or nonaction. We are reminded by these historical observations to look to the performative character of contemporary discourses of privacy to examine the effects of these discourses in relation to the data economy, data sharing and DataSHIELD. Discourses of privacy that afford primacy to the protection of participant information over other public goods (improvements in health care or understanding of disease, for instance) provide a rationale for inhibiting data sharing. Discourses of privacy combined with rationales about the provenance and quality of data and the respective intellectual property rights of researchers, participants, bioresources, research and academic institutions, and industry can together provide a strong argument against data sharing. The point is not that these arguments are right or wrong, but that we must be attentive to the effects of those arguments in thinking about DataSHIELD. We focus, as per Dourish and Anderson [27]:

> on the practical and discursive elements of privacy and security, which ask not what privacy and security *are* but rather what privacy and security *do*. [27, p. 322]

In one sense, DataSHIELD is itself a product of discourses of privacy. Maintaining the confidentiality of individual-level data is a prime driver for DataSHIELD. DataSHIELD would not be necessary in the absence of privacy concerns. During the workshop, there was a more or less established, though not defined, social convention among participants with regard to what privacy meant. Throughout the workshop, participants referred to the importance of attending to research ethics and privacy in particular. Privacy was the DataSHIELD workshop's leitmotif. In the introduction to the workshop, P.B. described privacy and the Ethical, Legal and Social Implications (ELSI) restrictions to data sharing as the key reason for investigating DataSHIELD:

> P.B.: Many societies and studies are not very happy about the idea of releasing individual-level data.

While privacy was a key driver for DataSHIELD, it was certainly not the only one, nor possibly the most pressing one. Data release is fundamental to examining the new bioscience questions. But accessing data can be a slow and frustrating process. Using the example of a European biobank, P.B. stated:

> P.B.: We want to release data to people as soon as possible when they request it, but even when we're pushing as hard as we can it typically takes a couple of months to get [from the] initial application to having the data.

The perceived slowness of pace of data access (including the impact of constraints imposed by ethics committees and other governance bodies) was set against competitiveness in the field:

> P.B.: It's moving at such a pace that you can't afford to wait that long because people are publishing stuff with those sorts of timescales, and therefore, you get beaten by your competitors. So basically there is a big problem with being able to access the data rapidly enough.

Central drivers for researchers were scientific development, intellectual property, and career progression, and thereby the maintenance of funding streams for themselves and the teams they work with. Likewise, other drivers were described that had 'nothing to do with the ethics,' for example, problems of handling and managing the increasing scale of contemporary datasets [1, 38]. In this way, privacy concerns morphed, in part, into issues of re-

search practice. This transformation did not replace concerns about privacy. Rather, these coexisted, and the essential tension between allowing access to data (and hence robust science) and protecting privacy remained the central concern of the workshop.

The validity of adopting a DataSHIELD approach to pooled analysis when ethicolegal constraints prohibit the release of individual level data from at least one of the collaborating studies depends on two preconditions. First, the pooled analysis that is required must be able to be carried out using a parallelized approach based on a series of local analyses that can be linked together by passing analytic instructions and summary statistics between a single AC and a series of DCs. Second, neither the analytic instructions nor the summary statistics must carry information that could be viewed as being equivalent to individual-level data and might therefore be sensitive to, or might reveal the identity of, individual study participants. In relation to its information content, any analysis that is currently viewed as being ethicolegally acceptable if undertaken as part of a conventional SLMA should be viewed as equally acceptable under DataSHIELD. This is because the type of information that is transmitted is the same. The only procedural difference is that, under DataSHIELD, the AC controls the analysis rather than the local statistician at each study. But this highlights a way in which precondition 2 could potentially be violated. Namely, a malevolent statistician might deliberately set out to undertake a series of analyses that are jointly identifying. For example, this might represent a form of residual disclosure [39]. If a target participant in a particular study is known to be aged exactly 17 years and 312 days, and one wants to know whether he has schizophrenia, one might ask 2 sets of apparently innocuous questions: (1) how many people in the data set are aged less than 17 years and 312 days, and what is the prevalence of schizophrenia in that subset; (2) how many people in the data set are aged less than or equal to 17 years and 312 days, and what is the prevalence of schizophrenia in that subset? When asked in combination, it is possible for individually innocuous questions to reveal sensitive information. This may occur accidently or deliberately, and it is not in any sense specific to the use of DataSHIELD – residual disclosure can be problematic in *any* data set.

The need for reassurance about DataSHIELD's capacity to maintain privacy was a repeated theme of the workshop. Not for the participants themselves: they were already largely convinced of the value of the technology and had gathered to explore its development and potential. Rather, they expressed concern during the workshop about how they would 'persuade' others of the validity and value of DataSHIELD and the protection it afforded. In the context of very strict access restrictions to national databases in their country, one participant was acutely concerned about assuring a range of potential users, asking the developers (P.B., V.F., and P.L.):

> So what can I show people that don't have IT expertise to convince them that if we do this I will be able to make data available to researchers who we want to give data to and do collaborative analysis but in a secure environment?

This participant persisted in seeking clarification relevant to a range of users, asking:

> OK, but say you worked at our IT department? Your job is to make sure that ABSOLUTELY NOBODY access[es] the national databases.

How to 'convince' others is a relational issue requiring a solution that involves social practices rather than only technical or scientific ones. While this was not resolved, and was indeed beyond the scope of the workshop, it remained a recurring theme. Within the paradigm of the workshop, a resolution *was* achieved through privacy concerns being transformed into an issue of security. The visual ethnographic data demonstrated that, for the participants, protecting privacy meant enacting security. As one workshop participant put it:

> Raw data [of individuals] remains completely secure on the peripheral data computers … We are able to work with the raw data but rather than moving the raw data to the center to do the analysis to each of the data computers here … we simply link all those together by passing backwards and forwards summary statistics that relate to the data.

Another workshop participant summarized it this way:

> DataSHIELD allows the … tracking [of] data … without actually getting the data. So in a way you are making the data available without that machinery sending it and then have an easier way to deal with the problem [of privacy]. You don't have to give out whole genomes [to researchers in other settings].

DataSHIELD produced, in the view of the participants, an application of the norm of privacy that was practical, flexible and operationalizable in researchers' everyday activities, and one that fulfilled the requirements of ethics committees. At the meta-level, there was a practical understanding of privacy as the 'nontraceability' and the 'nonidentifiability' of individual participants who had previously agreed to make their data available to researchers. But this needed to be operationalized and converted into scientific practice.

Further, there was recognition that enacting security required resilient systems against two types of potential

breaches. First, as identified by the participants, it meant that DataSHIELD should be resilient against independent hackers. Secondly, it meant that DataSHIELD should be robust against unscrupulous scientists who have legitimate access to the data but abuse that right. One participant noted, 'We've got a series of recommendations for implementation and, you're right, there is a difference between a full hacker and that [unscrupulous user]' and another participant, 'a full hacker is obviously something different.' The distinction that the researchers made is important, as each requires different sorts of security interventions and consequently different ways of protecting privacy. While the independent hacker problem is seen to require piloting, the unscrupulous scientist problem needs the introduction and implementation of formal and informal sanctions from the scientific community.

P.L.: We are going to pilot trying to hack it. If we find that [there are weaknesses] we'll basically have to have these functions rewritten in something else that is fundamentally more secure.

Another workshop participant stated:

So you've got researchers who may be [inadvertently abusing the system] … there'd be a sanction against them.

As an analogy, high-quality security measures may be used to prevent theft from a bank account if the culprit is an outsider who has no permitted access to that account, but no technology can provide absolute protection against the actions of a bank official with legitimate rights of access to a bank account but who abuses those rights. Rather, social measures – including sanctions – are necessary to inhibit such behavior. In the same way, dealing with unscrupulous scientists who misuse their legitimate access to sensitive data requires a social solution to a technical problem. This is the same problem facing other technologies, including SLMA.

The process that started off as a socioethical problem (i.e. the protection of privacy) was operationalized and converted into a technical issue (DataSHIELD), then through the technical problems of how to deal with unscrupulous users of DataSHIELD we are transported back into the socioethical realm (i.e. the implementation of professional sanctions against unethical scientists). The resilience of DataSHIELD against unscrupulous scientists is possible via socially negotiated sanctions that the community introduces and upholds. Sanctions gain an added importance in an internationally transdisciplinary community, such as the population biobanking community. Biobanking involves researchers from a wide range of disciplines who may never meet via e-mail, never mind face to face [38, 40]. Therefore, the usual discipline-based sanctions against unscrupulous users may not work. Moreover, it is the question of translation into research practice that was a central concern to participants; and that translation was to be achieved via enrolment (cf. Callon [41]): that is, persuading other stakeholders (policymakers, funders, ethics committees and other researchers) of the safety and effectiveness of DataSHIELD.

Developing DataSHIELD

By the end of the workshop, the group had demonstrated that an analysis run via DataSHIELD could precisely replicate results produced by a standard ILMA where all data are physically pooled and analyzed together. With the individual data not accessible and not accessed, the ethical concept of privacy was transformed into an issue of security. Provided security is maintained, invasion of privacy, and thus the disciplinary ethics of privacy may be largely elided via the use of DataSHIELD. Nonetheless, at this stage of development we recommend that DataSHIELD be implemented only if a number of safeguards are in place: (1) Ethical approval to use DataSHIELD should be obtained from all of the ethical committees that oversee the collaborating studies. (2) All participating scientists and statisticians who might access any output from – or influence any input to – DataSHIELD should sign a formal confidentiality agreement. (3) All information passed to and from each DC should be permanently recorded so that, should a breach of security occur – either by accident or by design – it can be identified and/or investigated post hoc. (4) No new class of model should be fitted using DataSHIELD until the information content of its summary statistics has been comprehensively explored and is thoroughly understood. (5) From the scientific perspective, any data to be pooled under DataSHIELD should be adequately harmonized so that a pooled analysis is both valid and meaningful.

The workshop also demonstrated that the development of DataSHIELD was based on social practices as well as scientific and ethical motivations. Protecting a social norm, and communicating it to other stakeholders (e.g. policy makers and ethics committees), occurs partly through demonstrations of scientific validity and effective security safeguards and partly through relationships of trust in which stakeholders are persuaded of the scientific value and security of the technology. These participants needed to know how to 'persuade' their organization, funder and other researchers. Privacy here was about trust: trust in the technology and ways of commu-

nicating the trustworthiness of that technology. Arguably, privacy is always about trust. 'Reassuring' and 'convincing' are, however, fundamentally relational concerns not purely technical or scientific ones. But, though the solutions are inevitably social, it is undoubtedly a false dichotomy to construct (and limit) privacy, trust, and ethics to one sphere (the public, ethics committees, funders, etc.) and scientists and research to another. As we demonstrate here, the social and ethical realm was ever-present in the scientific one, rather than the two perspectives living in incommensurable worlds. The effective execution and development of DataSHIELD is dependent on the social and the technoscientific. Future development of DataSHIELD requires formal proof of concept testing, piloting in a variety of settings, active tests of the potential for malicious use and hacking, *and* formal exploration of the social dimensions of DataSHIELD to understand and develop acceptable approaches to implementing DataSHIELD. Development of DataSHIELD demonstrates the need to move beyond constructing the priorities of science and society as incommensurable. Scientific and ethical concerns are the concerns of both the scientists and the public.

Acknowledgements

The research program of the Data to Knowledge for Practice (DKP) Group at the University of Leicester is supported by the BioSHaRE-EU project (European Commission, FP7, #261433), Wellcome Trust Supplementary Grant #086160/Z/08/A, and joint MRC/Wellcome Trust Project Grant #G1001799/#WT095219MA. Development of DataSHIELD and Opal software at the Ontario Institute for Cancer Research is funded under the BioSHaRE-EU project (European Commission, FP7, #261433). DataSHIELD was originally conceived under an international research program jointly led by the P^{3}G Project (Public Population Project in Genomics – funded by Genome Canada and Genome Quebec) and PHOEBE (Promoting Harmonization of Epidemiological Biobanks in Europe – funded by European FP6, LSHG-CT-2006-518418). Vincent Ferretti is a recipient of Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the government of Ontario. The authors gratefully acknowledge Prof. Dr. Peter Dabrock, Prof. Dr. Herbert Gottweis and Dr. Andréa Vermeer for their support and organization of the PRIVATE Gen Workshop 'Privacy and Post-Genomics Medical Research: Challenges, Strategies, Solutions' supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF).

References

1 Murtagh MJ, Wallace SE, Kaye J, Demir I, Fortier I, Harris JR, Cox D, Laflamme P, Ferretti V, Sheehan NA, Hudson TJ, Cambon-Thomsen A, Knoppers BM, Brookes AJ, Burton PR: Navigating the perfect [data] storm. Norsk Epidemiol 2012;20:203–209.

2 Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P: Size matters: just how big is big?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 2009;38:263–273.

3 Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ: The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? Int J Epidemiol 2003;32:51–57.

4 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–9367.

5 Gottweis H, Lauss G: Biobank governance in the post-genomic age. Per Med 2010;7:187–195.

6 Hoeyer K: The ethics of research biobanking: a critical review of the literature. Biotechnol Genet Eng Rev 2008;25:429–452.

7 Taylor MJ, Townend D: Issues in protecting privacy in medical research using genetic information and biobanking: the PRIVILEGED project. Med Law Int 2010;10:253–268.

8 Wallace S, Lazor S, Knoppers BM: Consent and population genomics: the creation of generic tools. IRB 2009;31:15–20.

9 Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, Laflamme P, Tobin MD, Macleod J, Little J, Fortier I, Knoppers BM, Burton PR: DataSHIELD: resolving a conflict in contemporary bioscience – performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 2010;39:1372–1382.

10 Sutton AJ, Kendrick D, Coupland CA: Meta-analysis of individual- and aggregate-level data. Stat Med 2008;27:651–669.

11 Warren SD, Brandeis LD: The right to privacy. Harv Law Rev 1890;4.

12 Gormley K: One hundred years of privacy. Wis L Rev 1992;1:1335.

13 General Medical Council: Confidentiality. 2009. http://www.gmc-uk.org/guidance/ethical_guidance/confidentiality.asp.

14 von Egdell W, et al: All England Law Reports, Court of Appeal, Civil Division. 1989, 1, pp 835.

15 Herring J: Medical Law and Ethics, ed 3. Oxford, Oxford University Press, 2010.

16 Campbell v Mirror Group Newspaper Ltd.: UK House of Lords, House of Lords. 2004, 22, pp 995.

17 Laurie G: Reflexive governance in biobanking: on the value of policy led approaches and the need to recognise the limits of law. Hum Genet 2011;130:347–356.

18 Nowotny H, Scott P, Gibbons M: Introduction: Mode 2 revisited: the new production of knowledge. Minerva 2003;41:179–194.

19 Nowotny H, Scott P, Gibbons M: Re-thinking science: knowledge and the public in an age of uncertainty. Cambridge, Polity Press, 2001.

20 Gibbons M, Nowotny H: The potential of transdisciplinarity; in Thompson Klein J, Grossenbacher-Mansuy W, Häberli R, Bill A, Scholz RW, Welti M (eds): Transdisciplinarity: Joint Problem Solving among Science, Technology, and Society–An Effective Way for Managing Complexity, ed 1, Berlin, Birkhäuser Verlag, 2001, pp 67–80.

21 Thompson Klein J, Grossenbacher-Mansuy W, Häberli R, Bill A, Scholz RW, Welti M: Transdisciplinarity: Joint Problem Solving among Science, Technology, and Society – An Effective Way for Managing Complexity. Berlin, Birkhäuser Verlag, 2001.

22 Bataille V, Boniol M, De Vries E, Severi G, Brandberg Y, Sasieni P, Cuzick J, Eggermont A, Ringborg U, Grivegnée AR, Coebergh JW, Chignol MC, Doré JF, Autier P: A multicentre epidemiological study on sunbed use and cutaneous melanoma in europe. Eur J Cancer 2005;41:2141–2149.
23 Suchman LA: Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge, Cambridge University Press, 1987.
24 Szymanski MH, Whalen J: Making Work Visible: Ethnographically Grounded Case Studies of Work Practice. Cambridge, Cambridge University Press, 2011.
25 Pink S: Doing Visual Ethnography: Images, Media and Representation in Research. London, Sage Publications, 2007.
26 Heath C, Hindmarsh J, Luff P: Video in Qualitative Research: Analysing Social Interaction in Everyday Life. London, Sage Publications, 2010.
27 Dourish P, Anderson K: Collective information practice: exploring privacy and security as social and cultural phenomena. Hum-Comput Interact 2006;21:319–342.
28 Solove DJ: Conceptualizing privacy. Calif Law Rev 2002;90:1087–1155.
29 Cohen JE: Examined lives: informational privacy and the subject as object. Stanford Law Rev 1999;52:1373.
30 Goffman E: On face-work: an analysis of ritual elements in social interaction. Psychiatry 1955;18:213–231.
31 Goffman E: The Presentation of Self in Everyday Life. Garden City, Doubleday and Company, 1959.
32 Dalsgaard S: Facework on facebook: the presentation of self in virtual life and its role in the US elections. Anthropol Today 2008;24: 8–12.
33 Nafus D, Tracey K: Mobile phone consumption and concepts of personhood; in Katz J, Aakhus M (eds): Perpetual Contact: Mobile Communication, Private Talk, Public Performance. Cambridge, Cambridge University Press, 2002.
34 Siegel RB: 'The rule of love': wife beating as prerogative and privacy. Yale Law J 1996;105: 2117–2207.
35 MacKinnon CA: Toward a Feminist Theory of the State. Cambridge, Harvard University Press, 1991.
36 Butler J: Excitable Speech: A Politics of the Performative. New York, Routledge, 1997.
37 Callon M: What does it mean to say that economics is performative? CSI Working Papers Series, Nr. 005, 2006.
38 Murtagh MJ, Demir I, Harris JR, Burton PR: Realizing the promise of population biobanks: a new model for translation. Hum Genet 2011;130:333–345.
39 Gomatam S, Karr A, Reiter J, Sanil A: Data dissemination and disclosure limitation in world without microdata: a risk-utility framework for remote access analysis servers. Stat Sci 2005;20:163–177.
40 Knoppers BM, Hudson TJ: The art and science of biobanking. Hum Genet 2011;130: 329–332.
41 Callon M: Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay; in Law J (ed): Power, Action and Belief: A New Sociology of Knowledge. London, Routledge & Kegan Paul, 1986, pp 196–233.