International Journal of Epidemiology, 2014, 1929–1944
doi: 10.1093/ije/dyu188
Advance Access Publication Date: 26 September 2014
Original article

Data Matters

DataSHIELD: taking the analysis to the data, not the data to the analysis

Amadou Gaye,[1] Yannick Marcon,[2] Julia Isaeva,[3] Philippe LaFlamme,[2] Andrew Turner,[1] Elinor M Jones,[4] Joel Minion,[1] Andrew W Boyd,[1] Christopher J Newby,[5] Marja-Liisa Nuotio,[6,7] Rebecca Wilson,[1] Oliver Butters,[1] Barnaby Murtagh,[8] Ipek Demir,[9] Dany Doiron,[2] Lisette Giepmans,[10] Susan E Wallace,[8] Isabelle Budin-Ljøsne,[3] Carsten Oliver Schmidt,[11] Paolo Boffetta,[12] Mathieu Boniol,[12] Maria Bota,[12] Kim W Carter,[13] Nick deKlerk,[13] Chris Dibben,[14] Richard W Francis,[13] Tero Hiekkalinna,[6,7] Kristian Hveem,[15] Kirsti Kvaløy,[15] Sean Millar,[16] Ivan J Perry,[16] Annette Peters,[17] Catherine M Phillips,[16] Frank Popham,[18] Gillian Raab,[14] Eva Reischl,[17] Nuala Sheehan,[8] Melanie Waldenberger,[17] Markus Perola,[6,7,19] Edwin van den Heuvel,[20] John Macleod,[1] Bartha M Knoppers,[21] Ronald P Stolk,[10,22] Isabel Fortier,[2] Jennifer R Harris,[3] Bruce HR Woffenbuttel,[22,23] Madeleine J Murtagh,[24†] Vincent Ferretti[2,25†] and Paul R Burton[2,24†*]

[1]School of Social and Community Medicine, University of Bristol, Bristol, UK, [2]Maelstrom Research Group, Research Institute of the McGill University Health Centre, McGill University, Montreal, Canada, [3]Norwegian Institute of Public Health, Oslo, Norway, [4]Department Statistical Science, University College London, London, UK, [5]Department of Infection, Immunity and Inflammation, Health Sciences, University of Leicester, Leicester, UK, [6]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland, [7]Unit of Public Health Genomics, National Institute for Health and Welfare, Helsinki, Finland, [8]Department of Health Sciences, University of Leicester, Leicester, UK, [9]Department of Sociology, University of Leicester, Leicester, UK, [10]Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands, [11]Institut für Community Medicine, University Medicine of Greifswald, Greifswald, Germany, [12]International Prevention Research Institute, Lyon, France, [13]Telethon Kids Institute, University of Western Australia, Perth, WA, Australia, [14]School of Geosciences, University of Edinburgh, Edinburgh, UK, [15]Norwegian University of Science and Technology, Levanger, Norway, [16]HRB Centre for Diet and Health Research, Department of Epidemiology and Public Health, University College Cork, Cork, Ireland, [17]Research Unit of Molecular Epidemiology, Research Center for Environmental Health, Neuherberg, Germany, [18]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Glasgow, UK, [19]University of Tartu, Estonian Genome Center, Tartu, Estonia, [20]University Medical Center Groningen, Medical Statistics, Groningen, The Netherlands, [21]Centre of Genomics and Policy, McGill University, Montreal, Canada, [22]University Medical Center Groningen, LifeLines Cohort Study, Groningen, The Netherlands, [23]Department of Endocrinology, University Medical Center Groningen, Groningen, The Netherlands, [24]School of Social

and Community Medicine, University of Bristol, Bristol, UK and [25]Ontario Institute for Cancer Research, Toronto, Canada

*Corresponding author. E-mail: paul.burton@bristol.ac.uk

†These authors contributed equally to this work.

Accepted 15 August 2014

Abstract

Background: Research in modern biomedicine and social science requires sample sizes so large that they can often only be achieved through a pooled co-analysis of data from several studies. But the pooling of information from individuals in a central database that may be queried by researchers raises important ethico-legal questions and can be controversial. In the UK this has been highlighted by recent debate and controversy relating to the UK's proposed '*care.data*' initiative, and these issues reflect important societal and professional concerns about privacy, confidentiality and intellectual property. DataSHIELD provides a novel technological solution that can circumvent some of the most basic challenges in facilitating the access of researchers and other healthcare professionals to individual-level data.

Methods: Commands are sent from a central analysis computer (AC) to several data computers (DCs) storing the data to be co-analysed. The data sets are analysed simultaneously but in parallel. The separate parallelized analyses are linked by non-disclosive summary statistics and commands transmitted back and forth between the DCs and the AC. This paper describes the technical implementation of DataSHIELD using a modified R statistical environment linked to an Opal database deployed behind the computer firewall of each DC. Analysis is controlled through a standard R environment at the AC.

Results: Based on this Opal/R implementation, DataSHIELD is currently used by the Healthy Obese Project and the Environmental Core Project (BioSHaRE-EU) for the federated analysis of 10 data sets across eight European countries, and this illustrates the opportunities and challenges presented by the DataSHIELD approach.

Conclusions: DataSHIELD facilitates important research in settings where: (i) a co-analysis of individual-level data from several studies is scientifically necessary but governance restrictions prohibit the release or sharing of some of the required data, and/or render data access unacceptably slow; (ii) a research group (e.g. in a developing nation) is particularly vulnerable to loss of intellectual property—the researchers want to fully share the information held in their data with national and international collaborators, but do not wish to hand over the physical data themselves; and (iii) a data set is to be included in an individual-level co-analysis but the physical size of the data precludes direct transfer to a new site for analysis.

Key words: DataSHIELD, pooled analysis, ELSI, privacy, confidentiality, disclosure, distributed computing, intellectual property, bioinformatics

Key Messages

- DataSHIELD provides a solution when ethico-legal considerations prevent or impede data-sharing and analysis.
- It promotes and facilitates collaborations by empowering data owners and affording them better control over their data.
- DataSHIELD has the potential to protect the intellectual property of researchers in institutions and countries with limited resources, thus enabling more balanced collaborations with wealthier partners.
- It also improves the governance and management of data by allowing them to be maintained locally.

Introduction

The analysis of complex interrelated datasets containing demographic, social, health-related and/or biological information derived from large numbers of individuals has become pivotal to the investigation of disease causation and to the evaluation of healthcare programmes and interventions. However, the daunting sample sizes needed to provide adequate statistical power[1–3] often exceed the provision of any one single study. Furthermore, if major research funders are to optimize return on their investment of public or charitable money, it is crucial that researchers other than those who originally created a particular data set are able to access and work with those data.[4] These two imperatives underpin the active encouragement of 'data sharing'—across several studies, or from a single data source—which is central to contemporary bioscience.[5] The data to be shared may be derived from large epidemiological studies, from smaller research projects and/or from healthcare or administrative records. They may originally have been intended for research or for direct support of patient care or public health. There is no doubt that liberating and integrating such information to support medical research has the potential to generate enormous future health benefits. But substantive challenges exist, and the sharing of data—particularly individual-level data, also known as *microdata*[6]—raises important societal and professional concerns.

In the UK, these concerns were recently highlighted by controversy surrounding the *care.data* project.[7,8] At a societal level they include real and perceived frailties of information governance when a research database containing potentially sensitive personal information about individuals is made accessible to any third party including researchers.[4] However, these broader societal concerns are closely—though not precisely—mirrored in the disquiet of some professional health researchers regarding the unfettered sharing of valuable scientific data that they believe exist primarily because they have made a substantial investment of their own time, effort and scientific thought to creating and managing them. In both instances, individuals for whom the data to be shared are valuable and potentially sensitive (personally, or as intellectual property) worry that, once they have been physically 'shared', there will be a significant loss of control over their subsequent exploitation. In support of this thesis, we have noted[9] that researchers are often more than willing to share the information contained in their data—because this enhances the quality and quantity of their own scientific output by providing opportunities for national and international collaboration. But they are sometimes less keen to hand over the physical data themselves,[9] because even with ethically and legally binding safeguards in place, the loss of governance control over the data themselves and the intellectual property they represent can be seen as seriously problematic. This is particularly so for data creators with limited resources for managing and scientifically exploiting their own data—e.g. researchers in developing countries. Effective and acceptable solutions must be found to all of these problems if we are to optimize evidence-based progress in stratified and conventional medicine.

Many technical and policy measures can be enacted to render data sharing more secure from a governance perspective and less likely to result in loss of intellectual property. For example, data owners might restrict data release to aggregate statistics alone, or may limit the number of variables that individual researchers might access for specified purposes. Alternatively, secure analysis centres, such as the ESRC Secure Data Service,[10] and SAIL,[11] represent major informatics infrastructures that can provide a safe haven for remote or local analysis/linkage of data from selected sources while preventing researchers from downloading the original data themselves. However, to complement pre-existing solutions to the important challenges now faced, the DataSHIELD consortium has developed a flexible new way to comprehensively analyse individual-level data collected across several studies or sources while keeping the original data strictly secure. As a technology, DataSHIELD uses distributed computing and parallelized analysis to enable full joint analysis of individual-level data from several sources—e.g. research projects or health or administrative data—without the need for those data to move, or even be seen, outside the study where they usually reside.[12] Crucially, because it does not require underpinning by a major informatics infrastructure and because it is based on non-commercial open source software, it is both locally implementable and very cost effective.

Co-analysis of data from several studies/sources is often conducted using study-level meta-analysis (SLMA),[13–15] using conventional meta-analysis to combine results generated by each study separately.[16,17] In contrast, individual-level meta-analysis (ILMA)involves the physical transfer of data from each study to produce a single central database that is then analysed as if it were a conventional multicentre data set.[16,17] Unfortunately, both SLMA and ILMA present significant problems.[12,16,17] Because SLMA combines analytical results (e.g. means, odds ratios, regression coefficients) produced ahead of time by the contributing studies, it can be very inflexible: only the pre-planned analyses undertaken by all the studies can be converted into joint results across all studies combined. Any additional analyses must be requested *post hoc*. This hinders exploratory analysis,[16] for example the investigation of sub-groups, or interactions between key variables. In contrast, ILMA is very flexible, but ethico-legal considerations can impede access to individual-level data. Thus, research may be delayed if formal data access procedures are protracted, or may have to be postponed while participants have reconsented.[18,19] ILMA may even be impossible if consent forms prohibit individual-level data being sent to external researchers, or if privacy legislation precludes sharing of data across national or jurisdictional boundaries.[12,20,21]

DataSHIELD circumvents these problems. First, it can be set up to be mathematically equivalent to ILMA,[12,22,23] while avoiding the attendant governance, legal or societal concerns.[21,24] Individual-level data never cross, and are never visible outside, the firewall of their home study.[12,20,24] Jones *et al.*[22] explain why fitting a generalized linear model (GLM) under DataSHIELD produces exactly the same results—not just a good approximation—as a GLM fitted to a single database containing the individual-level data from all studies combined. This is confirmed empirically in the current article by the comparison of the output of a GLM model fitted initially via DataSHIELD on all studies separately, and then through R on the pooled data (i.e. the separate data sets stacked together into one table). Second, however, DataSHIELD can also be configured to mimic a secure SLMA but without the need to ask individual studies to undertake their own analyses. Under DataSHIELD, any non-disclosive analysis may therefore be requested at any time without physically sharing data. DataSHIELD can also protect intellectual property when data producers are keen for external researchers to query and work with their data but do not wish to lose ultimate control by physically transferring their data. This can even apply to a single study—single-site DataSHIELD—which may be viewed as being a particularly simple and cost-effective way to construct a 'secure data enclave' within which data can be comprehensively analysed but not accessed. For all of these reasons, DataSHIELD encourages 'true', equal-status collaboration.

Figure 1 illustrates the basic IT infrastructure for a hypothetical DataSHIELD implementation for co-analysing six studies. The individual-level data themselves remain on 'data computers' (DCs) at their home bases. A central 'analysis computer' (AC) is used to issue commands to enact and control the analysis. As a by-product of its underlying structure, DataSHIELD can enhance governance and data management because data are locally maintained by their producers who typically know them best; that is, it encourages storage, updating and sharing of complex multi-class data from ongoing studies through a federated rather than a centralized architecture. However, this does not deny the important complementary role of large centralized repositories specializing in archiving particular classes of data, such as the European Genome-Phenome Archive[25] or the UK Data Service.[10] As an additional consequence of its structure, DataSHIELD can also avoid the need to move very large data sets. Finally, because all data remain unobserved at their home repository, DataSHIELD can mitigate some of the dilemmas arising from findings of actionable clinical significance in individuals.[26] Specifically, external researchers cannot, in principle, produce results pertaining to individual participants. Rather, individual clinical results can only be generated by investigators working with data from their own study and these investigators should be covered by formal internal policies.

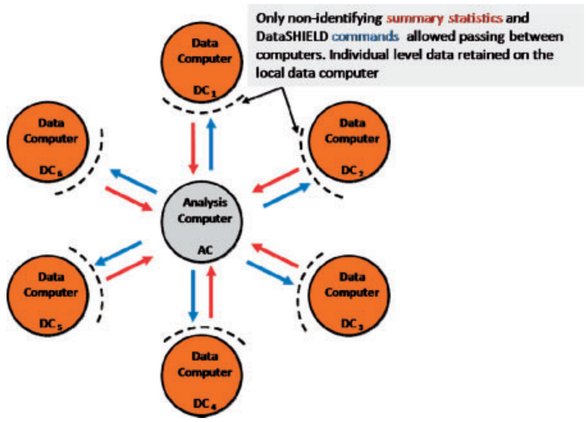
Only non-identifying summary statistics and DataSHIELD commands allowed passing between computers. Individual level data retained on the local data computer
Data Computer DC1
Data Computer DC2
Data Computer DC3
Data Computer DC4
Data Computer DC5
Data Computer DC6
Analysis Computer AC

Figure 1. Typical DataSHIELD setting for a pooled individual-level analysis.

DataSHIELD offers both opportunities and challenges. It has been known for several years that it works in principle,[12,22] but its practical implementation and utilization on an IT platform that can be used by non-expert researchers has proved to be challenging. This paper describes the application platform that has now been developed. It explains each of the fundamental steps in a typical DataSHIELD analysis and outlines the key elements of the infrastructure that underpins these steps. Illustration is based on a real-world setting in which DataSHIELD is currently being used to analyse data for a pan-European consortium: the Healthy Obese Project (HOP).[27] Finally, we briefly discuss a potential future role of DataSHIELD in circumventing some of the privacy and confidentiality concerns arising—as under *care.data*[7,8]—when progress in biomedical science might be accelerated if researchers could easily access and co-analyse data held in multiple sources, including healthcare, social or governmental data, that may have been administratively generated.

Methods

The IT infrastructure

The IT infrastructure required to carry out a DataSHIELD analysis comprises three main components: a computer server at each source study hosting an Opal database;[28]

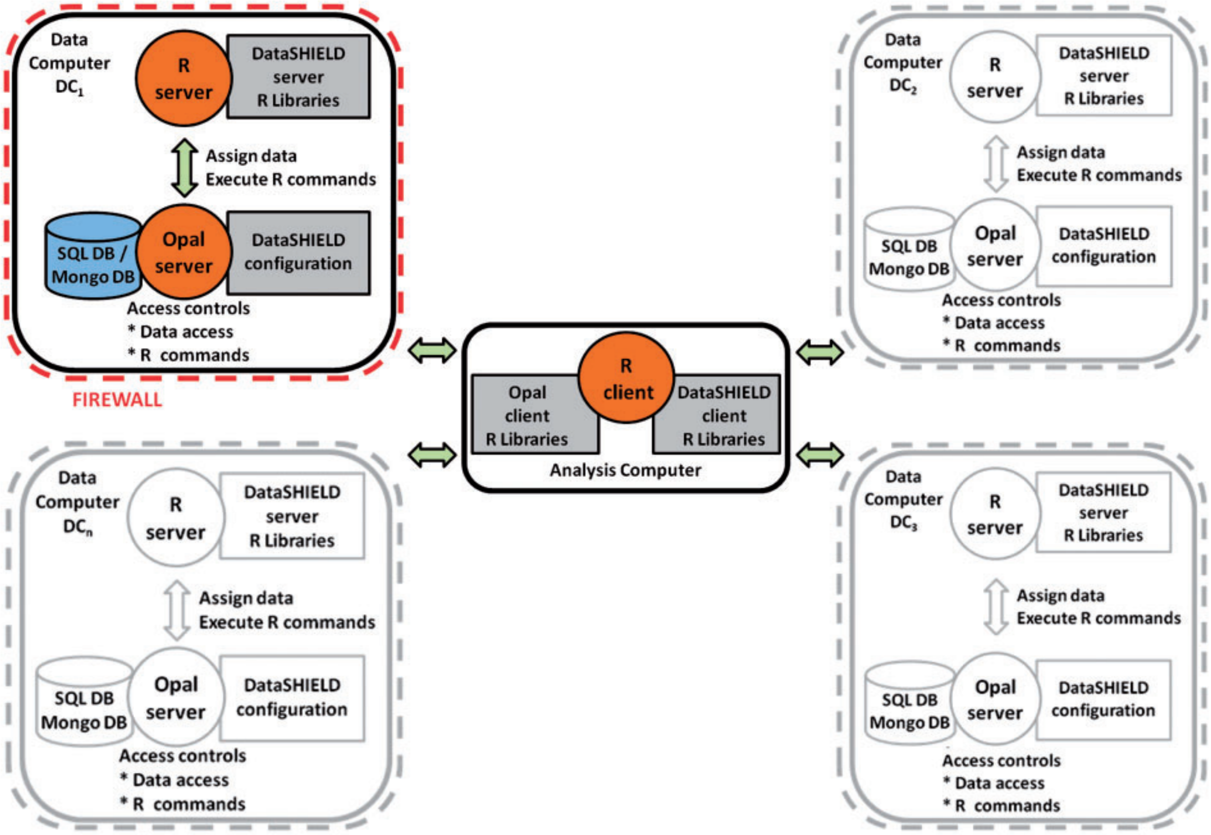

Figure 2. Overview of the IT infrastructure required for a DataSHIELD process. The settings are the same in all DCs so only one is highlighted in this figure.

the statistical programming environment(R[29]; and DataSHIELD-specific R libraries installed on the data servers (data computers = DCs) and on the client computer (analysis computer = AC). Opal is a core database application for biobanks and epidemiological studies developed by the Maelstrom Research group[30] in collaboration with OBiBa, an international software development project creating open-source software for Biobanks.[31] Opal, R and DataSHIELD are open source and freely available.

Instances of Opal, the R server and the DataSHIELD server-side R libraries are implemented behind the firewall of each data owner's DC (Figure 2). The AC is used to enact and control the distributed analysis. The DataSHIELD client-side R libraries are installed on the AC (Figure 2). A DataSHIELD platform consists of at least one AC communicating with a number of DCs or with just one DC (i.e. single-site DataSHIELD).

DataSHIELD process explained

DataSHIELD as described in this article is intended for the pooled analysis of 'horizontally partitioned' data, i.e. contributing sources hold the same variables but on different individuals (see Figure 3b). A new version of DataSHIELD is currently being developed for 'vertically partitioned' data where various sources hold different variables on the same individuals (see Figure 3c). This uses an overlapping range of secure approaches to secure data integration and retains the same fundamental principle: leave the data where they are but analyse them as if they were combined in one database.

As for any co-analysis, shared data must be harmonized first. The harmonization phase of the HOP project[32] within BioSHaRE-EU[33] (described in detail elsewhere[27,34,35]) is functionally independent of DataSHIELD itself (Table 1, step 0).

DataSHIELD functions

The fundamental building blocks of DataSHIELD are its client-side and server-side functions. As illustrated in Figure 2, server-side functions reside in the modified R environments located behind the firewall of the DC at each individual study. It is the server-side functions that actually process the individual-level data at the distinct repositories. The outputs from server-side functions (non-disclosive study-level statistics) represent the only information that ever leaves a DC, and this is why we can claim that DataSHIELD allows full analysis of individual-level data without those data ever having to be moved, or even

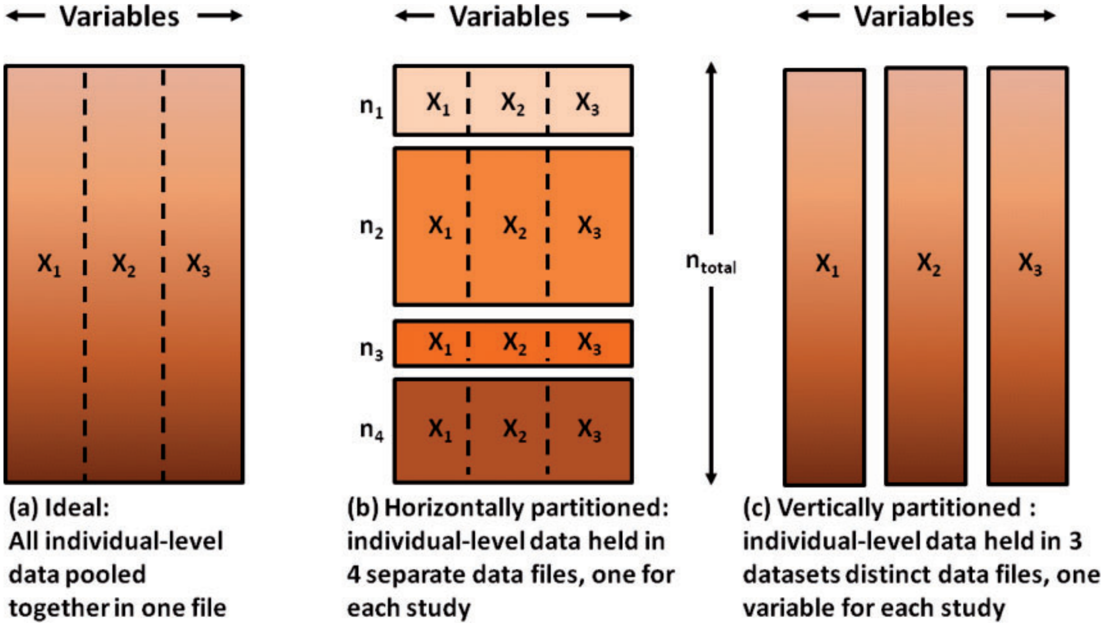
← Variables →
X_1
X_2
X_3
(a) Ideal:
All individual-level data pooled together in one file
← Variables →
n_1
n_2
n_3
n_4
X_1
X_2
X_3
(b) Horizontally partitioned:
individual-level data held in 4 separate data files, one for each study
n_{total}
← Variables →
X_1
X_2
X_3
(c) Vertically partitioned :
individual-level data held in 3 datasets distinct data files, one variable for each study

Figure 3. Graphical view of pooled data (a), horizontally partitioned (b) and vertically partitioned data (c).

rendered visible, outside their study of origin. Client-side functions reside on the conventional R environment on the AC. Client-side functions call and control server-side functions and combine information across different repositories when required. All DataSHIELD functions require approval under a technical and governance process including external independent evaluation.

DataSHIELD secure analyses

An iterative analysis (e.g. fitting a generalized linear model [GLM]) is illustrated in Figure 4: its steps are detailed in Table 1.The same process is triggered simultaneously in all four DCs. The process iterates through steps 5–8 until the combined coefficient estimates remain unchanged between two iterations (according to a pre-defined tolerance criterion). Once convergence is achieved the AC uses the final score vectors and information matrices from all data sets to provide definitive estimates of regression coefficients, their standard errors and other non-disclosive model outputs. One-step analyses are analogous to iterative analyses but do not require repeated loops. For example, to construct a contingency table, each study generates its own table in one step—this is inherentlynon-disclosive—and the AC integrates these to produce a combined table.

Disclosure control—examples

Some functions that are not intrinsically disclosive can nevertheless be problematic in certain settings. Thus, a contingency table with 1–4 observations in any one cell is often viewed as providing a potential disclosure risk.[21] To address this problem under DataSHIELD, each DC tests any contingency table it creates and will only return a full table to the AC if all cells are empty or contain at least five observations. All the AC knows is that it has received an incomplete table which is so constructed that nothing disclosive can be inferred—Sub-setting—e.g. by sex, age or phenotypic sub-type—is crucial in statistical analysis. But repeated sub-setting may produce sub-groups that are so small that results based on that subset (e.g. a mean) might potentially be disclosive. Under DataSHIELD, therefore, it is not possible to generate a subset data set containing 1–4 observations. However, this rule may be relaxed or made more stringent at the request of the principal investigator who is seen as taking responsibility for the overall analysis. The DataSHIELD project is currently working on governance rules for sub-setting.

Results: DataSHIELD at work

Analyses of data from the Health Obese Project

The Healthy Obese Project (HOP)[27,32] is part of the BioSHaRE-EU project.[33] It aims to identify individuals who are ‘healthy obese’ (HO), defined as having a body mass index >30 in the absence of any of the common metabolic sequelae of obesity—e.g. hypertension, hypercholesterolaemia, impaired glucose tolerance or diabetes—in order to study the biological and environmental correlates of HO. Since HO is relatively uncommon, any single study containing all requisite measures is likely to have inadequate statistical power. DataSHIELD provides an effective way to enact a secure federated co-analysis of the multiple studies involved in HOP. This section briefly describes how

Table 1. Detailed explanations of the steps in DataSHIELD process

Step	Explanation	Input data	Output data	Output location	Visibility
(0) Preliminary and prerequisite step	Strictly speaking, this step is not part of a DataSHIELD analysis process; it is, however, a prerequisite for any valid analysis that pools multiple data sets. Each contributing study (e.g. $STUDY_1$) identifies the requisite variables and creates any new harmonized variables needed in the combined analysis. These harmonized variables are then transferred from the study's database ($Original.DB_{STUDY1}$) to adatabase linked to the Opal server ($Analysis.DB_{STUDY1}$)	All variables held in $Original.DB_{STUDY1}$	Variables required for combined analysis. No data that may potentially be directly identifying [e.g. a full UK postcode, a full date of birth, or an ID that is equivalent to an ID available elsewhere (e.g. a national health system number)] unless such a variable is essential to the required analysis	A new analysis SQL database linked to an Opal server: $Analysis.DB_{STUDY1}$, located on a server controlled by the same researchers who run the original study	Invisible outside $STUDY_1$
(1) Login to collaborating servers	The user logs into the collaborating servers through secured web services, using the credentials provided to them. This authentication ensures that only users authorized by the access body (e.g. an analysis access panel put in place by the consortium) can actually carry out an analysis	A command with a specific public/private key pair	No data are returned, the connection is established	Not applicable	No individual-level server-side data are ever visible to the user after login
(2 and 3) Request and transfer of the shared data to the analysis zone	(2) The user sends a command to request the specific data to analyse. This could be all the variables or specific variables stored in $Analysis.DB_{STUDY1}$. (3) DataSHIELD extracts data from $Analysis.DB_{STUDY1}$ and transfers it ($Assigned.Data_{STUDY1}$) to the local R instance ($R.Environment_{STUDY1}$) of $STUDY_1$ controlled bythe researchers of $STUDY_1$	All or some of the variables in $Analysis.DB_{STUDY1}$	A data frame (an R data structure) with all variables or part of the variables in $Analysis.DB_{STUDY1}$	$R.Environment_{STUDY1}$ behind the firewall controlled by the same scientists who run $STUDY_1$	Individual-level data invisible outside $STUDY_1$. Aggregated data invisible outside $STUDY_1$ except via approved DataSHIELD commands[24]
(4) Starting the analysis (i.e. sending command to fit a GLM model)	The researcher sitting at the analysis computer (AC) sends an R command to every study telling it to fit one iteration of a generalized linear modelling fitting procedure (the iterative reweighted least-squares algorithm), including first-'guessed' estimates at what the ultimate set of regression coefficients will be	A short set of instructions completely unrelated to any data in any study which contains the model to fit and an arbitrary string of numbers representing the first-guessed coefficient estimates	Instructions about the model to fit and the coefficient estimates are received by $R.Environment_{STUDY1}$	$R.Environment_{STUDY1}$	The model to fit and the coefficient estimates are visible outside $STUDY_1$ but are non-sensitive

(Continued)

Table 1. Continued

Step	Explanation	Input data	Output data	Output location	Visibility
(5) Carrying out the analysis locally (i.e.enact one iteration of a GLM fit)	Each data computer responds to the instructions sent from the AC in step 4 by running a single iteration of aGLM fit. This fitting is carried out in $\mathit{R.Environment}_{STUDY1}$ using the first coefficient estimates as starting position. Two mathematical products of this analysis are called the score vector and the information matrix	Instructions as in step 4	A score vector (e.g. $\mathit{Score.Vector}_{STUDY1}$) and an information matrix (e.g. $\mathit{Information.Matrix}_{STUDY1}$) are calculated by each study	$\mathit{R.Environment}_{STUDY1}$	$\mathit{Score.Vector}_{STUDY1}$ and $\mathit{Information.Matrix}_{STUDY1}$ carry no individually identifying or sensitive data and are only visible outside study_1via a legal DataSHIELD Command
(6) Summary statistics returned to the analysis computer	DC_1 transmits $\mathit{Score.Vector}_{STUDY1}$ and $\mathit{Information.Matrix}_{STUDY1}$ to the analysis computer	$\mathit{Score.Vector}_{STUDY1}$ and $\mathit{Information.Matrix}_{STUDY1}$ are sent	$\mathit{Score.Vector}_{STUDY1}$ and $\mathit{Information.Matrix}_{STUDY1}$ are received	Analysis computer	$\mathit{Score.Vector}_{STUDY1}$ and $\mathit{Information.Matrix}_{STUDY1}$ are now visible to outside world but they carry no individually identifying or sensitive information
(7) Combining the summary statistics returned by the DCs	The analysis computer adds up the score vectors and information matrices from all DCs, divides the first sum by the latter (technically, a matrix multiplication) and uses the result to update the coefficient estimates using the conventional updating algorithm called the Iterative Reweighted Least Squares (IRLS) algorithm[36]	Score vectors and information matrices from all DCs	New coefficient estimates	Analysis computer	All visible to outside world, but carry no sensitive information
(8) Repeat step 4 with updated coefficients	The same process as in step 4 re-starts; the analysis computer commands the DCs to fit the same model with the updated coefficient estimates	Same as per step (4)	Same as per step (4)	Same as per step (4)	Same as per step (4)

Keep repeating steps 5-8 until the model is almost unchanged (judged by appropriate convergence criterion) between iterations—the model is then said to have converged

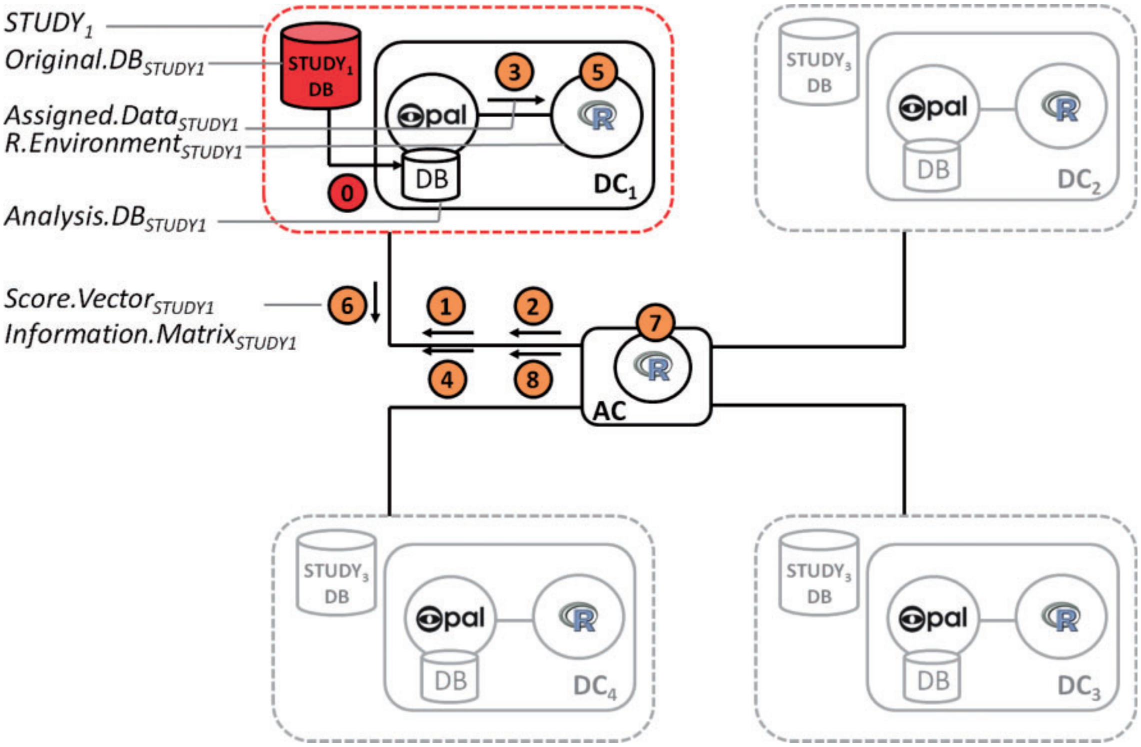

Figure 4. Overview of a DataSHIELD process. Each of the 8 steps and the terms used to refer to the key components and data exchanged between AC and DCs are detailed in Table 1.

DataSHIELD was implemented for this application. At the time of writing, HOP involves 10 studies across eight European countries (Table 2) sharing 96 harmonized variables.

Figure 5 schematizes the DataSHIELD analysis of HOP data. Under HOP, communications between the AC and the DCs pass through the BioSHaRE-EU[33] MICA[37] web portal. This ensures that links can only be made through a designated IP name. Such a portal is not a pre-requisite for a DataSHIELD analysis, but it further enhances security. In the Hop settings the Analysis Computer is just used to login to the HOP portal where the client functions are installed and from where the actual analysis is ran.

Examples of DataSHIELD commands

Although the examples in this section are real as they use real data from the HOP project, they are included here for illustrative purposes only. For the sake of conciseness—and to maintain consistency across all examples—we include only four of the available studies in these examples. Throughout this section the DataSHIELD commands, in bold and italic font, are preceded by explanations and followed, where there is any, by the output of the command, in italic font.

Histogram plots

Figure 6 illustrates the output from a DataSHIELD histogram plot of HDL cholesterol for each of the four studies (Figure 6A) and for the pooled data (Figure 6B). The DataSHIELD function ***ds.histogram*** filters the information returned from each study to remove bars based on a count of between 1 and 4. This means that potentially disclosive outliers are not shown on the plot. It however reports the number of invalid cells in the original grid density matrix used to produce the graph. For all DataSHIELD commands, the 'type' argument indicates whether to report results for each study separately (**type='split'**) or across all studies, the default behaviour.

```
ds.histogram('D$LAB_HDL',type = 'split')
ds.histogram('D$LAB_HDL')
ncds: Number of invalid cells (cells with
  counts > 0 and < 5) is 53
finrisk: Number of invalid cells (cells with
  counts > 0 and < 5) is 72
micros: Number of invalid cells (cells with
  counts > 0 and < 5) is 55
kora: Number of invalid cells (cells with
  counts > 0 and < 5) is 75.
```

Quantiles

The DataSHIELD function **ds.quantileMean** returns means and critical quantiles for quantitative variables. Unlike the conventional summary function in R, the

Table 2. Healthy Obese Project collaborating studies and shared number of participants at the time of this work

Study name	Host institution	Location	Participants
Cooperative Health Research in South Tyrol Study (CHRIS)	European Academy of Bolzano	Bolzano, Italy	1583
Cooperative Health Research in the Region of Augsburg (KORA)	Helmoltz Center Munich	Augsburg, Germany	3080
LifeLines Cohort Study (LifeLines)	University Medical Center Groningen	Groningen, The Netherlands	94516
Mitchelstown Study Population (Mitchelstown)	Living Health Clinic in Mitchelstown	Cork, Ireland	2047
Microisolates in South Tyrol Study (MICROS)	European Academy of Bolzano	Bolzano, Italy	1060
National Child Development Study (NCDS)	University of Leicester	Leicester, UK	7210
FINRISK 2007 Study (FINRISK 2007)	National Institute for Health and Welfare	Helsinki, Finland	5024
Nord-Trøndelag Health Study (HUNT)	Norwegian University of Science and Technology	Levanger, Norway	78968
Prevention of REnal and Vascular ENd-stage Disease study (PREVEND)	University Medical Center Groningen	Groningen, The Netherlands	8592
The Study of Health in Pomerania (SHIP) joined HOP after the analysis reported in this paper, and so the text and figures refer to 9 not 10 studies	University Medicine of Greifswald	Greifswald, Germany	4308

DataSHIELD function does not return the minimum and maximum values because these may be disclosive. The results below were obtained by running the command on the quantitative age variable encoding age in years:

```
ds.quantileMean('D$AGE_YRS')
Quantiles of the pooled data
   5%      10%      25%      50%      75%      90%      95%     Mean
36.46855 38.14755 43.30267 49.82448 56.23977 59.86747 61.36912 49.46372
```

One and two-dimensional contingency tables

One-dimensional tables. The output below is generated by the DataSHIELD function **ds.table1D**, applied to a categorical variable holding BMI in three classes—for all studies combined. In addition to the counts in each category, the function also reports column percentages, row percentages and global percentages. To save space, only counts are shown here. The function also reports on the 'validity' of each study data set (full results being reported only for studies where the table is entirely non-disclosive, i.e. no table cells have counts between 1 and 4). As the last component of the output—$VALIDITY.WARNING—each source is flagged as having only valid data, or at least some invalid data.

```
ds.table1D('D$PM_BMI_CATEGORIAL')
$'TOTAL.VALID.DATA.COUNTS   for   variable
 PM_BMI_CATEGORIAL'
ncdsfinrisk micros kora TOTAL
1       2453  1777   539   972   5741
2       2905  2096   364  1279   6644
3       1733  1151   157   812   3853
TOTAL   7091  5024  1060  3063  16238

$VALIDITY.WARNING
[1] 'ALL STUDIES VALID'
```

Two-dimensional tables. The function **ds.table2D** generates two-dimensional contingency tables. Here, the categorical BMI variable is tabulated against gender. The function **ds.table2D** also produces column percentages, row percentages, global percentages and validity information. It also runs chi-square tests for homogeneity on (nc-1)*(nr-1) degrees of freedom for each study and for all studies combined, where nc is the number of columns and nr the number of rows.

```
ds.table2D('D$PM_BMI_CATEGORIAL',
 'D$GENDER')
$'COMBINED.VALID.DATA.COUNTS-
  PM_BMI_CATEGORIAL (rows) V GENDER (cols) '
          0      1    TOTAL
1       2036  3705   5741
```

```
2       3826  2818   6644
3       1807  2046   3853
TOTAL  7669  8569  16238

$CHI2.TESTS.FOR.HOMOGENEITY
             X2-value   df     p-value
ncds         350.12295  2   9.370602e-77
finrisk      139.05465  2   6.377738e-31
micros        34.21016  2   3.726980e-08
kora          98.49705  2   4.089196e-22
ALL VALID
 STUDIES
 COMBINED  604.93484  2  4.365851e-132
```

Generalized linear models (GLMs)

Because we wanted to directly compare the results of a GLM analysis under DataSHIELD with the corresponding results obtained from a conventional R-based GLM analysis—i.e. with the raw data from all sources physically combined in one database (Table 3)—our GLM example is based on four of the HOP studies that explicitly allowed their data to be physically shared within the HOP consortium, as well as to be analysed via DataSHIELD: NCDS,[38] KORA,[39] LifeLines[40] and Mitchelstown.[41]

The DataSHIELD GLM function, ***ds.glm***, is currently constructed to fit linear regression (Gaussian family, identity link), logistic regression (binomial family, logistic link) and Poisson regression (Poisson family, log link). It can easily be extended to encompass other combinations of errors and links. Because it is based around the conventional *glm* function in R, it can fit categorical factors as well as quantitative covariates, and can make use of the full array of R model-fitting operators in specifying the formula—e.g. * meaning all possible interactions between a categorical covariate and another covariate, or **-1** meaning remove the regression constant. Intermediate summaries of the fitting process can be printed out after each iteration but, for the sake of conciseness, they are not reported here; only the final results

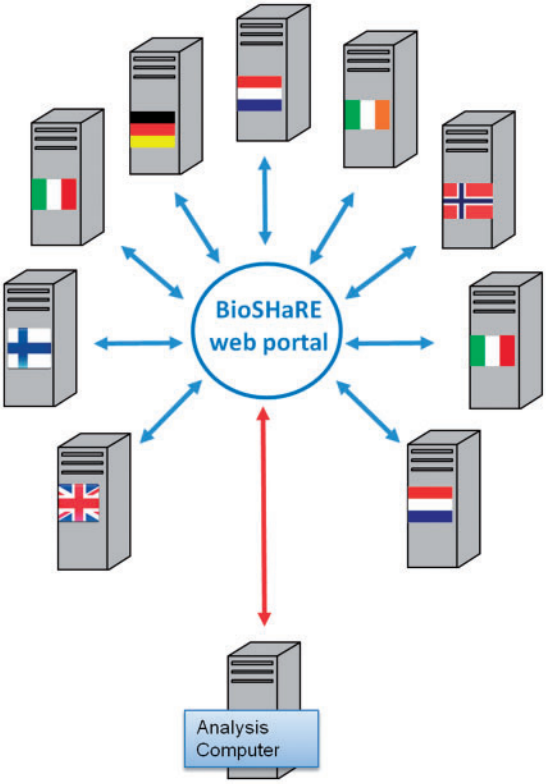

Figure 5. For the Healthy Obese Project, communications between AC and DCs were channelled through a trusted portal.

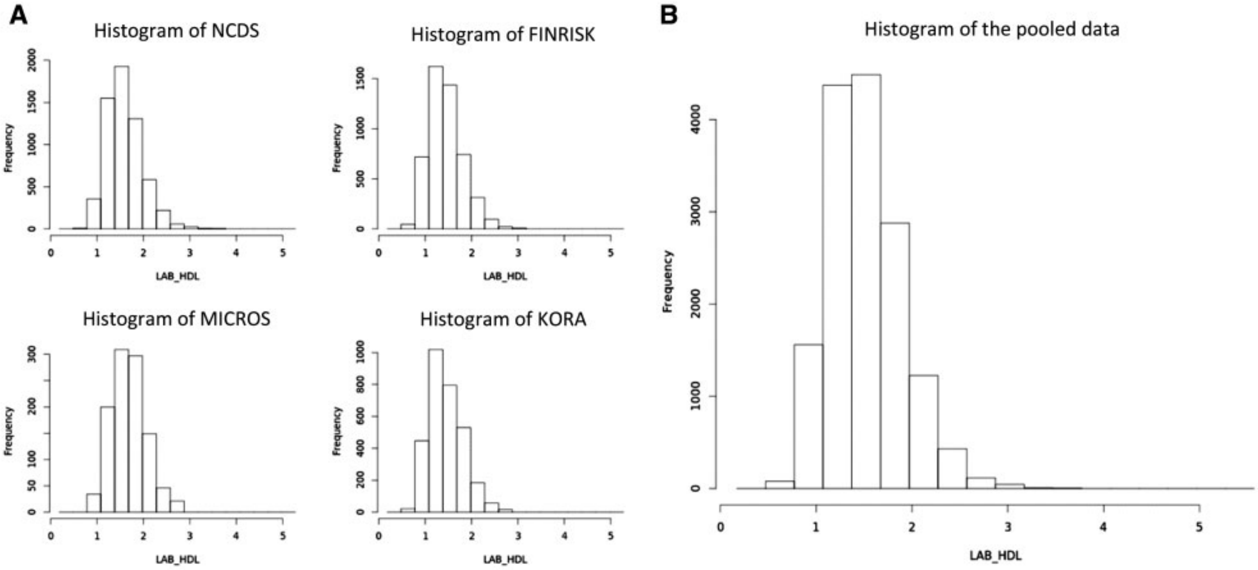

Figure 6. Histogram plots of the variable 'LAB_HDL' for each study (A) and for the pooled data (B).

Table 3. Comparison of the critical outputs of the same GLM model fitted using DataSHIELD (in light shading) and using standard R with the physically pooled data (in dark shading)

Parameter	Estimate	Estimate	SE	SE	*P*-value	*P*-value
(Intercept)	–4.17696	–4.17696	0.11274	0.11274	1.758e-300	<2e-16
Study.id.2	–0.11851	–0.11851	0.11692	0.11692	0.31078	0.31078
Study.id.3	–0.30955	–0.30955	0.08571	0.08571	3.041e-04	3.041e-04
Study.id.4	–0.40626	–0.40626	0.13438	0.13438	2.502e-03	2.502e-03
Age.50	0.06905	0.06905	0.00185	0.00185	1.033e-303	<2e-16
D$gender1	–0.45085	–0.45085	0.10936	0.10936	3.742E-05	3.74E-05
D$pm_bmi_categorial2	0.65056	0.65056	0.09179	0.09179	1.37E-12	1.37E-12
D$pm_bmi_categorial3	1.76134	1.76134	0.09343	0.09343	2.868E-79	<2e-16
D$gender1:d$pm_bmi_categorial2	0.15644	0.15644	0.12917	0.12917	0.22584	0.22584
D$gender1:d$pm_bmi_categorial3	0.24347	0.24347	0.12660	0.12660	0.05446	0.05446

DataSHIELD derived estimates rounded to same decimal places as standard R estimates. To avoid confusion, it should be noted that at a very early stage of the HOP analysis, the name of the categorical BMI variable was misspelt as '…CATEGORIAL…'. As that misspelling is now entrenched in all of the harmonized data sets etc. we chose not correct it for this paper.

SE, standard error.

(i.e. after the model has converged) are included in the output below. In order to enhance its illustrative value, the particular model we have fitted contains study-specific terms allowing for heterogeneity in the baseline risk of disease and have used the * operator to specify an interaction between GENDER and a three-level factor encoding BMI. In addition, we compare the estimates and confidence intervals from the GLM fitted using DataSHIELD with their equivalents from the same GLM fitted directly to a combined database into which the individual-level data from each study have been physically pooled. This provides empirical confirmation of the precise theoretical equivalence of the two approaches.[22] It should be noted that when variables are initially transferred from Opal into the DataSHIELD R environment at each source, they are by default placed in a data frame denoted 'D'. For the purposes of clarity here, these variables all have names that are capitalized, and the prefix 'D$' tells R to read them from the data frame. In contrast, all new variables created by transformation during the DataSHIELD session itself have been given lower-case names—these sit outside 'D'(at root level in the DataSHIELD R environment) and are not preceded by 'D$'.

```
glm.mod <- ds.glm(
formula=D$DIS_DIAB~study.id.2+study.id.3+
study.id.4+age.50+D$GENDER*D$PM_BMI_CATEGORIAL, family='binomial',
maxit=40)

$formula
D$DIS_DIAB ~ 1 + study.id.2 + study.id.3 + study.id.4 + age.50 +
    D$GENDER * D$PM_BMI_CATEGORIAL

$coefficients
                                Estimate   Std. Error    z-value     p-value
(Intercept)                    -4.17696342 0.112736796 -37.050578 1.757827e-300
study.id.2                     -0.11851073 0.116922767  -1.013581  3.107826e-01
study.id.3                     -0.30954982 0.085705184  -3.611798  3.040812e-04
study.id.4                     -0.40625599 0.134384195  -3.023094  2.502049e-03
age.50                          0.06904617 0.001853554  37.250693  1.032568e-303
D$GENDER1                      -0.45085471 0.109355479  -4.122836  3.742360e-05
D$PM_BMI_CATEGORIAL2            0.65056077 0.091792545   7.087294  1.367595e-12
D$PM_BMI_CATEGORIAL3            1.76133790 0.093433333  18.851280  2.868195e-79
D$GENDER1:D$PM_BMI_CATEGORIAL2  0.15644304 0.129168202   1.211158  2.258350e-01
D$GENDER1:D$PM_BMI_CATEGORIAL3  0.24346797 0.126597421   1.923167  5.445909e-02
```

Application in other settings

In this paper we have illustrated the use of DataSHIELD in a setting involving research-focused analysis of data that were originally collected for research purposes. But it could potentially be of equal value in settings involving co-analysis of data from multiple health service or other administrative databases, or the joint analysis of research data with administrative data. It is for these purposes that major infrastructural projects[7,8] like *care.data*[7,8] and secure data-sharing infrastructures[10,11] have been proposed and developed. For example, the aim of *care.data* was to amalgamate medical information on individuals from

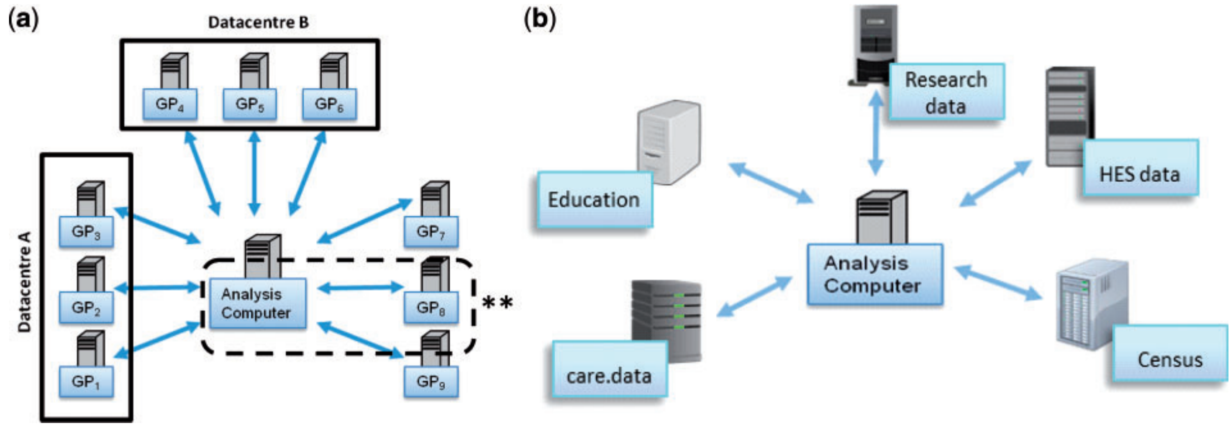

Figure 7. Illustration of DataSHIELD set-up for the analyses of: (a) horizontally partitioned data (similar data, different individuals) held in GP databases and/or data centres. (**Single-site DataSHIELD); and (b) vertically partitioned data requiring record linkage between different types of data on the same individuals held in a variety of data archives.

various administrative sources, including general practice (GP) records, into a single research database held by the Health and Social Care Information Centre and made available to approved researchers. Though some of the concerns that led to the suspension of *care.data* related to the possibility of commercial entities such as pharmaceutical and insurance companies being approved as customers, a central concern was protection of patient confidentiality. If they are ever to succeed, projects like *care.data* must therefore overcome two fundamental data-sharing challenges. First, they must find a safe and appropriate way to allow researchers to analyse data drawn from particular healthcare or other administrative data sources, including GP medical records, wherein the risk of breaching patient confidentiality is reduced to an absolute and acceptable minimum. Depending how many sources need to be accessed, this could potentially be achieved through a conventional—or single-site—application of horizontal DataSHIELD (Figure 7a). The second challenge is to securely combine information relating to individuals from a primary source (e.g. from a particular research project, or from GP medical records) with other health or administrative records on the same individuals, using record linkage and co-analysis. This is essential if some of the required data are not directly available from the primary source (e.g. hospitalization data, or education data). In such a setting, a vertical implementation of DataSHIELD can play a useful role (Figure 7b), although it is out of the scope of this paper to discuss vertical DataSHIELD in detail. In principle, DataSHIELD could provide a means to reassure the public that their data were being used in a secure manner. However, important challenges remain: (i) ongoing technical refinement of the functionality of the vertical implementation of DataSHIELD; (ii) extensive discussion with data-providing agencies—including government—and relevant governance committees to ensure that they are all comfortable with application of DataSHIELD to potentially sensitive administrative data; and (iii) consideration of whether, in any particular setting, the agencies involved will be willing and able to devote the time and resources required to prepare and document data ready for vertical DataSHIELD use. Our future work includes a focus on addressing these challenges.

Discussion

DataSHIELD enables co-analysis of several collaborating studies or data sources as if the data from all individuals in all studies were directly accessible but, in reality, these data remain completely secure behind the firewalls of their host computers. This is of significant value in several settings: (i)where ethico-legal or governance restrictions proscribe individual-level data release, or make permission for such release excessively time-consuming to obtain; (ii) a research group is particularly vulnerable to losing intellectual property (e.g. in a developing nation) but wishes to freely share the information held in its data without physically sharing the data themselves; and (iii)the underlying data are too large to be physically shared.

All components of the combined platform (Opal/DataSHIELD) are open source and available without restriction or payment. Both the installation and the configuration require minimal specialist IT expertise: researchers with no IT background have already installed Opal without major difficulties by following the wiki documentation available online.[42] DataSHIELD is therefore attractive for researchers with limited resources. An extensive suite of functions already exists, but development work continues and we recently started developing a Graphic User Interface that requires no prior knowledge of R to run a DataSHIELD analysis.[43] The newest software release of Opal incorporates an important enhancement. Specifically, DataSHIELD analysis is now truly parallelized: every command is sent simultaneously to all DCs—previously, each command necessarily completed on one DC before being sent to the next. This substantially speeds up analysis, particularly with many studies or time-consuming functions. If processing speed is particularly critical, further time may be saved by distributing the data from a large study across several Opal servers. If there are actual problems with the Opal instance at a given DC, then a message is sent to the data owner to correct that problem (e.g. the version of the libraries currently installed is not up to date, or the server is down). Crucially, if one or more of the data servers are unusable, the user can temporarily exclude them from analysis while they are repaired or updated.

Because in DataSHIELD potentially disclosive commands are not allowed, some analyses that are possible in standard R are not enabled. In essence, there are two classes of limitation on potential DataSHIELD functionality: (i) absolute limitations which require an analysis that can only be undertaken by enabling one of the functionalities (e.g. visualizing individual data points) that is explicitly blocked as a fundamental element of the DataSHIELD philosophy. For example, this would be the case for a standard scatter plot. Such limitations can never be circumvented and so alternatives (e.g. contour and heatmap plots) are enabled which convey similar information but without disclosing individual data points; (ii) current limitations which are functions or models that we believe are implementable but we have not, as yet, undertaken or completed the development work required. As examples, these latter include generalized linear mixed models[44] (including multi-level modelling[45,46]) and Cox regression.[47]

Despite its potential utility, implementation of DataSHIELD involves significant challenges. First, although set-up is fundamentally straightforward, application involves a relatively steep learning curve because the command structure is complex: it demands specification of the analysis to be undertaken, the studies to use and how to combine the results. In mitigation, most complex server-side functions are now called using simpler client-side functions and we are working on a menu-driven implementation. Second, like anyco-analysis involving several studies, data must be adequately harmonized[27,34,35] and the proposed work must comply with governance stipulations in every study. Third, good research governance demands that any published analysis can precisely be replicated. We are therefore developing systems to automatically identify the particular DataSHIELD release used for a given analysis. In addition, each data provider must unambiguously record the particular freeze of data they contributed. These fundamental issues apply in many settings other than DataSHIELD, but because the project could be damaged if early users were to encounter serious scientific or governance problems, application has so far been restricted to research groups with whom we are fully collaborating. This means we can provide active advice and support relating both to implementation and application. We plan to enable independent use as early as possible. Fourth, in undertaking a standard DataSHIELD analysis it is assumed that the data truly are horizontally partitioned, i.e. contributing sources hold the same variables but on different individuals (see Figure 3b). So far DataSHIELD has been applied in settings where individual participants in different studies are from different countries or from different regions so it is unlikely that any one person will appear in more than one source. However, going forward, that cannot always be assumed. We have therefore been considering approaches to identify and correct this problem based on probabilistic record linkage. In the genetic setting the BioPIN[48] provides an alternative solution. Ongoing work is required. Fifth, despite the care taken to set up DataSHIELD so that it works properly and is non-disclosive, it is possible that unanticipated problems (accidental or malicious) may arise. In order to identify, describe and rectify any errors or loopholes that emerge and in order to identify deliberate miscreants, all commands issued on the client server and enacted on each data server are permanently logged.

Data sharing platforms, such as *care.data*, that enable powerful integrative analysis of research data as well as data generated by activity in the health service, from disease or death registries or from other administrative or governmental sources, have the potential to generate great societal benefit. Most crucially, they can provide an important route for production of the raw ‘evidence’ needed for ‘evidence-based health care’. But, to be pragmatic, many of the routinely collected healthcare and administrative databases will have to undergo substantial evolution before their quality and consistency are such that they can directly be used in high-quality research without extensive preparatory work. By its very nature, such preparation—which typically includes data cleaning and data harmonization—cannot usually be undertaken in DataSHIELD, because it involves investigating discrepancies and/or extreme results in individual data subjects: the precise functionality that DataSHIELD is designed to block. Such work must therefore be undertaken ahead of time by the data generators themselves—and this is demanding of time, resources and expertise that—at present - many administrative data providers may well be unwilling and/or unable to provide. That said, if the widespread usability of such data is viewed as being of high priority, the required resources could be forthcoming. Then the primary challenge will be to find effective solutions to the professional and societal challenges presented by the need to ensure that all work with individual-level data is rendered adequately secure. These solutions must respect and protect individual autonomy and confidentiality while facilitating the scientific progress from which everybody benefits. This conundrum is well recognized as demonstrated in the series of articles under the heading Dealing with Data in *Science* in 2011,[49] and more recently in a review article exploring the combination of multiple healthcare databases for postmarketing surveillance of drug and vaccine safety.[50] Furthermore, these challenges and potential solutions provide a crucial focus for professional organizations aimed specifically at enhancing our capacity to make effective use of the rapidly accumulating body of available data in the arenas of health and social care, governmental administration and biomedical and social research. These organizations include major pan-European infrastructural projects in large-scale biomedical sciences such as: *ELIXIR*[51] and *BBMRI* (the Biobanking and Biomolecular Resources Research Infrastructure[52,53]); *EAGDA* (the Expert Advisory Group in Data Access) set up by four major UK funders (Wellcome Trust, MRC, ESRC and Cancer Research UK); the Public Population Project in Genomics and Society[54] and most recently, the Global Alliance for Genomics and Health.[55] DataSHIELD provides a radically different way to keep sensitive data from multiple sources completely confidential while maintaining their full scientific utility; it could prove to be an invaluable complement to other more conventional approaches.

No single approach can provide a perfect universal solution to the challenges arising from the complex interplay between professional and societal wishes, needs and concerns as healthcare and research data become ever richer, increasing both their power for good and their potential risk of disclosure. However, DataSHIELD provides important opportunities that neatly complement other approaches. It has already been proven to work in principle,[12,22] and this paper now addresses the equally taxing problem: how to make it work in practice. DataSHIELD now provides a real opportunity to follow the advice of Kahn in Dealing with Data[56] to move the ‘computation to the data, rather than the data to the computation’.[56]

Funding

This work was supported through funds from: the European Union’s Seventh Framework Programme BioSHaRE-EU, grant agreement HEALTH-F4-2010-261433; the Welsh and Scottish Farr Institutes funded by MRC; joint funding from MRC and Wellcome Trust comprising a strategic award underpinning the ALSPAC project and an infrastructural grant entitled *The 1958 Birth Cohort Biomedical Resource – facilitating access to data and samples and enhancing future utility*; and the BBMRI-LPC project (EU FP7, I3 grant).

Conflict of interest: None declared.

References

1. Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005;**366**:941–51.
2. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;**5**:e1000477.
3. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. *Nat Protocols* 2007;**2**:2492–501.
4. Walport M, Brest P. Sharing research data to improve public health. *Lancet* 2011;**377**:537–39.
5. Burton PR, Hansell AL, Fortier I *et al*. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2008;**38**:263–73.
6. Gomatam S, Karr A, Reiter J, Sanil A. Data dissemination and disclosure limitation in a world without microdata: a risk-utility framework for remote access analysis servers. *Stat Sc* 2005;**20**:163–77.
7. Hoeksma J. The NHS’s care.data scheme: what are the risks to privacy? *BMJ* 2014;**348**:g1547.
8. McCartney M. Care.data: why are Scotland and Wales doing it differently? *BMJ* 2014;**348**:g1702.
9. Demir I, Murtagh MJ. Data sharing across biobanks: epistemic values, data mutability and data incommensurability. *New Genet Soc* 2013;**32**:350–65.
10. UK.Data.Service. *About Secure Access*. http://ukdataservice.ac.uk/get-data/secure-access/about/what-is.aspx (7 March 2014, date last accessed).
11. Ford DV, Jones KH, Verplancke JP *et al*. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;**9**:157.
12. Wolfson M, Wallace SE, Masca N *et al*. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 2010;**39**:1372–82.
13. Newton-Cheh C, Johnson T, Gateva V *et al*. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009;**41**:666–76.
14. Repapi E, Sayers I, Wain LV *et al*. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2010;**42**:36–44.
15. Zeggini E, Weedon MN, Lindgren CM *et al*. Replication of genome-wide association signals in U.K. *Samples reveal risk loci for type 2 diabetes*. Science 2007;**316**:1336–39.
16. Petitti DB. *Meta-analysis, Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. 2nd ed. New York: Oxford University Press; 2000.
17. Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual- and aggregate-level data. *Stat Med* 2008;**27**:651–69.
18. Burman W, Daum R, Janoff E *et al*. Grinding to a halt: the effects of the increasing regulatory burden on research and quality improvement efforts. *Clin Infect Dis* 2009;**49**:328–35.

19. Malfroy M, Llewelyn CA, Johnson T, Williamson LM. Using patient-identifiable data for epidemiological research. *Transfus Med* 2004;**14**:275–79.
20. Burton P, Wolfson M, Masca N, Fortier I. Datashield: Individual-level meta-analysis without sharing the data. *J Epidemiol Commun Health* 2011;**65**:A37.
21. Wallace SE, Gaye A, Shoush O, Burton PR. Protecting personal data in epidemiological research: DataSHIELD and UK law. *Public Health Genom* 2014;**17**:149–57.
22. Jones EM, Sheehan NA, Masca N, Wallace SE, Murtagh MJ, Burton PR. DataSHIELD-shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk Epidemiologi* 2012;**21**:231–39.
23. Jones EM, Sheehan NA, Gaye A, Laflamme P, Burton P. Combined analysis of correlated data when data cannot be pooled. *Stat* 2013;**2**:72–85.
24. Murtagh MJ, Demir I, Jenkings KN *et al*. Securing the data economy: translating privacy and enacting security in the development of DataSHIELD. *Public Health Genom* 2012;**15**:243–53.
25. EGA. *European Genome-Phenome Archive*. https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/european-genome-phenome-archive-ega (13 March 2014, date last accessed).
26. Wallace SE. The needle in the haystack: international consortia and the return of individual research results. *J Law Med Ethics* 2011;**39**:631–39.
27. Doiron D, Burton P, Marcon Y *et al*. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol* 2013;**10**:12.
28. OBiBa. Opal [Opal is OBiBa's core database application for biobanks or epidemiological studies]. 2012. http://www.obiba.org/node/63 (24 June 2014, date last accessed)
29. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *J Comput Graph Stat* 1996;**5**:299–14.
30. Maelstrom. *Maelstrom Research*. https://www.maelstrom-research.org/ (4 March 2014, date last accessed).
31. OBiBa. *Open Source Software for Biobanks*. http://www.obiba.org/?q=node/1 (04 March 2014, date last accessed).
32. Healthy Obese Project. *Healthy Obese Project*. 2013. https://www.bioshare.eu/content/healthy-obese-project (19 March 2014, date last accessed).
33. BioSHaRE-EU. BioSHaRE.eu. https://www.bioshare.eu/ (19 June 2014, date last accessed).
34. Fortier I, Burton PR, Robson PJ *et al*. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39**:1383–93.
35. Fortier I, Doiron D, Little J *et al*. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol* 2011;**40**:1314–28.
36. Kuk A, Cheng Y. The Monte Carlo Newton-Raphson Algorithm. *J Stat Comput Sim* 1997;**59**:233–50.
37. OBiBa. *Mica*. http://obiba.org/node/174 (4 March 2014, date last accessed).
38. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* 2006;**35**:34–41.
39. Wichmann H, Gieger C, Illig T. KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* 2005;**67**:S26.
40. Stolk RP, Rosmalen JG, Postma DS *et al*. Universal risk factors for multifactorial diseases. *Eur J Epidemiol* 2008;**23**:67–74.
41. Kearney PM, Harrington JM, Mc Carthy VJ, Fitzgerald AP, Perry IJ. Cohort Profile: The Cork and Kerry Diabetes and Heart Disease Study. Int J Epidemiol *2013* 2013;**42**:1253–62.
42. OBiBa. *Opal documentation*. 2014. http://wiki.obiba.org/display/OPALDOC/Home (27 June 2014, date last accessed).
43. Gaye A, Burton WY. *DataSHIELD Online Interactive Terminal*. 2014. https://www.bioshare.eu/datashieldgui/ (29 June 2014, date last accessed).
44. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88**:9–25.
45. Goldstein H. Multilevel mixed linear modelling analysis using iterative generalized least squares. *Biometrika* 1986;**73**:43–56.
46. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998;**17**:1261–91.
47. Cox DR. Regression models and life-tables. *J R Stat Soc* 1972;**B;34**:187–220.
48. Nietfeld JJ, Sugarman J, Litton JE. The Bio-PIN: a concept to improve biobanking. *Nat Rev Cancer* 2011;**11**:303–08.
49. Hanson B, Sugden A, Alberts B. Making data maximally available. *Science* 2011;**331**:649.
50. Trifirò G, Coloma P, Rijnbeek P *et al*. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Int Med* 2014;**275**:551–61.
51. Elixir. *Elixir, Data For Life*. 2014. http://www.elixir-europe.org/ (27 June 2014, date last accessed).
52. BBMRI-ERIC. *Managing Resources for the Future of Biomedical Research*. http://bbmri-eric.eu/ (27 June 2014, date last accessed).
53. BBMRI-LPC. *Helping Europeans Get Healthier*. http://www.bbmri-lpc.org/ (27 June 2014, date last accessed).
54. Public Population Project in Genomics and Society. *P3G HOME*. http://p3g.org/.
55. Global Alliance 4 Genomics and Health. *Web site*. *2014*. http://genomicsandhealth.org/
56. Kahn SD. On the future of genomic data. *Science* 2011;**331**:728–29.