Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2010;**39**:1372–1382

doi:10.1093/ije/dyq111

DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data

Michael Wolfson,[1] Susan E Wallace,[2,3] Nicholas Masca,[4] Geoff Rowe,[1] Nuala A Sheehan,[4] Vincent Ferretti,[3,5] Philippe LaFlamme,[3,6] Martin D Tobin,[4] John Macleod,[7] Julian Little,[3,8] Isabel Fortier,[3,8,9] Bartha M Knoppers[2,3] and Paul R Burton[3,4,8,10]*

[1]Statistics Canada, Ottawa, Ontario, Canada, [2]Centre of Genomics and Policy, Faculty of Medicine, Department of Human Genetics, McGill University, Montreal, Quebec, Canada, [3]Public Population Project in Genomics (P3G), Montreal, Quebec, Canada, [4]Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK, [5]Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada, [6]McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada, [7]Department of Social Medicine, University of Bristol, Bristol, UK, [8]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, [9]Department de Médecine Sociale et Préventive, Université de Montréal, Montreal, Quebec, Canada and [10]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

*Corresponding author. Departments of Health Sciences and Genetics, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK. E-mail: pb51@le.ac.uk

Accepted 27 May 2010

Background Contemporary bioscience sometimes demands vast sample sizes and there is often then no choice but to synthesize data across several studies and to undertake an appropriate pooled analysis. This same need is also faced in health-services and socio-economic research. When a pooled analysis *is* required, analytic efficiency and flexibility are often best served by combining the individual-level data from all sources and analysing them as a single large data set. But ethico-legal constraints, including the wording of consent forms and privacy legislation, often prohibit or discourage the sharing of individual-level data, particularly across national or other jurisdictional boundaries. This leads to a fundamental conflict in competing public goods: individual-level analysis is desirable from a scientific perspective, but is prevented by ethico-legal considerations that are entirely valid.

Methods Data aggregation through anonymous summary-statistics from harmonized individual-level databases (DataSHIELD), provides a simple approach to analysing pooled data that circumvents this conflict. This is achieved via parallelized analysis and modern distributed computing and, in one key setting, takes advantage of the properties of the updating algorithm for generalized linear models (GLMs).

Results The conceptual use of DataSHIELD is illustrated in two different settings.

Conclusions As the study of the aetiological architecture of chronic diseases advances to encompass more complex causal pathways—e.g. to include the joint effects of genes, lifestyle and environment—sample size requirements will increase further and the analysis of pooled individual-level data will become ever more important. An aim of this conceptual article is to encourage others to address the challenges and opportunities that DataSHIELD presents, and to explore potential extensions, for example to its use when different data sources hold different data on the same individuals.

Keywords Pooling, analysis, meta-analysis, individual-level, study-level, generalized linear model, GLM, ethico-legal, ELSI, identification, disclosure, distributed computing, bioinformatics, information technology, IT

Introduction

Most known associations between genetic variants and chronic diseases reflect weak effects with typical allelic odds ratios in the range 1.1–1.4.[1–3] The reliable identification of such effects demands vast data sets.[1–5] Case–control studies including thousands of cases are required even when interest focuses on the simplest situation: the detection of the direct effects of single nucleotide polymorphism (SNP) variants.[1–3] Furthermore, when, as is likely, scientific emphasis starts to focus on the study of gene–environment and gene–gene interactions and the exploration of causal pathways more comprehensively, tens of thousands of cases will often be required.[1] Tens of thousands of subjects can also be required to study a quantitative phenotype (e.g. measured blood pressure), because allelic effect sizes may be as small as one-tenth of a standard deviation, or even less.[6–8]

To achieve sample sizes as large as this, it is often necessary to pool data across multiple studies, and large collaborative consortia have been responsible for much of the recent progress in human population genomics.[6,8–16] Large-scale data pooling is equally important in other settings too: in mainstream epidemiology[17]—particularly in the analysis of formal networks of studies[18,19]—in public health and health-services research, and in comparative international analysis in the social sciences, including coordinated economic surveillance.[20,21] Such pooling not only supports the attainment of large sample sizes but can also be used to reduce bias arising from access to a restricted subset of data. But, regardless of its purpose, the sharing of data always raises important ethico-legal issues even when the analysis is mutually agreed. Data privacy, for example, is a hot topic in genomic epidemiology,[22,23] as well as being a concern for government, industry,[24,25] the media and even the general public.[26] Biomedical science has responded cautiously to these concerns, ensuring that all ethico-legal stipulations are met and that new issues are dealt with carefully, as and when they arise.[23,27]

Given this caution, it is perhaps surprising that there has been such striking recent progress in detecting genetic associations with complex diseases:[3,28] 'in the past three years genome-wide association studies (GWAS)…. have reproducibly identified hundreds of associations of common genetic variants with over 80 diseases and traits (http://www.genome.gov/gwastudies)'.[9] But, in one sense, genomic epidemiology has been fortunate. The class of pooled analysis that has underpinned many of the recent successes,[6,8–16] just happens to be consistent with the ethico-legal frameworks that large-scale bioclinical studies have had in place over many years. That is, most such studies are permitted to take part in collaborative GWAS based on study-level meta-analysis (SLMA).[29,30] Here, investigators from each study perform a separate GWAS, and then share the association statistics for each SNP with a designated analysis centre (AC); but the raw data encoding SNP and disease status are *not* shared.[6,7,11] The AC then performs a meta-analysis to estimate the genetic associations across the consortium as a whole. But, bioscience will inevitably move on from its current focus on simple associations between genetic variants and disease-related traits, to explore causal pathways more thoroughly: e.g. by incorporating gene–environment interactions. This will increase sample size requirements further,[1] making data pooling yet more essential. In addition, data analysis will become increasingly unpredictable and, therefore, exploratory. For example, in a conventional meta-analysis-based GWAS it is clear *a priori* that each study must generate summary statistics to reflect the association of the disease of interest with each of a large number of designated SNPs (e.g. 1 million). This is onerous but it can be pre-specified ahead of time. The required set of summary statistics is far more difficult to predefine if the analysis is to involve gene–environment interactions; environmental and lifestyle factors may be parameterized in many different ways, and identification of the

appropriate parameterization often demands initial exploratory analysis.

Analytic and ethico-legal considerations

Large-scale statistical pooling is typically achieved in one of two ways.[29,30] First, the individual level data from each of the original data sources can be aggregated to produce one combined data set. This is then analysed as if it were generated by a single study, though study-to-study heterogeneity may necessitate the inclusion of study-specific model terms. This approach may be called individual-level meta-analysis (ILMA). Secondly, appropriate summary statistics can be generated from separate analyses carried out on each independent study, and these then pooled in an SLMA. SLMA is quick and convenient when based on summary statistics that already exist or can be easily derived *de novo*. It is therefore the approach to meta-analysis that is often adopted in public health research, the meta-analysis of randomized controlled trials and, recently, in the pooling of GWAS studies.[6–8,29–31] But, it has important limitations. First, although it is very convenient to use summary statistics that are already in the public domain, it is important to recognize that they can be biased by selective reporting dependent on findings. In the field of genomic epidemiology this can be particularly problematic.[32] Secondly, even when summary statistics are derived *de novo*, SLMA can be restrictive.[30] The analysis of all but the simplest of biomedical problems demands a significant element of exploration, but analysis in a conventional SLMA is unavoidably restricted to questions that can be addressed using the particular set of summary statistics that was initially requested.[30] If an important new question arises, it can only be answered if the investigators are all prepared to produce the new summary statistics that are required. This can cause serious delays.

In consequence, ILMA would often be preferred to SLMA. But, ILMA raises major ethico-legal challenges. Most notably the sharing of individual level data, sometimes termed 'microdata',[24] may be prohibited in law. In many jurisdictions, individual-level data are treated as being fundamentally different to aggregate data, and some individual-level data cannot cross certain national boundaries.[33] Even when sharing *is* legal, it may be proscribed by the consents and ethical approvals under which the data were initially collected.[34] And, even when—*in principle*—microdata can be shared, that sharing can demand protracted applications for access via scientific oversight committees and ethical review boards.[35,36] But these barriers are there for a good reason; the relevant ethico-legal considerations reflect important values held by many societies. Individual-level data can disclose identity,[24] they may be highly sensitive[24] and they may yield unexpected scientific knowledge of great practical or theoretical value, which the original investigators, funders, national governments and even study participants might feel wary about passing on to a third party.[23] The fundamental importance of these issues is indicated by the fact that they are addressed by the ethico-legal and governance provisions of almost all major bioclinical studies. To illustrate, Box 1 provides exemplar language[37] from the ethico-legal documentation of a number of international biobanks and cohort studies, and from the Model Consent Form prepared by the Public Population Project in Genomics (P^3G).[38] The quotes are not ascribed to particular studies because anonymity was guaranteed as part of the formal agreement under which this ethico-legal documentation was originally shared with P^3G.

Resolving a real conflict between 'competing public goods'

Although ILMA offers many advantages in terms of analytic flexibility,[29,30] it is therefore clear that ethico-legal restrictions on the transfer of individual-level data to third parties mean that a conventional ILMA approach is often impractical. Since this conflict in 'competing public goods' was identified, it has been discussed extensively by the international biobanking community; for example, in forums provided by P^3G, Promoting Harmonization of Epidemiological Biobanks in Europe (PHOEBE) and Biobanking and Bio-molecular Resources Research Infrastructure (BBMRI). These discussions have led to the rapid evolution of a novel approach to analysis that could, in theory, circumvent the conflict identified. The proposed approach is named DataSHIELD (Data aggregation through anonymous summary-statistics from harmonized individual level databases). This conceptual article describes the approach proposed, demonstrates that it works in theory, explores its potential uses and extensions, and discusses some of the challenges to be faced in implementing it. It is our hope that by sharing the concept with the broader research community, we will encourage others to work with us in undertaking a pilot implementation.

Methods

The conceptual underpinning of DataSHIELD is straightforward. Modern distributed computing is used to realize the full benefits of ILMA without physically sharing any individual-level data. All data remain on the local computers at their studies of origin and the role of the AC is to coordinate a parallelized analysis of the individual-level data on all of those local computers simultaneously. Critically, the parallelized analysis is so framed that the only information passing back and forth between computers consists of short blocks of computer code specifying

Box 1 Examples of language used in relevant ethico-legal documentation including consent forms and information leaflets

Examples of language used in the ethico-legal documentation of selected international biobanks and cohort studies

(1) Language restricting the scope of data sharing

Use of data restricted to researchers participating in the original study

(a) 'All research data are confidential... they will only be used in medical research and [will] remain in the sole use of the participating researchers.'

Use of data restricted to researchers in one country

(b) 'Blood and DNA samples may...be distributed to laboratories...around [country] for further research.'
(c) 'Research using the anonymous samples will be done by [researchers] ...throughout [country].'

(2) Language ensuring data de-identification

(a) '[Project] will give researchers restricted access to... anonymous samples to conduct [research]...'
(b) 'Researchers authorised by [Project] will have access to ... coded information...'
(c) '[Project] researchers or their collaborators at other research institutions... may be allowed access to your DNA sample and medical information, but they will not get... links to your identity.'

Examples of language used in the P^{3}G 'Model Consent form'

(1) The need to obtain both scientific and ethical approval

(a) 'The [Project] gives approved researchers access to data and samples... All researchers will only have access to coded data or samples, in order to protect your privacy. They also have to obtain prior scientific and ethical approval as described above, and their research must fit the purpose of the resource/biobank.'
(b) 'The [Project] expects to receive requests and, if approved, provide access to data'.

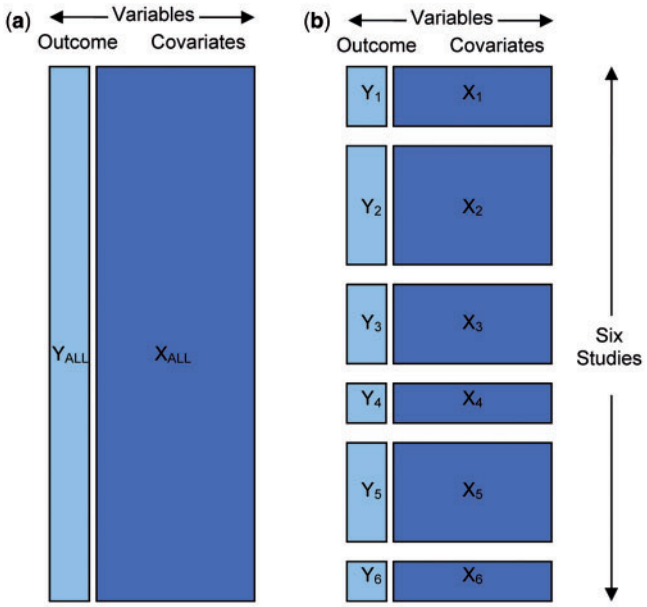

Figure 1 Schematic representation of structure of scientific problems that DataSHIELD is designed to address. (a) One file: all individual-level data pooled together in one large data file. (b) Partitioned: individual-level data held in six separate data files, one for each study

the next analysis required, and low-dimensional summary statistics used in estimating the mathematical parameters of the model (e.g. means or regression coefficients). These items disclose neither the identity, nor the characteristics, of individual study participants.

Figure 1 provides a schematic representation of the class of analytic problems that DataSHIELD is aimed at addressing; here, data are distributed across six sources. The aim is to estimate the statistical parameters that characterize the relationship between an outcome variable Y and one or more explanatory variables X. Here the data are horizontally partitioned:[25] i.e. each data set includes all of the variables (X and Y) but on different sets of individuals. A classical ILMA would involve stacking the data matrices from each study to produce one large data matrix (Figure 1a). Under DataSHIELD (Figure 1b), on the other hand, a series of parallel analyses are undertaken simultaneously—using X_j and Y_j in the jth study—and these analyses are synthesized in an appropriate manner to generate estimates pertaining to all six studies simultaneously.

Figure 2 provides a schematic representation of the type of IT infrastructure that might typically be

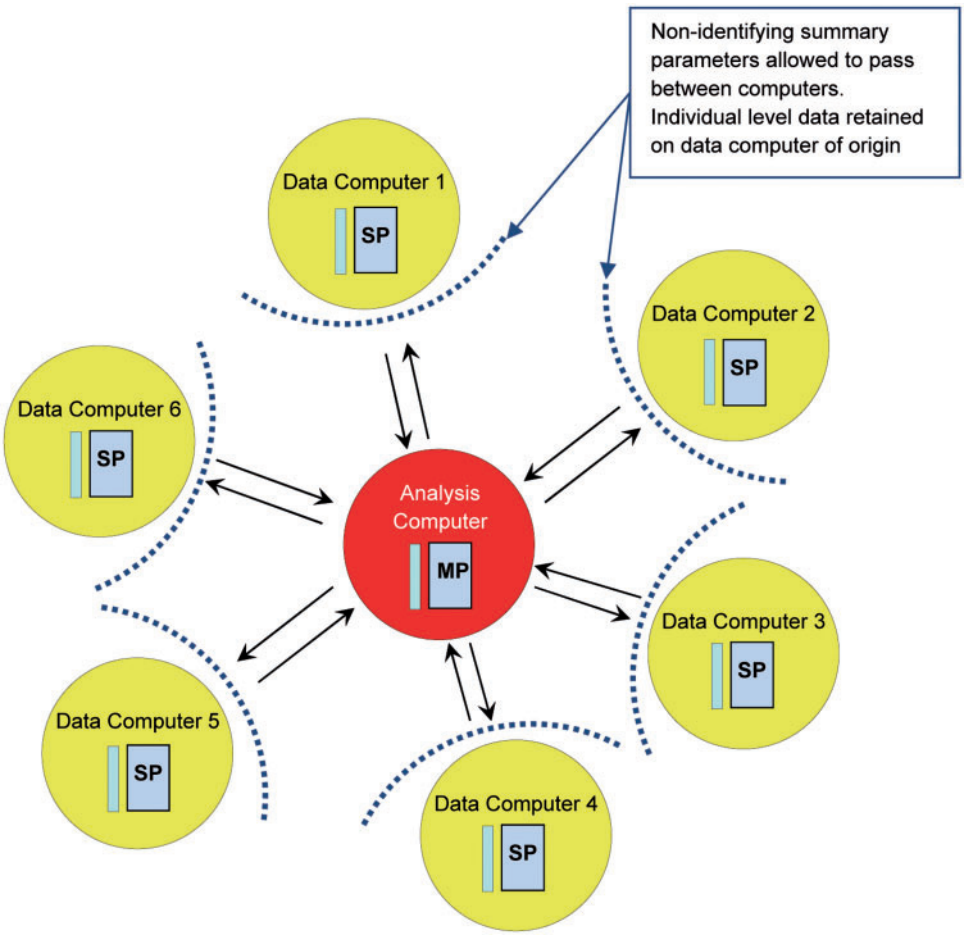

Figure 2 Schematic representation of the structure of DataSHIELD. The computer controlling analysis (heavily shaded circle) is sited at the analysis centre (MP: master process). The data computers (lightly shaded circles) are each sited at one of the study centres involved in the collaborative analysis (SP: slave process). The arrows indicate the flow of analytic instructions and summary statistics. All potentially disclosive individual-level data are secured on the local data computers

required to undertake a DataSHIELD analysis. The computers on which the individual-level data reside at each of the six centres are depicted as lightly shaded circles. One centre is designated the AC and it is a computer (the heavily shaded circle) at that centre that is used to coordinate and execute the analysis. Often, the AC will be one of the studies that are contributing data to the analysis. The analysis software/middleware in DataSHIELD will require two primary components: (i) a master process (MP) that resides on the coordinating computer at the AC; and (ii) a series of slave processes (SPs), each residing on the local data computers. This structure will enable analytic subroutines to be written by the AC, and then transmitted and activated in a suitable software environment (e.g. in 'R'[39]) on each of the data computers. As an analytic session proceeds, the analysis will evolve and the algorithm that is active on each SP will therefore change. It is the MP at the AC that will control which algorithms are running on which computers at which point in time.

Example 1

Using DataSHIELD to enhance the flexibility of SLMA

Perhaps the simplest application of DataSHIELD might entail the replication of a conventional SLMA. To illustrate this setting, data have been simulated for six hypothetical studies (for details see Supplementary Data: S1 available at *IJE* online) that have assessed peripheral systolic blood pressure (SBP in $mmHg^{-1}$) as a quantitative outcome variable and two explanatory covariates: AGE (years, centralized by subtracting the mean of 60 years); and an SNP (coded 0, 1 or 2, to reflect the number of copies of a minor allele). An illustrative analysis might involve fitting a multiple linear regression model to estimate a regression intercept ($b_{\text{intercept}}$) and regression coefficients b_{AGE} and b_{SNP} associated with the two covariates. Scientific interest might focus on b_{SNP} to provide an age-adjusted estimate of the increase in SBP associated with each additional copy of the minor allele.

Box 2 Exemplar code and output for Scenario 1

The statistician types:

```
regression.model<-lm(SBP~AGE+SNP)
results.matrix<-summary(regression.model)$coefficients[,1:2]
```

Thereby producing a results matrix for each study[a]: for example,

	Estimate	Std. Error[b]
(Intercept)	125.130	0.2629
AGE	0.203	0.0373
SNP	0.254	0.3907

[a]Here, the results shown are for simulated study 6
[b]Standard Error

If the parallelized analyses are to be undertaken in 'R',[39] the statistician at the AC might type the two lines of code at the top of Box 2. Using an appropriate scripting language such as Perl[40] this code could be packaged and transmitted to each of the SPs where it could be piped to R to fit the required regression model on the local data set. This will generate a results matrix (three rows, two columns) comprising an estimate and standard error for each regression coefficient (bottom of Box 2). Additional scripting instructions will then command each study to transmit its results matrix back to the AC. There, the study-specific results can be pooled using an appropriate form of SLMA, to produce parameter estimates and standard errors for all six studies combined. This analysis is detailed in Supplementary Data: S1 (available at *IJE* online).

This DataSHIELD analysis, as outlined, is mathematically equivalent to a conventional SLMA, and all individual-level data remain secure on their computers of origin. But, the first stage (estimation of regression coefficients and standard errors) is controlled remotely by the AC, rather than being carried out by the investigators at each study independently, at the request of the AC. This difference is crucial, because it means that once the initial regression model (Box 2) has been fitted, it is easy to fit a different model that may contain terms for which summary statistics might not, originally, have been requested; for example, one containing an interaction between the AGE and SNP covariates. This would be impossible in a conventional SLMA unless this supplementary analysis had explicitly been pre-specified. This demonstrates that, in principle, DataSHIELD permits SLMA to be undertaken more flexibly. But it offers far more than this. Perhaps most crucially, it allows researchers to make efficient use of an important and versatile class of mathematical models in a manner that is mathematically identical to a full ILMA.

Example 2

Using DataSHIELD to undertake ILMA without sharing the data

Many important analyses in contemporary biopopulation science can be framed as generalized linear models (GLMs).[41] This broad class of models incorporates many forms of regression—e.g. multiple linear regression, logistic regression, Poisson regression and many types of survival analysis. It also subsumes numerous other analytic procedures including *t*-tests, analysis of variance and estimation based on contingency tables.[41] GLMs are usually fitted iteratively using the iteratively reweighted least squares (IRLS) algorithm.[42] An initial guess at the required regression coefficients is progressively refined, over a number of iterations, until maximum likelihood estimates are obtained. Conveniently, in the present context, updating the coefficient estimates at any given iteration depends solely on an information matrix and a score vector, both of which can be obtained by fitting a single iteration of the same GLM to the individual-level data from each of the collaborating studies one at a time, and by summing them in the AC. The two sums may then be used to update the regression coefficients at that iteration[42] (for details see Box 3 and Supplementary Data: S2, available at *IJE* online). The regression coefficients and standard errors that are obtained in this manner are *identical* to those that would be obtained by fitting the same GLM to the pooled individual data from all studies combined, but the AC never has access to the individual-level data.

Results

The mathematics underpinning the IRLS algorithm guarantee that the DataSHIELD approach, as implemented in Example 2, will produce the same results as fitting the equivalent GLM to the individual-level data from all studies combined (for details, see Supplementary Data S2 and S4 at *IJE* online). Box 3 provides a concrete example to confirm this claim. It outlines the analysis of a second simulated data set consisting of six hypothetical studies set up to investigate the relationship between the risk of acute myocardial infarction, body mass index (BMI) and an SNP. Full details of the simulation, analysis, computer code and results are provided in Supplementary Data: S3–S6 at *IJE* online). In contrast to the simple model used in Example 1, this GLM incorporates an interaction term to reflect heterogeneity in the magnitude of the increase in risk of myocardial infarction for a given increase in BMI.

As proof of principle, the estimated regression coefficients and standard errors reported at the bottom of Box 3 are precisely the same, rounding error aside, as those derived from a conventional logistic regression model fitted to a single data set comprising the

Box 3 Simulated data example

DATA: six case-control studies

Study	Cases	Controls	Total
1	962	1038	2000
2	1486	1514	3000
3	761	739	1500
4	142	158	300
5	1036	964	2000
6	360	340	700

MODEL FITTING USING DataSHIELD:

Data items that are transmitted between computers are highlighted in **bold.** For additional details see Supplementary Materials. The DataSHIELD analysis is predicated on the assumption that all ethico-legal requirements have been met and that the data are adequately harmonized.

MODEL:

Outcome variable:

CC *case control status* - cases fulfil formal criteria for an acute myocardial infarction; controls are healthy and population based: 1=case, 0 = control

Explanatory covariates:

BMI *body mass index*: [kg/m^2] centralised by subtracting 23 kg/m^2. Increasing BMI is known to have greater health consequences in studies 4, 5 and 6, and the model must therefore allow for study-to-study heterogeneity. This is achieved by adding an interaction term to the model (**BMI.456**: taking the value 0 in studies 1-3; and the value of BMI in studies 4-6).

SNP *single nucleotide polymorphism*, minor allele frequency = 0.3:
0=no copies of minor allele; 1=one copy; 2=two copies

Model formula:

LP = $b_{Intercept}$ + b_{BMI}×BMI + $b_{BMI.456}$×BMI.456 + b_{SNP}×SNP

CC~binomial(1,exp[LP]/(1+exp[LP]))

This is a conventional logistic regression model with an additive genetic effect. An aim of analysis is to derive maximum likelihood estimates for the four regression coefficients: $\hat{b}_{Intercept}$; $\hat{b}_{BMI}$; $\hat{b}_{BMI.456}$; $\hat{b}_{SNP}$

Step 1: Analysis Centre (AC) writes code to run one iteration of model in R and transmits this to all six data servers
Transmission AC → DS (all 6 DS) **short block of computer code** (see Supplementary Materials for details).

Step 2: AC provides initial guess for four regression coefficients (here, all 0) and passes vector to all six data servers
Transmission AC → DS (all 6 DS) **vector of regression coefficients: [0, 0, 0, 0]**

Step 3: Analysis Centre tells each data server to run the computer code using the specified vector of regression coefficients and the data held locally on that server
Transmission AC → DS (all 6 DS) **instruction to run the control code once**

Step 4: Control code on each data server generates one matrix and one vector – and these summary statistics are both transmitted to the Analysis Centre.
Transmission DS → AC (each DS transmits one matrix and one vector, but each study transmits different values). *e.g.* in this example, the **matrix and vector** generated by study 5 and transmitted to the AC are:

$$(1) = \begin{bmatrix} 500 & 70.56657 & 70.56657 & 297 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 297 & 65.39412 & 65.39412 & 382 \end{bmatrix}; \quad (2) = [36,\ 487.2951,\ 487.2951,\ 149]$$

Step 5: AC sums both components across all six studies. The overall sum of the vectors is multiplied by the inverse of the overall sum of the matrices to produce a vector containing four update terms (one for each regression coefficient). These are added to the current vector of regression coefficients to produce refined estimates

Step 6: Return to step 2, and transmit the vector of refined estimates of the regression coefficients to all data servers
Transmission AC → DS (all DS) **vector of regression coefficients**: *e.g.* in this example, the updated vector transmitted at the start of the second iteration is: **[**-0.32183281, 0.02228647, 0.03911561, 0.53516954**]**

Steps 2-6 are repeated until all estimates stabilize from iteration to iteration. At this point, the refined coefficients passed on in step 2 represent the maximum likelihood estimates of the regression coefficients. Furthermore, standard errors for these coefficients can be obtained by calculating the inverse matrix of the sum of the matrices in step 5, and taking the square root of the elements down the diagonal. In this example the final results obtained are:

Parameter	Coefficient	Standard Error
$\hat{b}_{intercept}$	-0.3296	0.02838
$\hat{b}_{BMI}$	0.02300	0.00621
$\hat{b}_{BMI.456}$	0.04126	0.01140
$\hat{b}_{SNP}$	0.5517	0.03295

Box 4 A conventional logistic regression analysis [glm() in 'R'] on pooled data from all six studies combined

	Estimate	SE	z-value	Pr(>|z|)
Coefficients:				
(Intercept)	−0.32956	0.02838	−11.612	<2e-16
BMI	0.023	0.00621	3.703	0.000213
BMI.456	0.04126	0.0114	3.62	0.000295
SNP	0.55173	0.03295	16.746	<2e-16

individual level data from all six studies combined (Box 4). But (see Box 3 and Supplementary Data: S3–S5 at *IJE* online), information flow between the data sources and the AC is restricted to: (i) repeated instructions from the AC to the data computers to execute each new iteration of the GLM; (ii) non-disclosive summary statistics (one matrix and one vector) passed back from each data computer to the AC at the end of each iteration; (iii) the updated vector of regression coefficients—again non-disclosive—passed from the AC to the data computers at the start of each new iteration. None of these items is disclosive of identity or of sensitive information.

Discussion

This article demonstrates that if all ethico-legal and informatics challenges can be overcome then, in principle, DataSHIELD should enable a full pooled analysis of individual-level data from multiple sources to be undertaken, even when ethico-legal considerations might otherwise obstruct the physical sharing of that individual-level data. At present, DataSHIELD is no more than a concept and there is a quantum leap between proving that the mathematics work and actually implementing the approach in practice. The principal challenges are in developing the IT systems required, in determining whether ethical review committees agree that there is a real problem to be solved and that DataSHIELD provides a workable solution to that problem, and in implementing the local infrastructures at individual biobanks and cohort studies (staff and equipment) to enable its use. These challenges are substantive and it might be argued that publication should await successful implementation. The European Union has recently awarded funding under Framework 7 (the BioSHARE-EU project) to enable preliminary work to develop and pilot the required IT systems and to explore the relevant ethico-legal and social issues. Given that the implementation work will now definitely take place, it is critical to enrol studies, as pilot sites, to work with us in implementing and trialling the method. Furthermore, the preliminary work will include exploring the fundamental problem with research ethics committees and determining whether they view DataSHIELD as a viable solution. We hope this article will assist studies, biobanks and research ethics committees to determine whether they wish to contribute to such a project.

Development to date has been undertaken by an international group that includes leading bioinformaticians and ethico-legal experts. On the basis of active discourse between these experts and the broader international biobanking community (via P^3G and BBMRI), the prevailing viewpoint seems to be that there is a real problem to be overcome and that the fundamental challenges both in the IT and ethico-legal domains can, in principle, be overcome. For example, there is a broad consensus amongst ethico-legal experts that the physical sharing of individual level data between research groups must be subject to appropriate governance and that it is an inescapable fact that the formal documentation and oversight systems in certain studies (see Box 1) proscribes or discourages such sharing. The real challenge, therefore, is to explore whether DataSHIELD provides a workable solution. Bioinformaticians believe that the IT interface should be set up in a manner that actively prevents the AC from tunnelling into the local systems to extract data or other information and/or from fitting models that reveal identifying or sensitive data either directly or by logical deduction. It is therefore commonly argued that the DataSHIELD interface should parse all incoming and outgoing messages and then block and record any request, or series of requests,[43] that might, by accident or design, lead to the transmission of inappropriate information. Encouragingly, it seems to be the view of most IT experts that an interface with these characteristics can, in principle, be constructed, and will be feasible to use in practice. This optimistic viewpoint is supported by the fact that secure single-site interfaces already exist allowing external users to specify analyses and then to extract results—but, crucially, no more than results. For example, such an interface is at the heart of the UK's Economic and Social Research Council Secure Data Service.[44] Provided this optimism proves to be well founded, the majority view amongst ethico-legal and biobanking experts with whom DataSHIELD has been discussed seems to be that DataSHIELD might then be seen as being equivalent to conventional SLMA. This is because, in both settings, information flow between data providers and the AC is restricted entirely to analytic instructions and non-identifying summary statistics. If research ethics committees hold the same viewpoint, any study that is currently able to contribute to a conventional SLMA-based meta-analysis (including GWASs) should, in principle, be permitted to make use of DataSHIELD, and the formal ethical and governance requirements

should be equivalent. Ultimately, however, the only definitive proof that DataSHIELD will work and will be accepted by ethics review boards is to implement it for real—the publication of this conceptual article is an important step towards that aim.

The mathematics underpinning DataSHIELD is neither novel, nor difficult to implement.[29,30,41,42] For example, the fitting of a GLM requires no more than a partitioned modification of the conventional IRLS algorithm[41,42] (see Supplementary Data: S2–S5 at *IJE* online). Rather, the originality of the method lies in the basic concept itself. Interestingly, a similar idea has previously been floated in the technometrics literature,[25] and although this means that we cannot claim precedence, it strengthens the academic foundation of the proposal. Critically, the approach seems not to have been noted by statisticians, bioinformaticians or ethicists working in the field of biomedical research and it has neither been promoted nor applied in this important domain. From a technical perspective, our implementation via GLMs might be viewed as a special case of what the technometrics paper refers to as '*secure maximum likelihood estimation*'.[25] But, the maximum likelihood case is considered only in broad generality in that paper, and there is no specific focus on generalized linear models.[25] Furthermore, our implementation via GLMs circumvents some of the 'complications' that the technometrics authors note could arise in the more general case.[25] Our article therefore brings an exciting and potentially important new concept to the attention of the biomedical research community, and illustrates the practical implementation of that approach via a broad class of models (GLMs) that already has a wide range of applications in bioscience.

The extensive discussion of DataSHIELD since its initial proposal has resulted in a number of important extensions to the concept. The first is to expand the remit of the approach to work with data sets that are vertically[25] rather than horizontally partitioned. In contrast to horizontal partitioning (Figure 1), under vertical partitioning the different data sources contain different data items on the same primary set of individuals. Such a scenario occurs commonly when a major cohort study, such as ALSPAC (Avon Longitudinal Study of Parents and Children), links to secondary (often governmental) data sources to enrich the information that are available for analysis.[45] Critically, the data in such secondary sources are often sensitive and can be protected against misuse by prohibiting their physical release. This same problem arises regularly in cross-jurisdictional analyses being undertaken or overseen by, national statistics agencies such as Statistics Canada or Statistics UK. The mathematics underpinning the solution to the problem of vertical partitioning is '*substantially more complex*'[25] than that for horizontal partitioning but, in principle, a solution does exist in the form of an approach known as '*secure matrix products*'.[25] If this approach can successfully be implemented, this will markedly enhance the utility of the proposed DataSHIELD approach. The second extension that has been proposed is to take advantage of the approach to help bioscience deal with the pooled individual level analysis of data sets that cannot physically be shared, because of their vast physical size. As illustrative examples, such sources may include full genome sequence data or medical images on large numbers of subjects. Finally, we note that DataSHIELD can prove helpful in any meta-analytic setting where analysis at the level of individual patient records would be scientifically desirable, but ethico-legal considerations discourage ILMA. For example, a reviewer has noted that ILMA permits subgroups of subjects in a given study to be added or removed, which might be valuable when exploring the implications of an intention-to-treat analysis. Although care would have to be taken to ensure that such subgroups were not identified in a potentially disclosive manner, DataSHIELD could address this issue if the subgroups were appropriately flagged.

As an important aside, the genomics world is still grappling with the implications of the work of Homer *et al*.[23] A question that is regularly asked of DataSHIELD is whether it would protect against the form of inferential disclosure[24] described and explored by Homer *et al*. The simple answer is 'no', because disclosure under Homer *et al*. is based on summary statistics reflecting study-wide genotype distributions at each of many SNPs and is therefore totally unrelated to the third party release of individual-level data. This implies that the specific concerns raised by Homer *et al*.[23] cannot be invoked as being part of the rationale for controlling third party release of individual level data and, as a corollary, that these problems cannot be prevented by using DataSHIELD. But, this does raise an obvious follow-up question: 'Are there *other* circumstances where *summary parameters* can become identifying?'. This is relevant, because DataSHIELD relies on the transmission of summary statistics that are assumed to be non-disclosive. One recognized form of inferential disclosure is termed residual disclosure.[43] Here, the differences between a series of closely related summary statistics—that are themselves non-disclosive—permit precise inferences to be drawn about identity and attribute. It is therefore clear that other scenarios do exist in which summary data can become identifying and some of these may be, as yet, unknown. This emphasises the importance of introducing DataSHIELD cautiously. Because the particular set of summary statistics to be transmitted will vary from one class of problem to another, the potential risk of disclosure will require thorough investigation whenever a new class of models is introduced. Some types of model, such as GLMs,[41,42] are unlikely to be disclosive, not least because they are of low dimension: they typically have few parameters relative to the number of study participants. But the same may not be true of

other models, such as those containing large arrays of random effects.[46] This latter might restrict the fitting of generalized linear mixed models[46] (for example by excluding models where there is a random effect for any single subject). On the other hand, it may prove possible to hold the random effects on the local data computers, while transmitting non-disclosive parameters such as the local variance of the random effects. This requires extensive methodological work, but is an area that we believe would be of considerable theoretical interest to many biostatistics research groups.

Regardless of how data pooling is to be approached, two absolute criteria must always be fulfilled. First, all ethico-legal stipulations must be met. This implies that if it is unclear whether the governance rules of a particular study permit DataSHIELD to be used, that uncertainty must be resolved before DataSHIELD is implemented on that study. Secondly, the data to be amalgamated across studies must be sufficiently similar to allow them to be pooled. Two data sets may be said to be harmonized for a given set of variables in a particular scientific setting, if it is valid and feasible to pool them in that setting. DataSHIELD should not be used unless the studies to be pooled are harmonized. This requires a formal judgement to be made, and methods and tools exist to help scientists make this judgement in relation to pre-existing studies: these include the DataSHaPER (http://www.datashaper.org) in population genomics and epidemiology, and the methods advocated by the Luxembourg Income Study (http://www.lisproject.org) in economics. In addition, it is critical that IT systems are set up so data can be worked on using DataSHIELD.

To finish, we reiterate that our aim in placing DataSHIELD into the public domain at this juncture is to further stimulate active discussion amongst ethico-legal experts, bioscientists, epidemiologists, biostatisticians, health services researchers, social scientists, national statistical offices and IT professionals. It is our hope that interest generated by this article will encourage others to work alongside us in exploring the opportunities presented by this remarkably simple idea. If the key challenges can be identified and met—and there is no reason to believe that they cannot—DataSHIELD can provide an invaluable addition to the growing toolkit (http://www.P3G.org) that is facilitating the large-scale pooled analyses that are fundamental to current and future progress in contemporary biomedical and social science.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported as a core element of the research programs of the Public Population Project in Genomics (P^3G) funded by Genome Canada and Genome Quebec, and Promoting Harmonization of Epidemiological Biobanks in Europe (PHOEBE) funded under European Framework 6 (LSHG-CT-2006-518418). The methodological programme at the University of Leicester focusing on genetic statistics and large-scale data harmonization and pooling is also supported by Medical Research Council Project Grant (G0601625), Wellcome Trust Supplementary Grant (086160/Z/08/A), Leverhulme Research Fellowship (RF/9/RFG/2009/0062) and the Leicester Biomedical Research Unit in Cardiovascular Science (National Institute for Health Research). M.W. is Canada Research Chair in Population Health Modeling/Populomics. N.M. is funded by a British Heart Foundation Studentship (FS/06/040), J.L. is a Canada Research Chair in Human Genome Epidemiology.

Conflict of interest: None declared.

References

[1] 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* 2009;**38:**263–73.

[2] Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. *Nat Protocols* 2007;**2:**2492–501.

[3] 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.

[4] Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004;**429:**475–77.

[5] Khoury MJ. The case for a global human genome epidemiology initiative. *Nat Genet* 2004;**36:**1027–28.

[6] Newton-Cheh C, Eijgelsheim M, Rice KM *et al*. Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat Genet* 2009;**41:**399–406.

[7] Newton-Cheh C, Johnson T, Gateva V *et al*. Eight blood pressure loci identified by genomewide association study of 34,433 people of European ancestry. *Nat Genet* 2009;**41:**666–76.

[8] Repapi E, Sayers I, Wain LV *et al*. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2009;**42:**36–44.

[9] Hindorff LA, Sethupathy P, Junkins HA *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 2009;**106:**9362–67.

[10] Burton PR, Clayton DG, Cardon LR *et al*. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* 2007;**39:**1329–37.

[11] Zeggini E, Weedon MN, Lindgren CM *et al*. Replication of genome-wide association signals in U.K. Samples reveals risk loci for type 2 diabetes. *Science* 2007;**316:**1336–39.

[12] Frayling TM, Timpson NJ, Weedon MN *et al*. *A Common Variant in the FTO Gene is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity, Science* 2007; **316:**889–94.

[13] Easton DF, Pooley KA, Dunning AM *et al*. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;**447:**1087–93.

[14] Scott LJ, Mohlke KL, Bonnycastle LL *et al*. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;**316:**1341–45.

[15] Stacey SN, Manolescu A, Sulem P *et al*. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2007; **39:**865–69.

[16] Saxena R, Voight BF, Lyssenko V *et al*. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;**316:**1331–36.

[17] Friedenreich CM. Methods for pooled analyses of epidemiologic studies. *Epidemiology* 1993;**4:**295–302.

[18] Slimani N, Deharveng G, Charrondière RU *et al*. Structure of the standardized computerized 24-h diet recall interview used as reference method in the 22 centers participating in the EPIC project. *Comp Meth Programs Biomed* 1999;**58:**251–66.

[19] Harris JR, Willemsen G, Aitlahti T *et al*. Ethical issues and GenomEUtwin. *Twin Res* 2003;**6:**455–63.

[20] Lynch J, Davey Smith G, Harper S *et al*. Is income inequality a determinant of population health? Part 1. A systematic review? *Milbank Quart* 2004;**82:**5–99.

[21] Backlund E, Rowe G, Lynch J, Wolfson M, Kaplan G, Sorlie P. Income inequality and mortality: a multi-level prospective study of 521, 248 individuals in 50 US States. *Int J Epidemiol* 2007;**36:**590–96.

[22] Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. *Nat Genet* 2009;**41:**965–67.

[23] Homer N, Szelinger S, Redman M *et al*. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008;**4:**e1000167.

[24] 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. *Statistical Science* 2005;**20:**163–77.

[25] Karr A, Fulp W, Vera F, Young S, Lin X, Reiter J. Secure, privacy-preserving analysis of distributed databases. *Technometrics* 2007;**49:**335–45.

[26] GCNews. *Health Beats MoD on Equipment Losses*, 2008. http://www.smarthealthcare.com/equipment-losses (12 October 2009, date last accessed).

[27] P3G_Consortium, Church G, Heeney C, *et al*. Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* 2009;**5:**e1000665.

[28] Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Investigat* 2008;**118:**1590–605.

[29] Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual- and aggregate-level data. *Stat Med* 2008;**27:** 651–69.

[30] Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. 2nd edn. New York: Oxford University Press, 2000.

[31] Khoury MJ, Little J, Gwinn M, Ioannidis JP. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol* 2007;**36:**439–45.

[32] Hattersley AT, McCarthy MI. What makes a good genetic association study? *Lancet* 2005;**366:**1315–23.

[33] Kaye J. Do we need a uniform regulatory system for biobanks across Europe? *Eur J Hum Genet* 2006;**14:** 245–48.

[34] Zink A, Silman AJ. Ethical and legal constraints on data sharing between countries in multinational epidemiological studies in Europe report from a joint workshop of the European League Against Rheumatism standing committee on epidemiology with the ''AutoCure'' project. *Ann Rheum Dis* 2008;**67:**1041–43.

[35] Malfroy M, Llewelyn CA, Johnson T, Williamson LM. Using patient-identifiable data for epidemiological research. *Transf Med* 2004;**14:**275–79.

[36] Infectious Diseases Society of America. Grinding to a halt: the effects of the increasing regulatory burden on research and quality improvement efforts. *Clin Infectious Dis* 2009;**49:**328–35.

[37] Wallace S, Lazor S, Knoppers BM. Consent and population genomics: the creation of generic tools. *IRB: Ethics & Human Research* 2009;**31:**15–20.

[38] Knoppers BM, Fortier I, Legault D, Burton P. The public population project in genomics (P3G): a proof of concept? *Eur J Hum Genet* 2008;**16:**664–65.

[39] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2008.

[40] Wall L, Christensen T, Orwant J. *Programming Perl*. 3rd edn. Sebastopol: O'Reilly Media Inc., 2000.

[41] McCullagh P, Nelder J. *Generalized Linear Models*. London: Chapman and Hall, 1989.

[42] Aitkin M, Anderson D, Francis B, Hinde J. *Statistical Modelling in GLIM*. Oxford: Clarendon Press, 1989.

[43] Statistics Netherlands, Statistics Canada, Germany FSO, University of Manchester. *Glossary of Statistical Disclosure Control, Incorporated in Paper Presented at Joint UNECE/ Eurostat Work Session on Statistical Data Confidentiality*. Geneva: UNECE/EUROSTAT, 2005.

[44] ESRC_Secure_Data_Service. http://www.esrc.ac.uk/ESRC InfoCentre/research/resources/SDS.aspx. 2009 (21 June 2010, date last accessed).

[45] 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.

[46] Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; **88:**9–25.