DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective

E.M. Jones[1], N.A. Sheehan[1], N. Masca[1], S.E. Wallace[1], M.J. Murtagh[1] and P.R. Burton[1,2]

1) Department of Health Sciences, University of Leicester, UK
2) Public Population Project in Genomics (P^3G), Montreal, QC, Canada

Correspondence: Paul Burton, Department of Health Sciences, University of Leicester, Room 317 Adrian Building, University Road, Leicester LE1 7RH, United Kingdom
E-mail : pb51@le.ac.uk Telephone: +44 (0)116 229 7251 Telefax: +44 (0)116 229 7250

ABSTRACT

Very large sample sizes are required for estimating effects which are known to be small, and for addressing intricate or complex statistical questions. This is often only achievable by pooling data from multiple studies, especially in genetic epidemiology where associations between individual genetic variants and phenotypes of interest are generally weak. However, the physical pooling of experimental data across a consortium is frequently prohibited by the ethico-legal constraints that govern agreements and consents for individual studies.

Study level meta-analyses are frequently used so that data from multiple studies need not be pooled to conduct an analysis, though the resulting analysis is necessarily restricted by the available summary statistics. The idea of maintaining data security is also of importance in other areas and approaches to carrying out 'secure analyses' that do not require sharing of data from different sources have been proposed in the technometrics literature. Crucially, the algorithms for fitting certain statistical models can be manipulated so that an individual level meta-analysis can essentially be performed *without* the need for pooling individual-level data by combining particular summary statistics obtained individually from each study. DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases) is a tool to coordinate analyses of data that cannot be pooled.

In this paper, we focus on explaining why a DataSHIELD approach yields identical results to an individual level meta-analysis in the case of a generalised linear model, by simply using summary statistics from each study. It is also an efficient approach to carrying out a study level meta-analysis when this is appropriate and when the analysis can be pre-planned. We briefly comment on the IT requirements, together with the ethical and legal challenges which must be addressed.

INTRODUCTION

Very large sample sizes are required for the estimation of effects which are known to be small, and for addressing intricate or complex statistical questions. This is particularly so for genetic epidemiology studies where associations between individual genetic variants and phenotypes of interest are generally weak, and when scientific focus is often on the detection of rarer variants and the study of gene-gene and gene-environment interactions (1,2). Sample sizes that are sufficiently large for adequately powered analysis are often only achievable by pooling data from multiple studies. This has led to the development of large collaborative consortia for genome-wide association studies (GWAS) which have been responsible for many of the more recent findings in genetic research (3,4).

To achieve sufficiently large sample sizes, one would ideally prefer to conduct an individual-level meta-analysis (ILMA) by combining the data from every individual in each participating study into one large data set, and analysing this as a single study (allowing for centre-to-centre heterogeneity). However, because individual level data can be highly sensitive, and in many cases can actually disclose subject identity, the sharing of individual-level data always raises important ethical, legal and social issues (5-12). Thus, the physical pooling of experimental data across a consortium is frequently prohibited by the specific ethico-legal constraints that govern agreements and consents for individual studies. Biomedical science is especially cautious about such issues, and guidelines are frequently being revised in response to new threats to security (13,14). Thus, although scientifically preferable, a conventional ILMA is often not viable in practice.

In contrast, a study level meta-analysis (SLMA) *is* often consistent with ethico-legal restrictions and enables the effective estimation of simple gene-disease associations, such as those targeted in a genome-wide association study analyses. Here, each collaborating group in the consortium performs an analysis on its own data, and shares the resulting association statistics with an agreed analysis group. A meta-analysis is then conducted on these study-level statistics to obtain associational estimates across the whole consortium. Crucially, raw data are not shared at any point between the different studies. An SLMA is an ideal solution to data sharing for simple analyses, for which an analysis can be pre-planned. But, for more complicated investigations, an exploratory investigation will often be required and so statistical analyses cannot be specified so easily in advance. In these cases, an SLMA can become unwieldy and restrictive as it depends on the summary statistics already available from the chosen studies. Crucially, important new research questions cannot be addressed until additional information is agreed upon and then extracted from the original studies.

DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonised Individual levEL Databases) has been proposed as a method which yields identical results to an ILMA for a particular – though broad – class of analyses, generalised linear models, but without the need for pooling individual-level data (11). DataSHIELD exploits modern distributed computing methods to perform a parallelised analysis using a remote access analysis server (5). Central to this idea is the fact that all the data from any individual study in a consortium remain on their own local 'data computer' (DC), and cannot be accessed or even seen by any of the other studies or by the statistician coordinating the analysis. A 'central' node, which could be one of the DCs or an additional computer, is designated as the 'analysis computer' (AC). The AC coordinates the analysis on all DCs simultaneously, by transmitting appropriate blocks of code to each DC. The code contains instructions on what analysis should be executed, and which low-dimensional summary statistics should be sent back to the AC. The fundamental idea is that these summary statistics can be manipulated independently of the local computers to facilitate a combined analysis. The DataSHIELD approach can be implemented using appropriate pre-specified statistical software, such as R, on each DC. A DataSHIELD analysis satisfies the strict requirement that individual-level data are not shared and, in this regard, may be viewed as ethico-legally equivalent to an SLMA. DataSHIELD thereby offers the potential to resolve tensions due to the conflicting goals of scientific progress and participant confidentiality whilst allowing the flexibility to ask new questions of the data (11).

The central concept of DataSHIELD, notwithstanding the various statistical and technical issues, is simple and is related to an approach that has been introduced in the technometrics literature (5,7,15). The approach described in (7) entails an iterative round-robin mechanism used to compute, for example, the relevant statistics for what is referred to as "secure summation". This is initialised by a study chosen at random, study A, which generates the statistic of interest, S_A, from its own data. To obscure this value, so that the owners of the next study in the sequence (study B) cannot infer anything about it, a large random number is generated, M, known only to the owners of study A, and ($S_A + M$) is passed on to study B. Study B then adds the value of its own statistic, S_B, to ($S_A + M$) and passes this on the next study, and so on. This process continues until, eventually, the overall sum returns to study A and the random number M is subtracted to give the correct grand-total. This approach can be used to carry out a secure linear regression (7).

DataSHIELD supports an efficient and rapid implementation of more sophisticated and interactive analyses than does the round-robin approach in (7); not least because the latter requires a higher level of cooperation among the participating groups, which are all actively involved in the statistical analysis. For example, because each group must command the fitting of each model itself, there must be prior agreement on the form that models of increasing complexity will take. However, the potential drawback of using a central server with independent client computers (7,11), is that certain calls, or combinations of calls, from the AC to the DCs could disclose sensitive individual-level information. Predictably, calls of this nature may breach ethical and legal guidelines and therefore cannot be amongst the 'allowable' or 'permitted' operations. Much of the challenge of developing DataSHIELD lies in identifying potential disclosure and designing an IT infrastructure which can protect against it (5,7,11). This is under active development but falls outside the scope of this present paper. The focus of this paper is rather on the statistical underpinning of a DataSHIELD analysis. Specifically, we explain why DataSHIELD works for the class of generalised linear models using the iterative reweighted least squares (IRLS) algorithm for parameter estimation (16,17). In particular, we detail mathematically why physical pooling of the data from the independent studies is not necessary, and why the DataSHIELD approach mimics an ILMA. An example of a DataSHIELD-type analysis is then provided for a problem based on logistic regression. We discuss the future application of the method to new classes of analysis and comment on some of the perceived statistical challenges.

ILMA VIA DATASHIELD

Any mathematically permissible SLMA can be performed very efficiently using DataSHIELD and is more straightforward to carry out than a traditional SLMA as it can be fully automated. In this case, the summary statistics required from the parallel analyses are simply the final analytic results from each centre which are transmitted once from each study to the analysis centre. However, the potential of DataSHIELD extends much further, and can in theory be used to fit a generalised linear model (GLM), yielding *identical* results to those obtained from an ILMA but *without* pooling individual-level data. This is made possible via a trivial modification of the iterated reweighted least squares algorithm (IRLS) (16,17), which iteratively requests non-identifying summary statistics from each study, and combines these appropriately to estimate the parameters of the GLM. The mathematics behind this idea is straightforward, but it appears not to have been exploited, thus far, by biomedical researchers. We will now illustrate how this aspect of DataSHIELD ‘works’ to provide *identical* results to an ILMA analysis in a standard GLM setting (with fixed effects). We focus on the situation where data are *horizontally partitioned*, that is, where the collaborating studies have all measured the same outcome and covariates on all their study participants and that information relating to any given individual can be found in one, and only one, study.

Fitting a generalised linear model using DataSHIELD

First, consider the case of a single study where *N* independent observations are collected on a dependent variable *Y* and on a set of q covariates for each individual, summarised by the matrix $\boldsymbol{X^T} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$, with $\boldsymbol{x}_i^T = (x_{i1}, \dots, x_{iq})$ representing the q-dimensional vector of covariates for individual i. Assume that the relationship between Y and X for individual i can be summarised by the Generalised Linear Model (GLM)

$$\eta_i := g(\mu_i) = \boldsymbol{\beta^T x}_i,$$

(Identity 1)

where, as is consistent with standard notation (17), η_i is the linear predictor, g is the link function specified by the researcher, μ_i is the mean of Y_i with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$, and $\beta^T = (\beta_1, \dots, \beta_q)$ is the parameter vector we wish to estimate.

From the definition of a GLM, the random variable *Y* must be drawn from a distribution parameterised by θ and ϕ (the latter being a dispersion parameter) which belongs to the exponential family (*e.g.* Gaussian, binomial, Poisson), so that its probability density function, *f*, is of the form

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - c(\theta_i)}{\phi} + h(y_i, \phi)\right).$$

(Identity 2)

It can be shown (17) that

$$\mu_i := E[Y_i] = \frac{dc(\theta_i)}{d\theta_i}\ ;\ Var[Y_i] = \phi\frac{d^2c(\theta_i)}{d{\theta_i}^2} := \phi V_i\ .$$

(Identity 3)

For example, for a logistic regression in which $\mu = p$ (the proportion of positive responses),

$$\eta_i = g(p_i) = \log\frac{p_i}{1-p_i}\ ; c(\theta_i) = -\log(1-p_i)\ ;\ \phi = 1$$

so that

$$\frac{dg(\mu_i)}{d\mu_i} := g'(p_i) = \frac{1}{p_i(1-p_i)}\ ;\quad V_i = p_i(1-p_i)$$

(Identity 4)

The IRLS algorithm

The IRLS algorithm (16,17) is a method for finding approximate maximum likelihood estimates for $\boldsymbol{\beta}$. This is an iterative procedure, closely allied to the Newton Raphson method (17). We denote by $\boldsymbol{\beta}_t$ the vector $\boldsymbol{\beta}$ at the t^{th} iteration. The IRLS algorithm derives $\boldsymbol{\beta}_{t+1}$ via the identity:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{I}(\boldsymbol{\beta}_t)^{-\mathbf{1}}\boldsymbol{s}(\boldsymbol{\beta}_t)$$

(Identity 5)

where $\boldsymbol{I}$ is the expected information matrix and $\boldsymbol{s}$ is the score function. The quantities $\boldsymbol{I}$ and $\boldsymbol{s}$ are defined as

$$\boldsymbol{I}(\boldsymbol{\beta}_t) = \boldsymbol{X^T W}_t\boldsymbol{X};$$

$$\boldsymbol{s}(\boldsymbol{\beta}_t) = \boldsymbol{X^T W}_t\big(\boldsymbol{Y} - \boldsymbol{\mu}(t)\big)g'\big(\boldsymbol{\mu}(t)\big),$$

where $\boldsymbol{W}_t$ is a diagonal matrix with diagonal entries:

$$w_{ii}(t)^{-1} = V_i(t)g'(\mu_i(t))^2$$

(Identity 6)

relating to the i^{th} individual, i=1,...,N, and

$$g'\big(\mu_i(t)\big) = dg\big(\mu_i(t)\big)/d\,\mu_i(t)$$

The notation $\mu(t)$ is used since the quantity μ depends on $\boldsymbol{\beta}$, and hence on *t*. The process of updating $\boldsymbol{\beta}$ is repeated until the estimates ‘converge’ in the sense that some pre-established convergence criterion is met.

To see that this procedure can still be used when the data derive from different studies and are not pooled into one large study of size *N*, note first that we can write $\boldsymbol{I}$ and $\boldsymbol{s}$ in terms of vector summations across all *N* observational units (*i.e.* the *N* subjects in the pooled analysis):

$$\boldsymbol{I}(\boldsymbol{\beta}_t) = \sum_{i=1}^{N} w_{ii}(t)\,\boldsymbol{x}_i\boldsymbol{x}_i^T,$$

and

$$\boldsymbol{s}(\boldsymbol{\beta}_t) = \sum_{i=1}^{N}(y_i - \mu_i(t))g'\big(\mu_i(t)\big)w_{ii}(t)\boldsymbol{x}_i.$$

Next, note the following characteristics of the quantities $\mu_i(t)$, $g'\big(\mu_i(t)\big)$ and $w_{ii}(t)$:

1. We can estimate $g(\mu_i(t))$ via identity (1), since $\boldsymbol{\beta}_t$ is known at iteration (*t*+1) from the previous iteration.
2. Using this, it is then straightforward to estimate $\mu_i(t)$ via the known inverse link function $g^{-1}()$; *e.g.* for logistic regression, $\mu_i(t) := p_i(t) = exp(\boldsymbol{\beta_t}^T\boldsymbol{x}_i)$/(1 + $exp(\boldsymbol{\beta_t}^T\boldsymbol{x}_i)$), which is the expected probability of a positive response at iteration *t*.
3. Since we know the general form of the function g, along with the value of $\mu_i(t)$, it is possible to calculate $g'\big(\mu_i(t)\big)$. As an illustration, identity (4) denotes the general form of $g'\big(\mu_i(t)\big)$:= $g'\big(p_i(t)\big)$ in the logistic case, with $g'\big(p_i(t)\big)$ =1/$p_i(t)(1 - p_i(t))$.
4. Since the form of the probability density function of Y is known, and therefore the function $c()$ is also known (identity (2)), we can estimate $V_i(t) = d^2c(\theta_i)/d{\theta_i}^2$ for iteration t (identity (3)); in the particular case of logistic regression, identity (4) denotes the relevant function, with $V_i(t) = p_i(t)(1 - p_i(t))$.

5. Finally, $w_{ii}(t)$ can be calculated using $g'(\mu_i(t))$ and $V_i(t)$ via identity (6).

Calculating $\boldsymbol{I}(\boldsymbol{\beta}_t)$ and $\boldsymbol{s}(\boldsymbol{\beta}_t)$ is now straightforward and $\boldsymbol{\beta}_{t+1}$ can therefore be derived via identity (5), concluding the current iteration and enabling parameter refinement to progress. It should be noted that the vector summation notation (above) implies that $\mu_i(t)$, $g'(\mu_i(t))$ and $w_{ii}(t)$ may ultimately be calculated for each *individual* within each study, using only the highlighted identities and the observed data for that particular individual, $\{y_i, \boldsymbol{x_i}\}$. Hence, if the N individuals are split into S sub-studies with N_j observations in sub-study $j = 1, \dots, S$, the expected information matrix $\boldsymbol{I}$ and the score function $\boldsymbol{s}$ can still be extracted. In particular,

$$\boldsymbol{I}(\boldsymbol{\beta}_t) = \sum_{j=1}^{S}\sum_{i=1}^{N_j} w_{ii}(t)\,\boldsymbol{x}_i\boldsymbol{x}_i^T = \sum_{j=1}^{S} \boldsymbol{I}_j(\boldsymbol{\beta}_t),$$

and

$$\boldsymbol{s}(\boldsymbol{\beta}_t) = \sum_{j=1}^{S}\sum_{i=1}^{N_j} (y_i - \mu_i(t))g'(\mu_i(t))w_{ii}(t)\boldsymbol{x_i} = \sum_{j=1}^{S} \boldsymbol{s}_j(\boldsymbol{\beta}_t),$$

without the need to pool the data across all the studies. The IRLS algorithm therefore still works, and, in essence, we can mimic an individual level analysis without sharing individual level data at all. Of course, the data in each study must exist in the same format with the same scale for the outcome variable and covariates for this to work. But this is an essential requirement for almost any joint analysis anyway; the contributing studies must be harmonised in some sense. Where the data structure differs between biomedical studies but is amenable to active retrospective harmonization, the need for harmonization can often be fulfilled using an approach such as the DataSHaPER (18,19). It is also desirable that the same version of analytic software is installed on each DC, as different versions of the same package may use different algorithms, which could prevent or bias the analysis.

An illustration using logistic regression

To demonstrate the idea behind DataSHIELD, we give a simple example of a logistic regression analysis. The simulated data are based loosely on the outcome *myocardial infarction* (binary outcome), with systolic blood pressure X_1, body mass index X_2, and a genetic variant reflected in a single nucleotide polymorphism (SNP) genotype X_3 as covariates. If $N(\alpha, \gamma^2)$ denotes a normal (Gaussian) distribution with a mean of α and a standard deviation of γ, we assume that:

$$X_1 \sim N(120, 7^2);$$
$$X_2 \sim N(25, 3^2).$$

In addition, if $Bin(2, \pi)$ is the distribution of positive responses (0, 1 or 2) in a sample size of two independent observations, where π is the probability that each observation will be positive, we assume that the genotype is distributed as:

$$X_3 \sim Bin(2,0.3),$$

this implying that if the minor (rare) allele is denoted g and the common allele G, the minor allele frequency is 0.3 and the respective probabilities of the three SNP genotypes GG, Gg and gg are: $(1–0.3)^2 = 0.49$; $2(0.3)(1–0.3) = 0.42$; and $0.3^2 = 0.09$.

The occurrence of a myocardial infarction for individual i, Y_i, occurs with probability p_i, and is related to the covariates via

$$\text{logit}(p) = –0.5 + 0.1X_1 + 0.05X_2 – 0.3X_3$$

where the function 'logit' is defined as $\text{logit}(p) = \log(p/(1 - p))$.

The DataSHIELD approach is illustrated by simulating data for six hypothetical horizontally partitioned case-control studies from the above model with case-control ratios as detailed in Table 1.

Table 1. Six hypothetical case-control studies for the myocardial infarction example with horizontally partitioned data.

Study	Total observations	Cases	Controls
1	200	79	121
2	2000	701	1299
3	700	254	446
4	600	215	385
5	1000	382	618
6	100	38	62

An ILMA based on the pooled data from all studies combined was compared with a DataSHIELD-type analysis (11) based on the six studies separately. Both analyses were performed using the statistical package R (20). Table 2 summarises the outcome of each analysis. As can be seen, and as would be anticipated given the theory outlined earlier, the outcomes are not just similar, they are identical. Both approaches required four iterations to achieve 'convergence', with the convergence criteria for both analyses set at

$$\frac{|D_r - D_{r-1}|}{D_r + 0.1} < 10^{-8}$$

(the default convergence criterion of the glm() function in R (20)), where $D_r = –2\log L_r$, and L_r is the likelihood at the r^{th} iteration.

Comments

If one were to estimate the log-odds ratio for each of the variables (X_1, X_2, X_3) via an SLMA of the six studies in section 2.1, the estimates, along with their associated 95% confidence intervals, would be very similar to those from a GLM ILMA/ DataSHIELD analysis (11). In other words, the advantage of DataSHIELD is not that it provides inferences that are either more powe-

Table 2. Estimates of the coefficients from each analysis.

		ILMA using GLM		DataSHIELD using GLM	
Intercept	True values	Estimate	Std. error	Estimate	Std. error
	0.5	–0.442	0.043	–0.442	0.043
X_1	0.1	0.098	0.005	0.098	0.005
X_2	0.05	0.048	0.011	0.048	0.011
X_3	–0.3	–0.312	0.051	–0.312	0.051

rful or less biased than SLMA – though SLMA cannot, of course, be used at all for analyses that demand access to individual patient records. Rather, DataSHIELD provides a more flexible and less time consuming approach to analysis, as it is not restricted to the summary statistics that have been made available by each study ahead of the analysis. In principle, new research questions can therefore be explored in real time, limited only by the particular classes of analysis that DataSHIELD can perform, and by the security features required to prevent the release of participant-identifying summary statistics.

It should be noted that the constraint imposed by the need to pre-define summary statistics is often of limited impact in certain areas of biostatistics, for example, in meta-analysing clinical trials, or undertaking a genome wide association analysis. In both of these settings, the range of potential analyses is limited and best practice would often suggest that an analysis plan can, and *should*, be specified before the meta-analysis starts. But this is *not* the case when analysis is aimed at investigating genes, environment and interactions between them. Exploratory analysis is then essential, and should ideally be undertaken on the full data set combined – a restrictive need to limit investigation to models that have been specified *a-priori* will often be extremely inefficient from a scientific perspective.

Disclosure risks

DataSHIELD is currently under development and piloting. In order for it to become useful as a practical analysis tool, it must satisfy the specific governance requirements of each study to which it is applied. In particular, it is critical that there is appropriate protection of the privacy and confidentiality of individual study participants. This is not a trivial challenge, and in any case, it is never possible to design a totally secure system. Security breaches could for example arise if the IT system underpinning DataSHIELD was insufficiently secure, or if the people operating it behaved in a cavalier manner or with deliberate mal-intent. But despite their undoubted importance, these issues fall outside the remit of this particular paper. The first is a computer science concern and has to do with the general problems surrounding the set-up of remote access analysis servers and is discussed in detail elsewhere (5,7,15). The second relates to the ethical and legal concerns that necessarily arise from any collaboration involving the sharing of data. Both of these issues are being actively addressed under the DataSHIELD project, but neither will be considered further here. There are, however, a range of other ways in which supposedly secure information can inadvertently be transmitted which relate to the nature of the transmitted data themselves and to statistical inferences that can be based upon them. These issues *are* central to this paper.

First, individual data records may themselves be revealing, either directly, through identifying labels such as names, social security numbers, addresses ('identity disclosure' (5)) or indirectly, by releasing sensitive information such as rare disease status etc. ('attribute disclosure' (5)). DataSHIELD *does not* provide a way of by-passing the need to ensure that potentially identifying items such as these are not released by a study. For example, in developing the suite of secure 'aggregating' functions that are used to explore the overall distribution of data items across individual studies, one of the issues that is being considered is the minimum cell size in any table that may acceptably be viewed by the AC – *e.g.* should all cells containing fewer than 5 individuals be collapsed with other local cells. The IT system being developed aims to enforce this capacity.

Second, certain combinations of records or summary statistics, whilst individually innocuous, can lead to 'inferential disclosure' whereby the privacy of study participants is threatened, not by their own records, but by inference based on the statistical distribution over the whole database (5). Perhaps the simplest example of this is based on what is sometimes called 'Residual Disclosure'. For example, one might start by requesting a summary statistic based on all those participants in a particular study who were born *on or before* 12/6/1959, followed by a request for the same summary statistic on all of those born strictly *before* 12/6/1959. The difference between these two summary statistics will reveal information specifically about those born on 12/6/1959. Crucially, the size of that group may be very small – possibly a single individual – and this can pose an obvious threat to the privacy. For example, if only one participant has the specified date of birth, and the summary statistic in question is the prevalence of bipolar depression, one can then precisely infer whether that individual does or does not suffer from a condition they may well prefer to keep confidential. Of course, additional information, not available through DataSHIELD, or prior knowledge – for example someone's date of birth – would be necessary to then identify this individual.

To address the specific concern of residual disclosure, strict restrictions are at present placed on the potential for analysing subsets of data. But, in the future, a more flexible approach may involve ongoing parsing of each new analytic request and its comparison with earlier commands in order to place restrictions on the 'query' and 'answer' spaces, and implement risk and utility measures for each query (5). More generally, since DataSHIELD depends on the sharing of summary statistics to fit models, it is fundamental that we consider whether any violations of privacy rules are likely to occur, either directly from a single summary statistic, or via a combination of such measures. It has long been known that simple linear regression can be considered 'secure' (5,7,15) and this has now been extended to the broader class of GLMs via DataSHIELD (11). In this case, the summary statistics, either individually or in combination, that pass from the DC to the AC are unlikely to disclose sensitive information when the number of study subjects is large in comparison to the number of parameters in the GLM (11) – as is generally the case. This, however, may not be the case with more complex models of interest, and ensuring that privacy regulations are not violated via the exchange of summary statistics necessary to fit these models will require considerable effort. For example, fitting Generalised Linear Mixed Models (GLMM) (21) and Generalised Estimating Equations (GEE) (22) are of particular interest, as biobank data may be correlated. But, these pose new problems: for example, GLMMs typically rely on estimating potentially identifying subject-specific random effects to fit a model, and so the DataSHIELD algorithm would require that these be handled very carefully, and perhaps even retained on the DCs. One possibility is that DCs would send a summary statistic, such as the local variance of the random effects, to the AC rather than the individual values themselves.

Discussion

DataSHIELD offers an alternative approach to the analysis of large datasets obtained by combining several studies but where the physical sharing of individual data is not possible or is not permitted (11). Crucially, DataSHIELD enables one to mirror an analysis based on pooling individual data between several studies but *without* physically sharing those data. As a pivotal characteristic of the proposed approach, the present paper demonstrates why, for a broad class of analyses including those that require the fitting of a GLM, this mirroring is mathematically perfect. Investigation into the required statistics for other statistical models and more complex analyses is currently underway.

For the future, a number of important methodological challenges and potential applications are already evident.

In the context of contemporary developments in epidemiology, public health, and population genomics (23,24), a potentially important application of DataSHIELD is in *Mendelian randomisation* analysis (25,26). Here, the focus is on the use of a known genetic variant as an *instrument* to circumvent the problem of unobserved confounding of some exposure-outcome association of interest and not on the aetiological role of the genetic variant itself (25,27). Because the crucial gene-exposure association is often weak, vast sample sizes can be required for reliable effect estimates and some alleviation from the inherent problem of 'weak instrument bias' in these applications (26,28-30). This implies that meta-analysis will often be invaluable, and as methods for causal inference in this setting sometimes *demand* access to individual level data, ILMA will often be essential and if physical access to the individual level data is problematic, DataSHIELD may provide a viable solution.

Fitting models, however, is just one part of a thorough statistical analysis, and ideally DataSHIELD would also provide tools for the diagnostic checking of fitted models. Since such checks are commonly conducted using potentially identifying residuals, other ways of model checking must be used. Some basic checks for departures from the model are theoretically viable, however. For simple linear regression models, there are a number of statistics that can be derived via summation (15). This is more complex for a GLM, though it is possible to derive the diagonal elements, h_{ii}, of the 'hat matrix' H associated with any GLM via a DataSHIELD approach, since h_{ii} depends only on the covariate values of individual i and the matrix $\boldsymbol{I}(\boldsymbol{\beta})$, calculated during the fitting of the GLM via the IRLS algorithm (17). This means that it is possible, for example, to securely calculate the 'measure of leverage' (16), defined as the trace of the hat matrix, via DataSHIELD. However, such measures are hardly sufficient for model checking. In consequence, we are also investigating the use of contour or density 'maps' in place of residual plots – thereby avoiding the need to show individual plotting points. These could be produced for each study, and sent to the AC without compromising data privacy legislation, and combined to create an overall residual contour plot for the model as a whole. Patterns in this contour plot could then be used for detecting a poor fit of the model to the data. A useful alternative may also be provided by 'synthetic diagnostics' which entails simulating residuals along with dependent and independent variables that mimic their true counterparts. These are often sufficient to detect gross deficiencies in a model (31).

Thus far, and as discussed above, the use and development of DataSHIELD has focussed primarily on *horizontally partitioned* data where each study has measured the same attributes on different participants. Another situation where an approach such as DataSHIELD could be enormously useful is that of *vertically partitioned data*. Here information on the covariates on any specific individual may be held in different datasets (32). As an example, a particular cohort study may wish to undertake an analysis which incorporates data on individual study participants that can only be garnered from a remote (and potentially highly secure) data repository such as a Justice Department database. In this case, one cannot even begin without some element of data sharing as the different studies need to determine which subjects they all have in common and how to deal with incomplete records (7) – that is, some way to 'link' individuals in the cohort study with their corresponding data, if any such exist, in the secondary database. Presuming that this first hurdle can effectively be overcome, some adaptation of the 'secure matrix products' approach to linear regression on vertically partitioned data (7,32) by which off-diagonal blocks of the sample covariance can be computed with a limited degree of sharing is theoretically possible for other analyses. Furthermore, it is of course conceivable that many databases in the real world will be *"partially overlapping, vertically partitioned"* (32) a mixture of both horizontally and vertically partitioned data. This poses an even greater challenge, but it is discussed in more detail in Reiter et al (33).

It is, of course, possible that statistics hitherto regarded as 'safe' may unexpectedly be demonstrated to be 'unsafe'. A problem of this nature recently occurred in the GWAS context. In that field, meta-analyses are frequently performed using study-specific summary SNP-based statistics. Aggregate results from these, such as allele frequencies, were routinely published – without restriction – on the web. In 2008, a statistical test was proposed by Homer et al (34) making it possible to determine with high power whether an individual of interest, or proband, participated in a given genetic study, using only the summary allele frequencies of that study together with a full genome-scan profile of the proband. This caused enormous concern resulting in the withdrawal of most aggregate data from the internet with access restricted to approved researchers only (14). The implications for the development philosophy of DataSHIELD are that an apparently secure summary statistic used by DataSHIELD could at some indeterminate point in the future be demonstrated to be disclosing. It is therefore critical that the summary statistics needed by any new class of model are scrutinised carefully to ensure that their potential for disclosure is properly understood. In addition, users will be required to notify DataSHIELD if any insecurities are identified in which case all possible steps will be taken to block those loop-holes. Perhaps most importantly, all project members are forewarned by the Homer *et al* example that the unexpected might just happen. DataSHIELD must thus be under continuous development and be prepared to react to change.

DataSHIELD presents only one method of protecting the confidentiality of participant data, though others primarily focus on ensuring secure access to one repository (35). These methods must all be used strategically in a combined approach that ensures that *bona fide* researchers can rapidly access the data they need (as participants assume will be the case) whilst simultaneously making sure that confidential data remain secure and that all governance stipulations are satisfied in full. DataSHIELD has the potential to provide a key component of the armoury that is required, though it certainly cannot solve *all* problems. It is being designed and developed by an international consortium that is truly transdisciplinary (12,36) – including bioinformaticians, biostatisticians, epidemiologists, social scientists and ethico-legal experts – and the aim is to ensure that potential problems, challenges and opportunities are identified early and dealt with robustly. A number of pilot/development projects have been identified, are currently being set up, and are aimed at ensuring that DataSHIELD moves forwards both efficiently and effectively.

Acknowledgements

The work described in this manuscript is a core element of the research programs of: the BioSHaRE-EU project (HEALTH-F4-2010-261433) funded by the European Commission under the Seventh Framework Program (FP7/2007-2013); P3G (the Public Population Project in Genomics) funded as an International Consortium Initiative by Genome Canada and Genome Quebec; and PHOEBE (Promoting Harmonization of Epidemiological Biobanks in Europe) funded by the European Commission's Sixth Framework Program (LSHG-CT-2006-518418). The methodological program at the University of Leicester focusing on genetic statistics and large scale data access, harmonization and pooling is also supported by: MRC Project Grant (G0601625); a Joint Funded Grant from the Wellcome Trust (WT095219MA) and MRC (G1001799); Leverhulme Research Fellowship (RF/9/RFG/2009/0062); and the Leicester Biomedical Research Unit in Cardiovascular Science (NIHR).

References

1. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009; **38** (1): 263-73.
2. Khoury MJ. The case for a global human genome epidemiology initiative. *Nat Genet* 2004; **36** (10): 1027-8.
3. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009; **41**: 666-76.

4. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, et al. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2009; **42** (1): 36-44.
5. 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. *Statist Sci* 2005; **20** (2): 163-77.
6. Foster MW, Sharp RR. Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nature Rev Genet* 2007; **8** (8): 633-9.
7. Karr A, Fulp W, Vera F, Young S, Lin X, Reiter J. Secure, privacy-preserving analysis of distributed databases. *Technometrics* 2007; **49** (3): 335-45.
8. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics – re-shaping scientific practice. *Nature Rev Genet* 2009; **10** (5): 331-5.
9. Knoppers BM, Chadwick R. Human genetic research: emerging trends in ethics. *Nature Rev Genet* 2005; **6** (1): 75-9.
10. Hoeyer K. The ethics of research biobanking: a critical review of the literature. *Biotechnol Genet Eng Rev* 2008; **25**: 429-52.
11. Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, 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** (5): 1372-82.
12. Murtagh MJ, Thorisson GA, Wallace S, Kaye J, Demir I, Fortier I, et al. Navigating the perfect [data] storm. *Norsk Epidemiologi* 2012; **21**: 203-209.
13. P3G_Consortium, Church G, Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J, Bobrow M, Weir B. Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* 2009; **5** (10): e1000665.
14. Couzin J. Genetic privacy. Whole-genome data not anonymous, challenging assumptions. *Science* 2008; **321** (5894): 1278.
15. Karr A, Lin X, Sanil AP, Reiter J. Secure regression on distributed datasets. *J Comput Graph Stat* 2005; **14** (2): 263-79.
16. McCullagh P, Nelder J. *Generalized linear models*. London: Chapman and Hall, 1989.
17. Aitkin M, Anderson D, Francis B, Hinde J. *Statistical Modelling in GLIM*. Oxford: Clarendon Press, 1989.
18. Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010; **39** (5): 1383-93.
19. Fortier I, Doiron D, Little J, Ferretti V, L'Heureux F, Stolk RP, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol* 2011; **40**: 1314-28.
20. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2008.
21. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; **88** (421): 9-25.
22. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**: 1049-60.
23. Khoury MJ, Gwinn M, Bradley L, Little J, Higgins JP, Ioannidis JP. *Human Genome Epidemiology*, 2nd edn. New York: Oxford University Press, 2009.
24. Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 2005; **366** (9495): 1484-98.
25. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003; **32** (1): 1-22.
26. Sheehan NA, Didelez V. Commentary: Can 'many weak' instruments ever be 'strong'? *Int J Epidemiol* 2011; **40**: 752-4.
27. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Meth Med Res* 2007; **16** (4): 309-30.
28. Bound J, Jaeger D, Baker R. Problems with instrumental variable estimation when the correlation between the instruments and the exogenous variable is weak. *J Am Stat Assoc* 1995; **90**: 443-50.
29. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol* 2011; **40**: 755-64.
30. Keavney B. C reactive protein and the risk of cardiovascular disease. *BMJ* 2011; **342**: d144.
31. Reiter JP. Model diagnostics for remote access regression servers. *Stat Comput* 2003; **13** (4): 371-80-80.
32. Karr A, Lin X, Reiter J, Sanil A. Privacy preserving analysis of vertically partitioned data using secure matrix products. *J Offic Stat* 2009; **25**: 125-38.
33. Reiter J, Kohnen C, Karr A, Lin X, Sanil A. Secure regression for vertically partitioned overlapping data. Technical report no 146 (2004). Available at: http://wwwnissorg/sites/default/files/pdfs/technicalreports/tr146pdf 2004.

34. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008; **4** (8): e1000167.
35. ESRC Secure Data Service. Available at: http: //www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/data-services/sds.aspx. 2011.
36. Murtagh M, Demir I, Harris J, Burton P. Realizing the promise of population biobanks: a new model for translation. *Hum Genet* 2011; **130** (3): 333-45.