Running an analysis of combined data when the individual records cannot be combined: practical issues in secure computation

Gillian M Raab*, Chris Dibben* Paul Burton **

* Administrative Data Research Centre, University of Edinburgh

** University of Bristol

Abstract. In social or epidemiological research comparable data are often collected by agencies in different settings, e.g. in different countries or by different organisations. Disclosure concerns may prevent the agencies releasing their data to outside users. Comparison of results between the different agencies may be carried out by running separate analyses in the safe haven provided by each agency and comparing the published reports. This approach has several disadvantages. One can never be sure that the data sets and variables, which are nominally the same, are really comparable. An analysis that adjusts for covariates in each individual agency will not be identical to what one would obtain if the raw data were pooled. Tests for agency-by-covariate interactions are not readily carried out from published reports.

A similar situation has arisen in the analysis of genomic data, where a pooled analysis of small individual studies is required for adequate inference, but the individual centres do not wish to share their data. The DataSHIELD system was developed in response to this (see `www.datashield.org`) and implements a joint analysis by linking the computer in each centre to an analysis computer (AC). The AC holds no raw data, but receives summary statistics from each of the individual studies, combines them, and passes the combined summaries back to the individual centres. This allows joint analyses such as generalised linear models (GLMs) (McCulloch and Nelder, 1989) to be fitted by iterating this exchange of summary statistics. The interface between the AC and the other centres prevents any raw data being exchanged.

When disclosure concerns would not allow centre computers to be linked in this way it is possible to adapt this procedure by exchanging summaries between agencies by email. Routines in R have been developed to allow such analyses to be carried out via the E-DataSHIELD protocol. This paper describes the development of the *eds* package to carry out these analyses and presents an example of its use.

1 Introduction

1.1 background

When research projects based on different populations collect equivalent information a joint analysis that combines and compares the results from individual studies may often be desirable. The three Longitudinal Studies (LSs) in the UK, each held by a different National Agency, are an example of such a group of studies. Researchers often wish to carry out analyses that combine and compare the results from more than one of the LSs. Because confidentiality and disclosure control prevent the data from individual studies from being combined into one large pooled data set, the usual approach is to carry out equivalent analyses on the individual studies and summarise the results (e.g. Popham and Boyle, 2011). While this approach can deliver results, it is time-consuming and certain aspects are less informative than would be the case for an analysis carried out on the pooled data. The situation is analogous to that of meta-analysis, where it is recognised that bringing all the data together for a pooled analysis is greatly to be preferred over combining data from published studies (Blettner *et al.*, 1999).

1.2 Multi-party Computation

Situations such as that of the LSs are often described as having data held in distributed data bases. The data may be split between databases either by cases (horizontally partitioned data), by variables (vertically partitioned data) or by a mixture of the two. The LS problem we describe above is an example of pure horizontally partitioned data. Methods for analysing such data without pooling the records have been developed by computer scientists (see Kantarcioglu, 2008 for a review) and by statisticians (e.g. Karr *et al.* 2007). Computer scientists often refer to these methods as "privacy-protecting data mining".

The methods for horizontally partitioned data rely on the fact that for, most data summaries and statistical inferences, the results for the pooled data can be derived from summary statistics which can be calculated by summing contributions from each individual study. The individuals studies compute their summary statistics which are then added together to produce the inference from the pooled data. For inferences, such as GLMs or non-linear models, that require an iterative procedure the summary statistics from each study must be summed at each stage of the iterative procedure. In some cases the individual summaries may be considered non-disclosive and thus can be exchanged freely between the different studies. When this is not the case the summary statistics must be encrypted by a method that allows the encrypted items to be summed (homomorphic encryption) before their sum can be decrypted in a secure setting and then checked and made available to all parties. We refer to this as "secure multi-party computation".

Several implementations of these methods have been described (El Emam *et al.* (2013), Jiang *et al.* (2013), Wang *et al.* (2014), Gaye,A. *et al.* (2014), Chida *et al.* (2014)). All of these systems require that the computer which is used to carry out the analyses is able to communicate with those holding the sensitive information, in a manner which prevents any access to the raw data. The DataSHIELD project (`www.datashield.org`) described in Gaye *et al.* 2014 implements the Wolfson *et al.* 2010 proposals, via open-source software. The computers on which the individual studies reside (Data Computers,

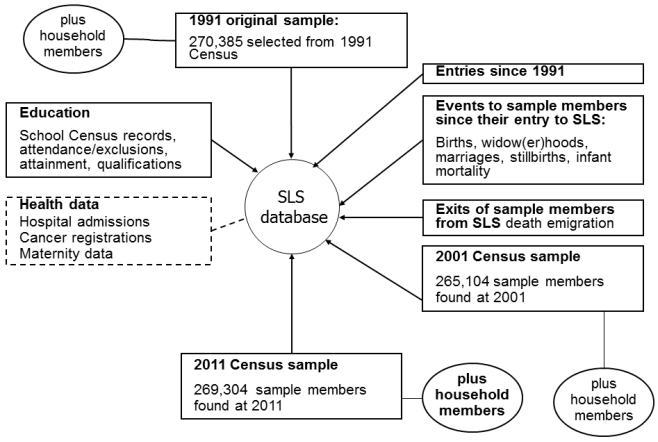

Figure 1: Structure of data sets linked into the SLS as of June 2014.

DCs) each communicate with an Analysis Computer (AC). Only non-disclosive summary statistics are passed from the DCs to the AC and the software can provide some checks on this. The DCs for individual studies never communicate with each other. This requires special software (also freely available) that restrict the types of data that can be passed between the computers, Such a setup would not be possible for data held in more secure settings, such as the LSs , where confidentiality requires that each LS is held on a secure server with no internet access. To overcome this we have developed similar methodology where non-disclosive summary statistics are exchanged between the AC and the DCs by email; we refer to this as E-DataSHIELD.

1.3 Structure of this paper

In the next section we describe the structure of the data held in the UK Longitudinal studies. Section 3 provides a brief description of the methods for multi-party computation and in particular its implimentation by the DataSHIELD project, which was the starting point for the development of E-DataSHIELD. Section 4 introduces the *eds* package for **R** and summarises the steps in an E-DataSHIELD analysis. We explain the modifications to the DataSHIELD protocol that were needed to make such a joint analysis feasible in practice. Section 4 presents an example of the use of the *eds* package on freely-available synthetic data from the ONS Longitudinal Study. The concluding section provides a discussion of future directions and of some of the limitations of the methods.

2 The UK longitudinal studies

The England and Wales Longitudinal Study (ONS LS) (Hattersley and Cresser, 1995), the Scottish Longitudinal Study (SLS) (Boyle *et al.*,2009) and the Northern Ireland Longitudinal Study (NILS)(O'Reilly *et al.*, 2009) are rich micro-datasets linking samples from the national Census in each country to administrative data for individuals and their immediate families across several decades. All of the longitudinal studies (LSs) have a similar structure. At their core are data from the UK decennial Census for the relevant country. Individuals are linked over time across Censuses and to administrative data on births, deaths, marriages, records of immigration and emigration from the relevant country and other sources. Figure 1 illustrates the data that are currently linked to the SLS, includ-

ing the Censuses held in 1991, 2001 and 2011. Negotiations are ongoing for more data sources to be linked and individual projects may add extra data. The SLS also includes detailed geographic identifiers that allow linkage to environmental data and other sources aggregated for small areas. The other LSs have somewhat different structures and links to different ranges of administrative data. The ONS LS includes data on 5 Censuses starting in 1971 while the NILS has data from the latest three Censuses[1].

The data are extremely sensitive and Census data are controlled by the Census Acts that prohibit public access for 100 years. Inclusion in the studies is by a number of "secret" birthdays spread throughout the year that are known only to a very few core staff in each study. There are four such birthdays in the ONS LS, 20 in the SLS and 104 in NILS, differing because of the different sampling fractions needed for each population size. No resident of the UK knows whether they are included in one of the LSs; this is justified ethically by the extremely secure conditions under which the data are held.

Only researchers who have undergone training in data security are permitted to have access to the data. Each of the LSs holds many thousands of variables in a series of files. No user has access to all of the data. Following an application to use a study the user specifies the data required and an extract is prepared. All analyses must to be carried out in a secure setting where rules, such as those relating to mobile phone use, are enforced by supervising staff. All outputs from an analysis are scrutinised by staff to ensure that there is no potential for disclosure of information about individuals or identifiable small subgroups and they are then emailed to a user in an encrypted form. A user's output is classified either as an **intermediate output** or as a **final output**. Intermediate outputs can only be shared with other members of the study team who have submitted their details signed confidentiality agreements. Final outputs can be published. The security requirements for final outputs (e.g. rules for the minimum cell size allowed in a table) are stricter for final outputs than for intermediate outputs.

3 Methods

3.1 Multi-party computation

The basic idea behind multi-party computation is that the final data summaries or analyses can be calculated from the set of individual summaries from each of the distributed databases. In the simplest case a mean can be calculated from the count of records in each study along with the totals for the variable of interest. For linear regression of a vector Y on a matrix of predictors X the regression coefficients for a combined analysis can be obtained by summing the terms $X'X$ and $X'Y$ from each of s studies to give the coefficients as $(\sum_{i=1}^{s} X_i'X_i)^{-1}/(\sum_{i=1}^{s} X_i'Y_i)$. For both of these examples the computer carrying out the analysis (AC) needs to obtain only a single summary from each of the computers holding the data from individual studies (DCs). In other cases, such as GLMs, where inferences are obtained by an iterative process, the AC must obtain the summaries at each iteration after supplying the current parameter estimates to each DC.

[1] see http://calls.ac.uk/ for more information about the three UK longitudinal studies and their support units

3.2 The DataSHIELD protocol.

In the DataSHIELD protocol the AC controls all the analyses using commands from the **R** package (R core team, 2015). Commands can be issued from the AC to enable the secure summaries to be obtained from each DC and then combined to provide estimates. Where simple descriptive statistics are requested, the interface between the DCs and the AC supresses certain disclosive items (e.g. small cell counts) according to user-defined settings. Models that require iterations are fitted on the AC by a single command to run all the iterations such as

```
fit1<-ds.glm(y~x1+fac2,start=c(0,.1,.3,.5),
family="binomial")
```

to run a logistic regression where y, x1 and fac2 are the names of columns in the data sets and fac2 is a factor with 3 levels.

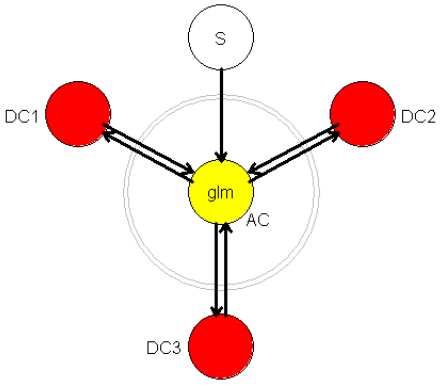

Figure 2: Datashield protocol.

All the computations are initiated by the AC running the GLM, including sending commands to each DC. No action is required at the DCs. The AC user has already set up communication with DCs, and the AC communicates with them to run the iterations of a GLM. Figure 2 illustrates the communication between the AC and the DCs for a project with three data providers. The first step of arriving at suitable starting values, illustrated by a white circle with the letter S, can be carried out by any or all of the DCs or even from prior information. It is then supplied to the AC. The double circle separating the AC and the DCs indicates the secure interface.

3.3 Secure multi-party computation

Where the summaries themselves are considered disclosive, they may be encrypted in a manner that permits their sums to be recovered (homomorphic encryption). A simple version that has been used for secure multi-party computation (Fienberg *et al.* (2006), Karr *et al.* (2007), and others) is known as secret sharing, or secure summation. First proposed by Benaloh (1987), it consists of passing the sum to be calculated around the data providers. The first data provider calculates their contribution to the sum and adds a random quantity to it. It is then passed to the next study who adds their sum to the total, and so on until it returns to the first one which subtracts the number added at the start to reveal the overall total. This is less secure than other more sophisticated forms of homomorphic encryption that have been used in some implemntations of secure multi-party computation (e.g. El Emam *et al.* 2013, Wang *et al.* 2013,), since it can be broken by collusion among the data providers.

In the DataSHIELD protocol there is no encryption of the summaries sent from the DCs to the AC. Although data such as information matrices and score statistics might seem to pose little privacy risk, it can be shown that individual data can be derived from them

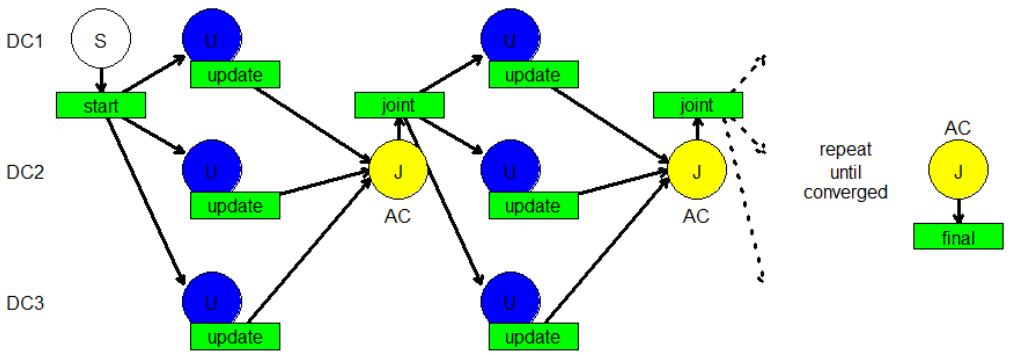

Figure 3: Schematic representation of the communications required for an iterative *eds* analysis.

(Sparks *et al.* (2008),El Emam *et al.* (2013) online suppliment). None of the encryption methods could be used with the simple communication structure for DataSHIELD illustrated in Figure 2. The lack of encryption in DataSHIELD may not be a great problem, however, since the individual DCs are not actively involved in the computations.

3.4 Developing the E-DataSHIELD (eds) package

Following a DataSHIELD workshop in 2011, it was decided to adapt the DataSHIELD protocol to enable combined analyses of the three LSs by exchanging vectors of coefficients and summary statistics by email. DataSHIELD routines are written in **R** (R code team, 2015) and the E-DataSHIELD routines were built as an **R** package. A preliminary version of the *eds* package was written and has been used for two real studies[2] that in each case combined information from two LSs. The initial version mirrored the functionality of DataSHIELD at that time, creating summary statistics that conformed to disclosure rules and fitting GLMs.

As with any statistical analysis, the bulk of the work on the projects was in data preparation and in assuring that the different data sets had exactly the same structure. The first stage of an *eds* analysis is for each data provider to run a routine that produces details of the data that will be used in the analyses. These are then exchanged between studies to check that there is agreement for all aspects of the data to be used in analyses. Once the data sources are harmonised the main analyses can be carried out.

Initial runs for fitting GLMs soon revealed some problems. Unlike DataSHIELD, the *eds* method requires the active participation of each data provider. Figure 3 represents all the transfers that are required in an iterative *eds* analysis, where the blue circles with a U indicate that the routine `eds.glm.update` is run on a DC and the yellow circles with J that `eds.joint.update` is run on the AC. The green rectangles represent the files that are emailed between computers. For GLMs, at each iteration the researchers at each DC have to

1. receive a vector of coefficients from the AC by email
2. transfer it to the secure server via a secure data stick (for some LSs this transfer can only be done by agency staff, not be the individual researcher)

[2]Both of these used Poisson regressions to compare survival between the LSs using the person-years of follow-up as an offset.

3. read it into their **R** workspace

4. run a routine to update the information matrix and score vector write them to a file

5. export the file from the secure setting via the secure data stick

6. email it to the AC.

This process was tiresome and it was difficult to keep track of what stage we were at, especially if there were delays in getting the files out of the safe settings. Several steps were taken to make it easier. It was essential to reduce the total number of iterations required. Instead of fitting only one model at a time, the routines were modified to allow a whole set of models to be fitted at once and for the iterations to stop once they have all converged. A routine was written to obtain good starting values automatically. Initially the analyst at each DC or AC had to choose a name for the file that was to be sent to the AC or DC. This was changed so that file names were defined automatically, and the files exchanged included a number of items to ensure that the correct data were being accessed at the appropriate stage of the iterations. These steps made the process much more workable.

The analysis computer could be one of the DCs, but in practice the process runs more smoothly when an external computer is used. The AC need have no access to data since it receives all of its data via emailed summaries from the DCs.

3.5 Comparing studies and including study by covariate interactions

Commonly, users of the LSs wish to check whether the relationships being investigated are homogeneous accross the LSs. This requires that a factor defining the study be included as a column in the data for each LS, and that it be defined as a factor with a number of levels corresponding to the total studies included. A routine was written to achieve this. Models including terms for studies cannot be fitted by any one DC from only their own data, so the routine for calculating starting values had to be modified. Models that excluded columns from terms involving the studies were fitted and the coefficients from omitted terms set to zero. All this is done automatically without the need for any user intervention.

3.6 Making an eds analysis more secure

Because the DCs are already involved in the analyses the use of secure computation is much easier for E-DataSHIELD than for the original DataSHIELD protocol. There is no barrier to DCs sending files by email to each other, rather than only with the AC. Moreover, security may be more important when the information has to come out of the secure settings to be transmitted by email. Thus a secure option was added to the *eds* routines. For this protocol the AC has to be the DC which holds data for `studyno = 1` . At each iteration the update at this DC adds new random numbers to every item in all the quantities exchanged (information matrices, score vectors and current combined deviances) that need to be updated by each DC. The value of the coefficients from the previous iteration is also included in the file sent. The random numbers are stored in a file

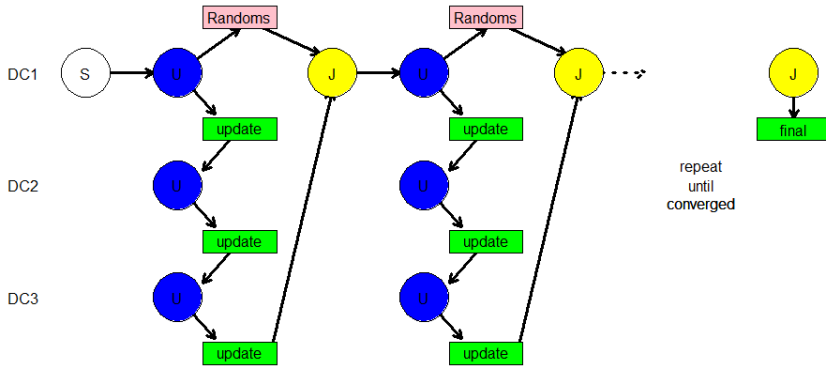

Figure 4: Schematic representation of an *eds* GLM fit using secure summation.

that is retained on this DC for the first study. The analyst at this computer then sends the file to the DC with `studyno = 2`, which adds their contribution to the sums, and so on until the last study which send it back to the DC with `studyno = 1` . Before the joint update is carried out, the object with the random numbers is retrieved and subtracted from the totals. A new set of random numbers is then added to the new quantities to be updated at the next iteration. Figure 4 illustrates the exchanges required for this protocol. When there are only two studies included, this protocol will not prevent each study learning the contribution from the other at the final iteration. But, more importantly, it will mean that each of the email files being exchanged will not by itself be disclosive of the results of a single study. This protocol will be more valuable when many small studies are being combined, as is the case for some genomic studies or those investigating the adverse effects of drugs. We intend to expand the *eds* package to incorporate secure computation for summary statistics as well as for model fitting in the near future. All these additions of random quantities are carried out automatically and transparently to the user, as is illustrated in the example in the next section.

4 Example of an *eds* analysis.

To illustrate the use of the *eds* package we use data from the synthetic spine, generated to be similar tothe data from the 2001 and 2011 Census of England and Wales held in the ONS Longitidinal Study. These data are available over the internet from the SYLLS (SYnthetic data estimations for UK LongitudinaL Studies) project at `http://calls.ac.uk/guides-resources/synthetic-ls-data/`. These data are made available as part of the *eds* package. They consist of three data sets each from a different region of England, including all women who were married or in a civil partnership at the time of the 2001 Census and who had not died or been widowed before 2011. A combined analysis is required of the factors that affect whether the women are still married in 2011. The three data sets are for the regions London (n=9047), South East England (n=13228) and South West England (n=8944). We illustrate just a very small set of three models. The variables that will be used in the example are given in the table below.

Variable	Description	Details or Categories
SynAge2001	Age group at 2001	10 year age groups from "10-19" to "70+"
Ethnic.Group	Ethnic group reported in 2001	"White" "Black" "Asian" "Mixed" "Chinese"
stillmarr	1/0 variable	Indicate that the member is still married in 2011
studyno	Region	1=London 2=SEast 3=SWest
stillmarr	1/0 variable	Indicate that the member is still married in 2011

4.1 Example without secure summation

Before the analysis each DC must run the function `eds.prepare` to add a factor with three levels containing the appropriate study number. The first stage in an *eds* analysis is to define the project via a start file. Here this is run on DC1 where the London data are held and starting values are calculated from a fit to just London, but setting terms involving `studyno` to zero. The following command defines the models we want to fit, where the symbol ¿ indicates a command submitted to **R**.

```
 >eds.make.start(London,project="spine", family="binomial",quietly=F,
+ starts="calc",models=c("study+age","Eth+age+stud","age+Ethnic*stud"),
+  formulae=c(stillmar~SynAge2001+studyno,
+  stillmar~SynAge2001+Ethnic.Group+studyno,
+ stillmar~SynAge2001+Ethnic.Group*studyno))
```

Some details of the models are printed and followed by

```
Start for project  spine  iteration  1
written to file  Start_project_spine_itno_1.R
Now email file to other computers for all to run first iteration
```

Each of the DCs then runs a command like

```
>eds.glm.update("spine",itno=1,data=SEast,quietly=T)
```

The information on the project name and iteration number (itno) tells the DC to look for the file which has just been transferred. This and the value of `studyno` tells the DC the filename to use for the result. A message reminds the user which file to send off.

```
Now email file to analysis computer
File written  Project_spine_Itno_1_Study_2.R
```

Once the AC has received all three files from the DCs the starting coefficients can be updated with the function:

```
> eds.joint.update(project="spine",itno=1,studies=1:3,quietly=T)
```

which gives the message

```
0  out of  3 models converged
File written:  Joint_fit_project_spine_Itno_2.R .
```

This file is sent to each of the DCs who repeat the updates above, but with `itno=2.` These steps are repeated until at iteration 5 we get the message:

```
3  out of  3 models converged
File written:  Joint_fit_project_spine_Itno_6.R
```

This file contains the results of the final joint fits which can be circulated to all the participating studies. Once sourced into R it produces an object called `result` which contains all the information about the fit and can be used to produce tables of results e.g.

```
> dev.table<-eds.deviance.table(result)
> dev.table
Deviance params    df dev.over.df Dev_diff df_diff Pval_diff
study+age       20911.93      9 30760   0.6798415     0.00       0        NA
Eth+age+stud    20661.25     13 30756   0.6717796   250.67       4     0.000
age+Ethnic*stud 20648.22     21 30748   0.6715306    13.03       8     0.111

> table_model2<-eds.summary(result,models=2)

Results for model  Eth+age+stud
Fitted to a  binomial  model
Formula  stillmar ~ SynAge2001 + Ethnic.Group + studyno
                       coef     se       z  pval
(Intercept)          0.6250 0.1991   3.14 0.0017
SynAge200120-29      0.3757 0.2010   1.87 0.0616
SynAge200130-39      0.7986 0.1986   4.02 0.0001
SynAge200140-49      1.2318 0.1997   6.17 0.0000
SynAge200150-59      2.1739 0.2049  10.61 0.0000
SynAge200160-69      3.0009 0.2340  12.83 0.0000
SynAge200170+        3.6384 0.3742   9.72 0.0000
Ethnic.GroupMixed   -0.4984 0.1318  -3.78 0.0002
Ethnic.GroupAsian    0.7037 0.0757   9.30 0.0000
Ethnic.GroupBlack   -0.7617 0.0750 -10.16 0.0000
Ethnic.GroupChinese -0.1180 0.1447  -0.82 0.4147
studyno2             0.2312 0.0453   5.11 0.0000
studyno3             0.2448 0.0513   4.78 0.0000
```

The results appear to show differences between ethnic groups and regions in the probability of staying married that would agree with what our prior conceptions of such differences might be. But we must remember that these data are not real, having been generated as a teaching tool, and the real data might give different results.

4.2 Example with secure summation

The code for running a secure analysis is very similar to the above, but with the parameter secure set to TRUE. The messages to the user tell them where to send the file, as shown in Figure 4. The start file is identical except for the extra parameter, `secure=T`. Updating the models at DC1 produces this output.

```
>eds.glm.update("spine",itno=1,data=London,quietly=F,secure=T)

Result read from file  Start_project_spine_Itno_1.R
Fitting  3 models for project  spine  at iteration  1
File with random objects is written as  Project_spine_Itno_1_Randoms.R

Result for project  spine  iteration  1  study  1
written to file  Project_spine_Itno_1_Study_1.R

Now email file to next data computer, study  2
```

The file is sent on to DC2 which updates the models and sends a new file to DC3 where in turn a new file is produced that is sent back to DC1. When the function `eds.joint.update` is runat the next iteration with `secure=T` the file with random objects, written at the last iteration , is automatically picked up and the values subtracted from the sums before the new coefficients are calculated. The process converges in five iterations to the same solution as above.

5 Discussion and future directions

We have shown that multi-party computation with information exchange by email is feasible in practice and that it can be made secure. The *eds* package is now being finalised and will shortly be submitted to the CRAN web site `https://cran.r-project.org/` to make it available to others.

A limitation of the method is the need for multiple exchanges by email when an iterative calculation is being carried out. Fienberg *(*et al.) (2006) have shown that, when all the data are categorical, a logistic regression can be fitted in a single iteration. This follows from the equivalence of logistic regression and log-linear models in this case because the appropriate marginal tables from form the sufficient statistics for a log-linear model. By a similar argument this should hold for a person-years analysis, fitted via a Poisson model, where all the predictors are categorical. In this case the total person-years for each margin will also be required. We hope to look into the feasibility of using this approach in *eds*. It would be particularly valuable for the LSs where most of the variables are categorical. It could not replace the current methods entirely as it would not work with numeric predictors.

Another limitation of this methodology is that no residuals are available to produce diagnostic plots, because providing residuals along with fitted values would disclose data values. Reiter (2003) has proposed a method to obtain residuals from remote-access servers by generating synthetic data and Reiter and Kohnen (2004) suggest a method of obtaining

residuals when modelling a categorical variable by partitioning the marginal distributions of the predictors. While both of these approaches might be tried, a simpler method would be to examine the fit of a model by checking whether a better fit is obtained by including additional terms, such as quadratic terms for numeric predictors or additional interactions for categorical variables.

6 Acknowledgements

We are grateful to the DataSHIELD project for having introduced us to these ideas and shared their code with us. Several other colleagues have assisted by providing feedback on the earlier version of the *eds* package: David Wright and Dermot O'Reilly from the NILS team, Fiona Cox, Kevin Ralston and Zhiqiang Feng from the SLS team and Rachel Stuchbury from the ONS LS team as well as David Walsh from the University of Stirling.

The DataSHIELD project is supported by funding from: the European Unions Seventh Framework Programme - BioSHaRE-EU (Biobank Standardisation and Harmonisation for Research Excellence in the European Union); a strategic award from MRC and Wellcome Trust for the ALSPAC project; and the Welsh and Scottish Farr Institutes, MRC funded E-Health Informatics Research Centres (EHIRCs).

References

Benaloh, J.(1987) Secret sharing homomorphisms: Keeping shares of a secret secret, In: Odlyzko, A.M. (ed.) *CRYPTO 1986. LNCS* **263**,251-260. Springer, Heidelberg.

Blettner, M,. Sauerbrei, W., Schlehofer,B., Scheuchenpfluga, T., Friedenreichd C. (1999) Traditional reviews, meta-analyses and pooled analyses in epidemiology, *International Journal of Epidemiology* **28**,1-9.

Boyle, P., Feijten, P., Feng, F., Hattersley, L., Huang, Z., Nolan, J., Raab, G.M.: (2009) Cohort profile: The Scottish Longitudinal Study (SLS). *International Journal of Epidemiology*, **38(2)**, 385–392

Chida, K., Morohashi, G., Fuji, H., Magata, F., Fujimura, A. (2014) Implementation and evaluation of an efficient secure computation system using R for healthcare statistics. *J Am Med Inform Assoc* **21** e326-e331.

El Emam, K. , Samet, S., Arbuckle, L., Tamblyn, R., Earl, C.,Kantarcioglu, M., (2013) A secure distributed logistic regression protocol for the detection of rare adverse drug events *J Am Med Inform Assoc* **20** 453-461.

Fienberg, S., Fulp, W., Slavkovic, A., Wrobel, T. (2006) Secure log-linear and logistic regression analysis of distributed databases. In: Domingo-Ferrer, J., Franconi, L. (eds.) *PSD 2006.* LNCS, **4302** 277-290. Springer, Heidelberg

Gaye, A., Marcon, Y., Isaeva, J.,*et al.* (2014) DataSHIELD: taking the analysis to the data, not the data to the analysis *International Journal of Epidemiology.* **43(6)**, 1929–1944 .

Hattersley, L., Cresser, R. (1995) The Longitudinal Study, 1971–1991: History, organisation and quality of data. *LS Series no.7*, The Stationery Office, London

Kantarcioglu M. (2008) "A Survey of Privacy-preserving Methods Across Horizontally Partitioned Data, *Privacy-Preserving Data Mining Models and Algorithms Series: Advances in Database Systems* , **34**, Aggarwal, Charu C.; Yu, Philip S. (Eds.).

Karr, A., Fulp, W., Lin, X., Reiter, J., Vera, F., Young, S. (2007), Secure, privacy preserving analysis of distributed databases. *Technometrics* **49**, 335–45.

McCullagh P, Nelder J.(, 1989) Generalized Linear Models. London: Chapman and Hall.

O'Reilly, D., Rosato, M., Catney, G., Johnston, F., Brolly, M., (2009) Cohort description: The Northern Ireland Longitudinal Study (NILS). *International Journal of Epidemiology*, **41(3)**, 634–641

Popham, F., & Boyle, P.J. (2011) Is there a Scottish effect for mortality? Prospective observational study of census linkage studies *Journal of Public Health*, **33(3)**, 453–458.

R Core Team: (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org

Reiter, J.P. (2003), Model diagnostics for remote access regression servers, *Stat. Comput.* **13** 71-380.

Reiter,J.P., Kohnen C.N.(2004) Categorical data regression diagnostics for remote servers. *Journal of Statistical Computation and Simulation* **75(11)** 889–903,

Sparks, R., Carter, C., Donnelly,, J, B. OKeefe, C. M.(2008) Remote access methods for exploratory data analysis and statistical modelling: in Privacy-Preserving Analytics *Computer methods and programs in biomedicine* **9(1)** 208-222.

Wallace SE, Gaye A, Shoush O, Burton PR.(2014) Protecting personal data in epidemiological research: DataSHIELD and UK law. Public Health Genom ;17:14957.

Wang, S.,Jiang, X.,Wua,Y., Cui, L.,Cheng, S., Ohno-Machado, L.(2013) EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed privacy-preserving online model learning *Journal of Biomedical Informatics* **46** 480-496

Wolfson, M., Wallace, S. E., Masca N.*et al.* (2010) DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data. *Int. J. Epidemiol.*, **39**, 1372–1382.

Jiang, W., Li, P., Wang, S., Wu, Y., Xue, M., Ohno-Machado, L., Jiang X. (2013) WebGLORE: a web service for Grid LOgistic REgression. *Bioinformatics* **15;29(24)**, 3238-40.