How are analyses performed in InterConnect and how is data security and privacy protected?

- Existing methods of performing **analyses** to address questions about **heterogeneity** in diet, physical activity and disease outcomes between people in different countries require research groups to physically **bring data together** in one place or follow an analysis plan and **share results.**

- **InterConnect** uses a different approach which is based on a platform called **DataSHIELD**; this avoids the challenges of existing methods for analysing data from multiple studies, such as **ethico-legal constraints** which limit researchers' ability to physically bring data together and the **analytical inflexibility** that is associated with conventional approaches to sharing results.

- The key feature of the **DataSHIELD** platform is that data from research studies **stay on a server** at each of the institutions that were originally responsible for the collection of the data. The platform allows an analyst to pass **commands** to each server and **results that do not disclose the identity of any study participants** are returned to the analyst. There is no access to the individual records within each study. Results generated in this way from each study can be combined to give overall results that are **mathematically equivalent** to having all the data pooled together in one place.

- While the InterConnect approach may seem the same as the conventional results sharing method, there are **significant advantages.** The analyst has the flexibility to **refine and rerun analyses quickly**, they can be sure that the analysis plan is **executed correctly** and there is no need for someone to conduct the analysis at each participating institution.

- The DataSHIELD platform inherently protects privacy by inhibiting any viewing, calculation, and analysis of data for an individual participant. Furthermore the **security and privacy** of data held on each server is maintained using standard web security methods. All communications between servers contain either **analysis commands** or **results that do not disclose the identity of any study participants**. Only users with appropriate permissions are able to perform analyses and participating institutions can remove access at any time.

Why is a different approach to analyses across cohorts needed?

The variation in the risk of diabetes and obesity between different countries and continents around the world is considerably greater than the variation in risk within individual countries. This population level heterogeneity in diet, physical activity and disease outcomes is largely unexplained because the existing methods for analysing data from different studies or cohorts (i.e. cross-cohort analyses) have significant limitations.

Physically bringing data together from cohort studies across the world is desirable from an analytical standpoint because it allows flexibility in the analyses that are conducted. This method is commonly referred to as individual-level meta-analysis (ILMA). However, this method is constrained by governance, ethical and legal challenges (Wallace et al, 2013):

- Governance issues can arise, for example, as data custodians become doubtful about control over the data that they have collected once it has been transferred to another site.
- Ethico-legal concerns can arise as the data on individuals within each study (so called 'individual-level data') may contain sensitive information about the individual's health, lifestyle, genotype, or sociodemographic factors that potentially can be used to identify these individuals in breach of their right to privacy.
- If the dataset is very large, the entire computational burden is placed on one location.

Thus, although scientifically preferable, a conventional ILMA is not always viable in practice.

The other approach that has been used for cross-cohort analyses is to combine analytical results that are produced by contributing studies and sent to an analytical lead or coordinating group. This method is commonly referred to as study-level meta-analysis (SLMA) since the results are being brought together at the study-level. While useful in some instances, it can be considered to be inflexible, resource-intensive and prone to mistakes:

- Only pre-planned analyses that are undertaken by all the studies can be combined to provide joint results from across all studies. Any additional analyses must be requested post hoc and performed by the data host. This can hinder exploratory analysis and take a long time to complete: the pace of analysis moves at the rate of the slowest contributing study.
- Each contributing cohort study has to commit to providing an analyst's time to prepare the data and follow the analysis plans.
- Data preparation and analytical plans may not be implemented correctly or interpreted consistently across all studies. If any mistakes or interesting results are found a recalculation is often the only solution, which can take a long time to complete.

How is InterConnect's approach different?

To address the constraints of the existing methods, InterConnect uses the DataSHIELD platform for analyses across different cohorts. Rather than physically pooling the data together for analysis as in a standard ILMA, the data stay at the host institution. The platform allows an analyst to pass commands to each server and results that do not disclose the identity of any study participants are returned to the analyst. There is no access to the individual records within each study. In effect, DataSHIELD ‘takes the analysis to the data’ to return summary results that can be combined to be analytically equivalent to ILMA but without any access to individual-level data.

While this may seem the same as the conventional results sharing approach there are significant advantages:

- The analyst has the flexibility to refine and rerun analyses quickly, without waiting for an analyst at each institution to follow an updated analysis plan. If an interesting or spurious result is found during analysis this can be investigated immediately.
- Each cohort study does not have the burden of providing an analyst’s time to prepare and analyse the data held at their institution.
- The analyst can be sure that data preparation and analysis plans are implemented exactly as designed, removing the need to assume that every study returns correct results to the analyst.

How does analysis with DataSHIELD work?

With the DataSHIELD platform, individual-level participant data from contributing studies are held securely on geographically-dispersed servers that are based at the institution that collected the study data. Analytical commands are sent as blocks of code from another server within the network and these commands request each study-based server to undertake an analysis and return results that do not reveal the identity of any of the study participants. Such results are called ‘non-identifiable summary statistics’; they give useful information about the individual-level study data but are distinct from it. The analyst sending the commands has no access to the individual records of study participants at all and cannot physically see such data. The analyses are performed locally so all data stays at source, within the governance structure of the originating study, and the Investigators who are responsible for the study remain in complete control of its use, by controlling who has access to the server to run analyses. The analytical flow is shown in Figure 1 below.

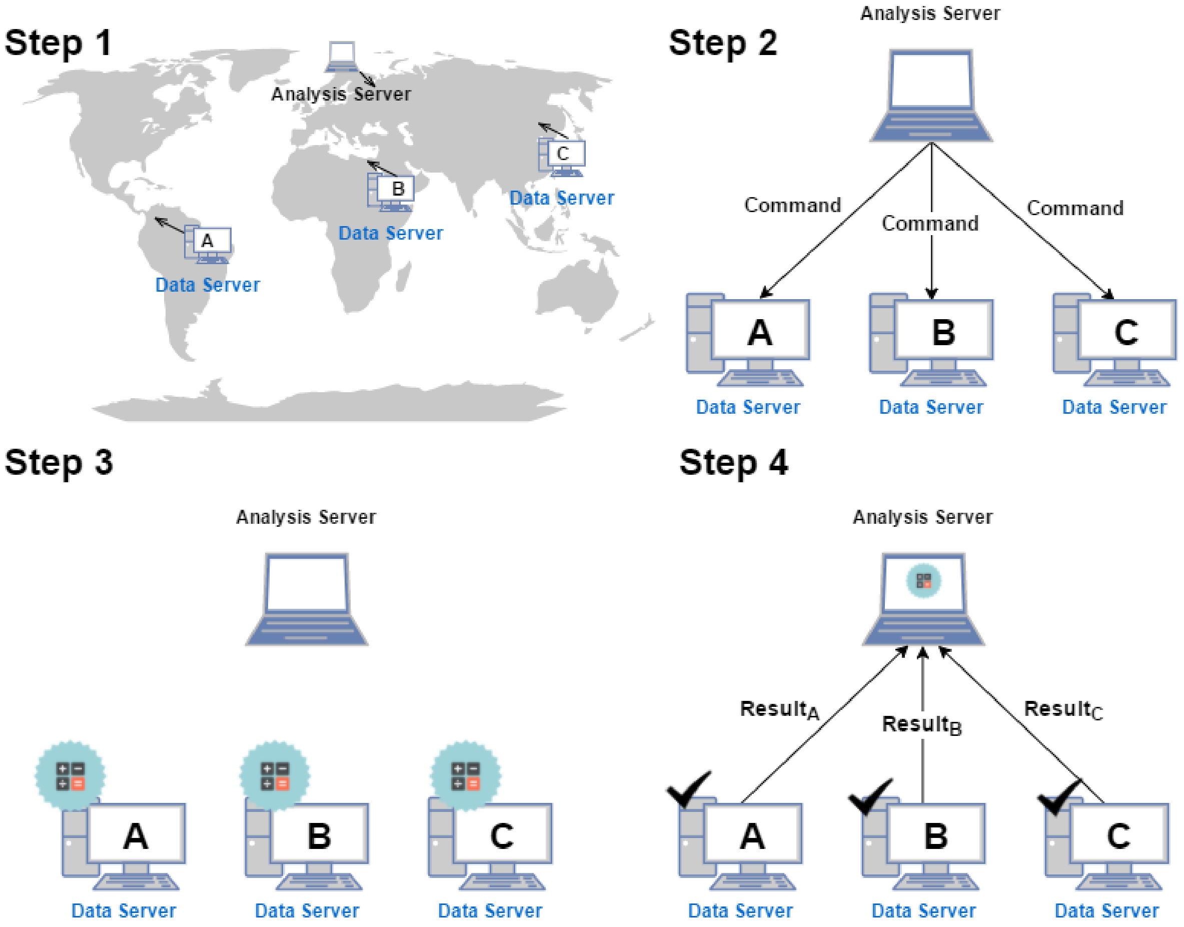

Figure 1: ***Step 1:*** *The analyst using the analysis server wishes to conduct a cross cohort analysis using data dispersed geographically on separate data servers A, B, and C.* ***Step 2:*** *If studies agree to collaborate and appropriate permissions granted, the analyst can use the analysis server to send analytical commands to the data servers.* ***Step 3:*** *The data servers run the commands on their data locally using installed statistical packages; meanwhile the analysis server waits for results to return.* ***Step 4:*** *Once the data servers have completed running the commands, the non-identifying summary results are returned for further aggregations and final calculations. Individual level data remain on each of the data servers and the analyst has no access to these at any stage of the analysis using DataSHIELD.*

How is a mean calculated with DataSHIELD?

A simple example of how an analysis is performed on the DataSHIELD platform can be demonstrated by calculating a mean. Suppose we have a list of 15 ages:

$$[20, 20, 30, 20, 30, 40, 25, 30, 40, 60, 70, 24, 16, 20, 50]$$

The average age would be the sum of all the ages in the list divided by number of ages in the list. This gives 30:

$$\frac{20 + 20 + 30 + 20 + 30 + 40 + 25 + 30 + 40 + 60 + 20 + 30 + 15 + 20 + 50}{15}$$

$$= \frac{450}{15} \qquad = 30$$

Now let us assume that this list of ages is split into 4 different groups, or studies, so that these numbers represent the age of participants in a study. This is what an analyst would see in a conventional ILMA.

Study	Ages in database
A	[20, 20, 30, 20]
B	[30, 40, 25, 30, 40]
C	[60, 20, 30]
D	[15, 20, 50]

By having each study submit only non-identifiable summary statistics to a central location it is possible to calculate exactly the same mean as pooling all the ages in a single list as shown previously. To do this, DataSHIELD first calculates the four sums at each study along with the number of relevant participants in the study. The table below shows the information a DataSHIELD user would be able to see from the participating studies that have given explicit permission to this user.

Study	Number of People	Sum of Ages
A	4	90
B	5	165
C	3	110
D	3	85

We then have all the non-identifiable summary statistics (split means) necessary to calculate the means at each individual site, as well as the global mean.

$$\frac{Sum\ Age_A + Sum\ Age_B + Sum\ Age_C + Sum\ Age_D}{NumPeople_A + NumPeople_B + NumPeople_C + NumPeople_D}$$

$$= \frac{90 + 165 + 110 + 85}{4 + 5 + 3 + 3}$$

$$= \frac{450}{15} \qquad = 30$$

The calculation of the mean is a simple example, yet highlights the key process in which non-identifiable summary statistics can be used to calculate results mathematically equivalent to those made by pooling all of the data in a central location (Jones, 2012). This extends readily into other calculations as well such as a regression as shown in the next section.

How is a regression and meta-analysis performed with DataSHIELD?

More complex analyses such as regression analysis using a generalised linear model can be used to look at the association between an outcome and exposures. For example, we might be interested in the association of exercise with the tendency of individuals to develop diabetes. In a typical cross-cohort analysis, associations are computed for each study and combined using meta-analysis, which allows a systematic assessment of the results provided by each study to give an overall conclusion on the association studied.

This type of analysis can be performed using DataSHIELD. The analyst uses the analysis server to send a command to the data hosting servers to perform regression for a model that is being tested. A generalized linear model is fitted to the data at each of the data servers. The model estimates, standard error, model, and other non-identifying results calculated at the data servers are sent back to the analytical server. Finally the analytical server produces a forest plot using these results. Figure 2 shows a simple diagram of this process.

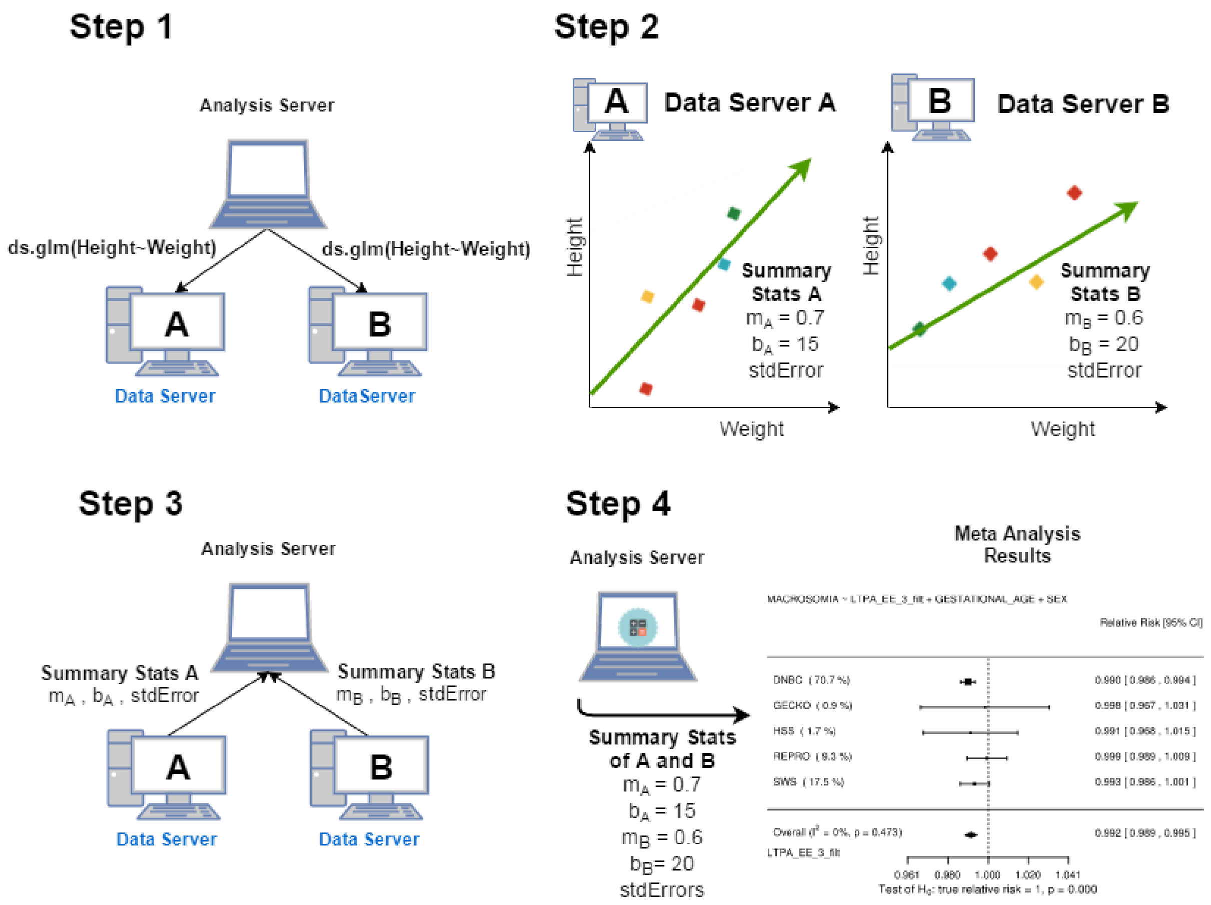

Figure 2: ***Step 1:*** *To investigate an association between variables, the analyst sends commands to participating studies to fit a Generalized Linear Model (GLM) to variables of interest.* ***Step 2:*** *The data servers at participating studies fit the generalized linear model; in the meantime the analysis server waits.* ***Step 3:*** *The studies return sufficient non-identifiable summary statistics of the regression to the analysis server.* ***Step 4:*** *The summary values are used to create forest plots from results obtained using models such as the Random Effects Model.*

What other types of analysis can be performed with DataSHIELD?

At the time of writing, work is in progress on providing functionality for time to event analysis using Cox regression. Additional functions can be developed in the future, provided they can be expressed as an algorithm that can be executed on each study server and will not reveal the identity of individual study participants during execution.

How are privacy and confidentiality maintained with the InterConnect network?

DataSHIELD is the interface between the analysts who are running cross-cohort investigations and the data hosted on servers of collaborating studies. It prevents analysts from accidentally or intentionally viewing individual-level data by prohibiting any calculations other than those necessary for non-identifiable summary statistics. It also records the analyses that have been performed to provide an audit trail. The non-identifiable summary statistics are untraceable to any one individual so that access to these statistics does not infringe upon the privacy of any individual (Gaye et. al., 2014). Therefore privacy and confidentiality are inherent to the DataSHIELD platform.

Calculations on DataSHIELD are limited to non-identifiable summary statistics through:

- Cell suppression (a technique that removes results that were generated from a small number of individuals and would increase the likelihood of revealing private data)
- Restricting the types of analysis and commands permitted which could lead to private data being revealed.

The study data remain housed locally, so data can easily be removed or restricted by the local study team at any time (e.g. in the case of withdrawal of participant consent). As such the control of the data remains in the hands of the local study data team.

The confidentiality and privacy provided by the DataSHIELD platform is also dependant on the security of the servers and their communication links, such that the features of the platform cannot be bypassed. These details are explained in the next section.

What security measures are in place on the InterConnect network?

The InterConnect network employs security measures to provide protection from external attack. The software tools in DataSHIELD use standard web security techniques, used on internet banking and e-commerce sites all over the world. Using Figure 3 below we can highlight some measures that are put in place:

1. Data exposure is minimised by including only a subset of the full dataset as required for the analysis agreed beforehand. Therefore sensitive data that are not needed for the analysis are not uploaded to the server.
2. The subset of data is only accessible to a genuine user authenticated to and designated to perform such analysis.
3. The analysis server is protected by a firewall that blocks external attempts to upload malicious programs

4. Each server hosting data at the collaborating studies' locations has firewall settings can be configured to only allow connections from the analysis.
5. Traffic between the servers is encrypted using industry standard tunnelling and encryption tools so that it cannot be read if intercepted.
6. Information sent between the servers consists of analytical commands and non-identifying summary statistics as produced through DataSHIELD.

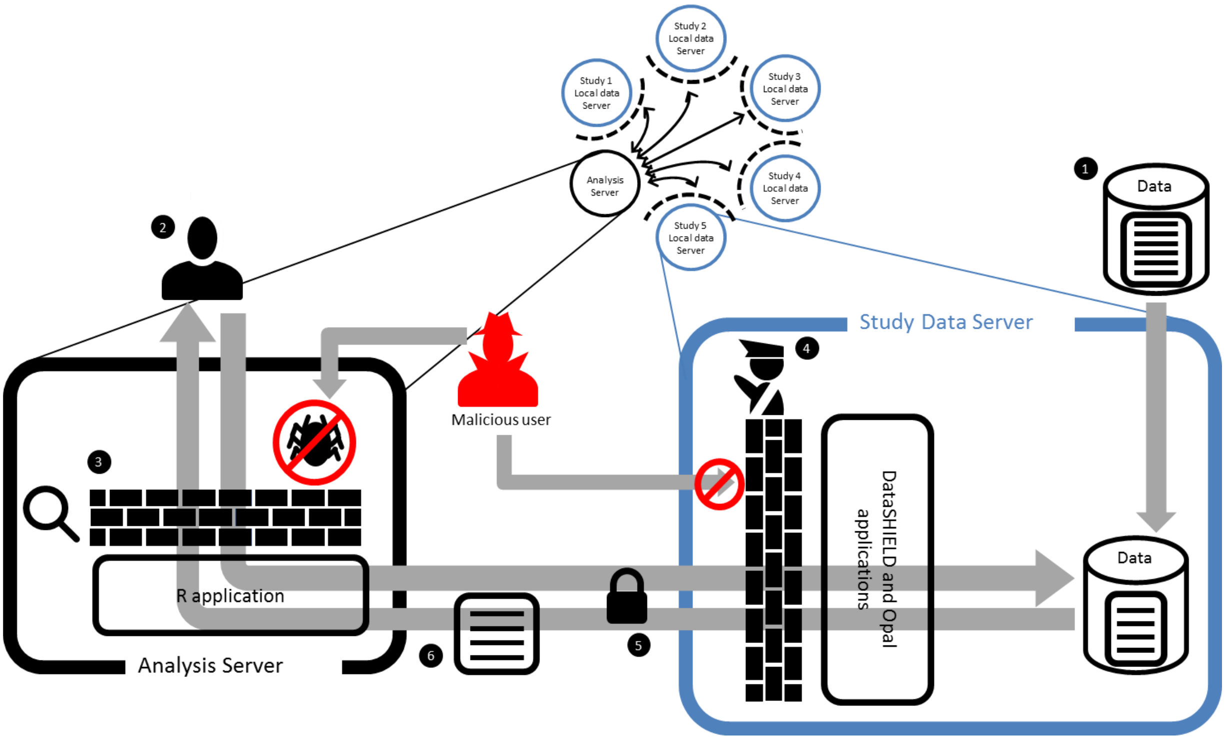
Study 2 Local data Server
Study 1 Local data Server
Study 3 Local data Server
Analysis Server
Study 4 Local data Server
Study 5 Local data Server
1
Data
2
Study Data Server
Malicious user
4
3
DataSHIELD and Opal applications
R application
Data
Analysis Server
6
5

Figure 3: *A Diagram highlighting the key security features used on the InterConnect network*

References

Gaye, A et al. "DataSHIELD: taking the analysis to the data not the data to the analysis", *Int J Epidemiol* 2014; 43 (6): 1929-1944

Wallace, S.E. et al. "Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law", *Public Health Genomics* 2014; 17(3):149-57

Jones, E.M. "DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective", *Norsk Epidemiologi*, 2012: 21 (2): 231-239