INTERCONNECT – IN BRIEF

InterConnect (www.interconnect-diabetes.eu) is a European Union FP-7 funded project that seeks to optimise the use of existing data to enable new research into the causes of diabetes and obesity.

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 physically bringing data together from cohort studies across the world is constrained by governance, ethical and legal challenges. To address this, InterConnect is taking a new approach to enabling cross-cohort analyses. Rather than physically bringing the data together for analysis, it is 'taking the analysis to the data'.

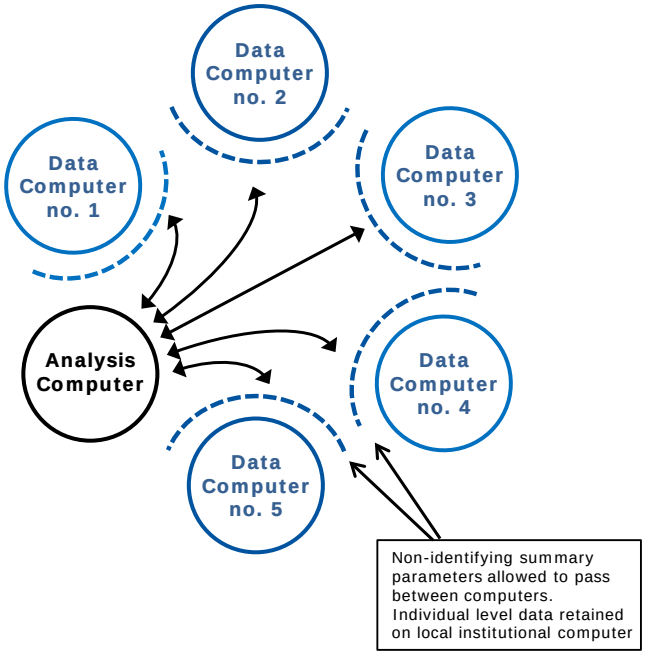
Our vision is to develop a global 'data access and results sharing' network that enables existing data to be readily used in cross-cohort analyses to understand population differences in the risk of diabetes and obesity.

A new approach to cross-cohort analyses

InterConnect provides a new approach to optimising the use of existing data which is secure, scalable and sustainable.

A fundamental aspect of the approach is a federated process. Individual participant data from contributing studies are held securely on geographically-dispersed, study based computers; analytical commands are sent as blocks of code from a computer within the network which requests each computer to undertake an analysis and return non-identifiable summary statistics (i.e. results, not data).

Analyses are performed locally so all data stays at source, within the governance structure of the originating study, and the study controls who can run an analysis.



Better access and new questions

Major governance, ethical and legal challenges limit the extent to which it is feasible to physically share data and so much cross-cohort collaboration relies on study-level meta-analyses and sharing of results.

In contrast, the InterConnect approach provides the opportunity to analyse individual level participant data from multiple studies through a secure, federated network. In this way, it is possible to perform an analysis that is equivalent to a meta-analysis of harmonised individual level data and so the approach is called 'federated meta-analysis'.

This enables researchers to move beyond traditional consortium-led or literature-based meta-analyses, which rely on study-level information obtained either from analysts within each individual study or published papers, to address new research questions.

Efficient and reusable

Sharing of results places a heavy analytical work load on each individual study that is participating in an analysis led by a third-party collaborator; this is ultimately unsustainable, particularly when additional information (e.g. covariance estimates) is needed from each study to enable conduct of a multivariate meta-analysis.

The InterConnect approach is radically different. The efforts of each study team are focused on preparation of the data and setting up an IT infrastructure which then enables them to participate in future analyses in a secure, scalable and sustainable manner:

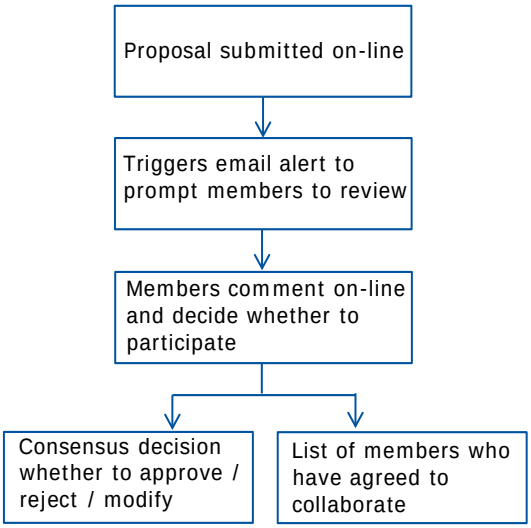
- Participation is by giving permission for local data access and, unlike results sharing, does not require the contributing cohort to perform any statistical analyses themselves.
- The infrastructure can be re-used on a long term basis, with the study team providing access to additional subsets of data depending on the specific analytical requirements of each new research question.
- Analyses are done in real-time on each study-specific server so once set up, there is no need to wait for outputs from individual study analysts.

It is therefore highly efficient and, unlike data sharing initiatives based on central data deposition, studies remain in complete control of their data, deciding whether to participate in each analyses on a case-by-case basis.

A democratic network

All those with responsibility for studies can play a central role, leading analyses across the network rather than simply providing data for others.

Transparent and democratic processes will be established to review proposals. A network-wide review will ensure the quality of the research proposal and serve to avoid duplication; studies will be able to decide whether or not to participate in each individual research proposal and will review them against their specific terms of consent and local data access policies. Publication rules will be developed to ensure appropriate recognition of the contributing parties.



Joining the network

InterConnect is conducting a number of initial exemplar projects to illustrate how the approach can make novel and otherwise challenging research questions tractable and to begin development of a collaborative 'data access and results sharing' network. To participate in the network, studies will be requested to:

- Make meta-data (i.e. questionnaires, data dictionaries and standard operating procedures) about the study available in English and answer questions to inform data harmonisation
- Provide, set up and maintain a secure server within their host institution; install open-source software and upgrades in collaboration with local IT staff
- Prepare the relevant data and upload onto the secure server; prepare and upload further variables for new analyses that the study chooses to participate in
- Review proposals for analyses and decide whether to participate, taking into consideration the original terms of consent and any local regulatory or data access / ethics committee processes
- In time, and where studies wish, develop proposals and lead cross-study analyses

INTERCONNECT – IN MORE DETAIL

Initiating the network via exemplar research questions

InterConnect is conducting a number of exemplar projects that address research questions of aetiological and public health interest and engage researchers in the approach. Such projects ensure that activities are linked to current and future scientific direction and enable us to understand the real-life issues that affect implementation.

Two exemplar projects are already well underway and you will be reading this document because you are considering whether or not to participate in a further project. The information below provides an overview of the work flow of an exemplar project and addresses some frequently asked questions (FAQs) and we encourage you to also look at the InterConnect website. We will invite study Investigators to a WebEx presentation to provide an opportunity for discussion.

Overview of an exemplar project work-flow

Registry

A registry has been set up with the initial purpose of enabling discovery of studies relevant to the risk of diabetes and obesity. It is being populated with publicly available information on the key characteristics of the study in a searchable format; the study team is simply asked to check that the information is correct.

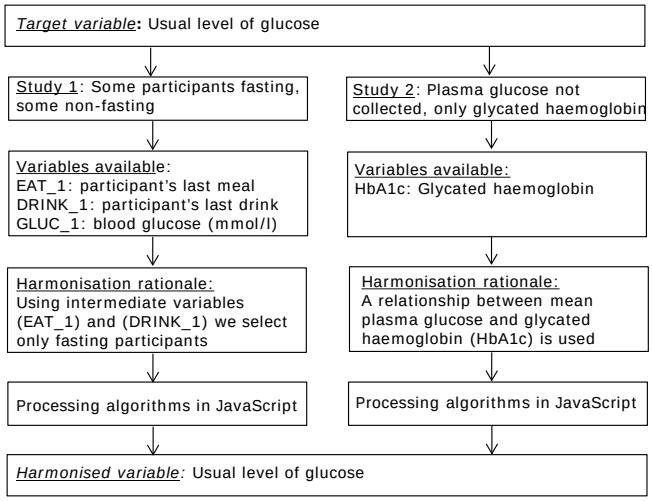
The functionality of the registry will be further developed for studies involved in exemplar projects. Additional information will include descriptions of relevant variables and harmonisation algorithms (see below) so that the harmonisation process is explicit and is also available for re-use in future projects.

Data harmonisation

Collaborating studies will be asked to provide meta-data (e.g. questionnaires, data dictionaries and standard operating procedures) about the study in English and answer questions to help the InterConnect core team develop a plan for data harmonisation.

Developing the data harmonisation plan involves defining the target variables that need to be derived from the data of participating studies and their format.

It will be a balance between uniformity (e.g. the exact question wording and data collection procedure used in all studies) and acceptance of a degree of heterogeneity between studies (e.g. slightly different wording or procedures used) in order to maximise the number of studies able to generate the target variables. An example of how one target variable can be derived from data from two studies is shown (right).



The harmonisation process compares the full definition and format of the target variable to the study-specific questions, collection procedures and data formats to determine their compatibility. Studies will be asked to respond to queries and check that they agree with the proposed harmonisation algorithm for each target variable. The core team will then code the algorithms in JavaScript ready for application.

The InterConnect core team will lead the harmonisation process but if any studies wish to be more closely involved or lead on aspects of it, they would be very welcome to do so.

IT set up

Studies will be asked to provide and set up a local, secure server and install free software which supports the data harmonisation algorithms and provides the environment for statistical computing (the 'R' programming language); both use the Linux operating system.

Set up will require input from local IT staff but is relatively straightforward and quick for those with Linux experience and technical advice will be provided. Further information on the set up process is provided in the FAQ below and the Standard Operating Procedures that will be provided.

There is flexibility as to when the study actually sets up the server so it can be scheduled in line with local priorities. However, studies will be asked early in the project to confirm that the set up is feasible since this is key requirement for participation.

Data upload

Once the harmonisation algorithms have been agreed, studies are asked to extract the data required to generate the target variables from their main database and load it onto their local, secure server.

Studies will also need to convert the sub-set of the data to be uploaded to the format defined by the software. Technical advice will be provided as required.

Analysis

The InterConnect core team will apply the JavaScript algorithm through remote access to the local data server to create the harmonised dataset for the study, hosted on the local institution's firewall-protected server.

The final step is the analysis of harmonised study datasets via securely encrypted remote connections via the software for federated analysis (DataSHIELD) which acts as an interface between the study database software and the statistical environment.

IT staff local to the study will set up and control the user accounts for remote access. The secure web-based link between the analysis computer and the data computers requires only a standard broadband internet connection.

The InterConnect core team will lead the analytical process for the exemplar projects but in due course others may wish to lead further analysis, through remote access to a shared analytical server maintained and supported by InterConnect.

Overview of the timelines for an exemplar project

The chart below provides a generalized view of the tasks and timelines relating to an exemplar project for illustrative purposes. Specific timelines may vary depending by individual project.

	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
Tasks completed by each study												
Complete meta-data request form				*+ study responds to queries*								
Confirms IT set up is feasible												
Sets up local IT		*(short task - schedule < month 8)*										
Prepare and upload data												
InterConnect core team tasks												
Plan data harmonisation and agree with studies												
Code harmonisation algorithms												
Support local IT set up and test connnectivity												
Apply harmonisation algorithms and conduct analyses												
Communication												
Regular TC / WebExs with studies*												

**Agree analysis plan, publicatoion policy, harmonisation approach etc*

FREQUENTLY ASKED QUESTIONS

About the network:

- *What types of studies will the network include?*

Studies will be relevant to the risk of T1DM, T2DM or obesity. The scope will be broad, only excluding studies that are recruited on the basis of diagnosed disease and case-control studies that are not related to diabetes, obesity or glycemic traits and are therefore unsuited to the overall research purpose.

- *How does InterConnect relate to other consortia?*

InterConnect is distinct from other consortia or networks. The main difference is that, in InterConnect, the majority of the work for an analysis is done by the team leading that analysis; other members are able to participate by preparing and uploading relevant data and providing the lead analyst with IT access to the data.

- *How was the software developed?*

InterConnect builds on the work of several research groups that have developed an integrated support platform for retrospective harmonisation and federated analysis of data; these have been tested in the FP7-funded BioSHaRE project (www.bioshare.eu) and made available as an open source toolkit by Maelstrom Research (www.maelstrom-research.org).

- *How will research questions and analyses be decided?*

All studies contributing data will be able to propose analyses and an online tool is being developed to support this. Proposals will uploaded and all members will be prompted to review and comment; if approved and if members with relevant study data wish to participate, a collaborative grouping will form to conduct the analysis. Study Investigators retain complete control of their data. If they do not wish to participate in an analysis, their data will not be included. This is achieved very simply; allowing your study data to be analysed is a proactive step requiring IT permissions to be put in place so it cannot happen without your permission.

- *What if I change my mind and want to withdraw my data from the network?*

Members are free to leave the network at any time. Investigators retain complete control of their data at all times and give active approval for analyses by giving permission for specific users (verified by password protected login or other credentials) to access specific data sets. Once partners are informed, technically withdrawing from the network is a simple matter of disconnecting the local server from the network which will take effect immediately.

- *What if I would like to contribute data but don't have the resource locally to set up a server or contribute to the harmonisation process?*

We encourage all studies to be active members by setting up a local server and then contributing to, or leading, network analyses. In some circumstances, it may be possible to request that one network member hosts data on behalf of another. However, the original data provider will still need to provide comprehensive meta-data to facilitate the harmonisation process and also consider individual requests for their datasets to be used in analyses.

- *What will happen to InterConnect when the European Union funding ends?*

In addition to benefits to researchers, the InterConnect approach also provides benefit to funders and wider stakeholders. Funders of research want to maximise the value of their investment in data and users of research evidence want to see the policy, social and economic benefits that can be realised by more effective data sharing. InterConnect is working with these groups to develop a shared vision for sustainability of the network.

About the IT set up, security and privacy:

- *Is the network secure?*

The software tools use standard web security techniques, used on internet banking and e-commerce sites all over the world. This includes the following features:

  - Encryption. Communication between users and servers cannot be intercepted and read by those who are not supposed to view that information.
  - Authentication. The identity of users can be verified either by login credentials or via a signed certificate so that they are only able to access data for which they have been given permission.
  - Connection management. A firewall is used to ensure that only the intended clients with specific IP addresses can connect to the servers. The study based local IT server is hosting a copy of the data from the underlying study database and only contains the sub-set of variables that are required for the analysis; there is no live link to the full study database.
  - Update management. To limit system security threats inherent to systems that are connected to the internet, software will be regularly patched and updated.

- *How is privacy and confidentiality maintained?*

The information produced and transferred out of the local data source via the federated analysis process is aggregated, non-disclosive summary statistics and untraceable to any one individual; access to this data does not infringe upon the privacy of any individual. Privacy and confidentiality is additionally addressed through a number of means:

  - Use of a number of data security methods during analysis, including cell suppression (a technique to avoid disclosure of sensitive tabular data), restrictions on the types of analysis permitted, and limits to risky commands which could lead to malicious statistical analyses.
  - Any breach of the data traffic between the analysis computer and local site can only yield aggregated summary statistics, thus not exposing any individual-level data to risk.
  - The study data remains housed locally and 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).

- *What are the ethical, legal or social issues (ELSI?)*

The ethical, legal and social issues are local. As with all research, those with responsibility of study data must ensure that the proposed analyses are consistent with the terms of consent for the original data collection and that all relevant institutional scientific, data access and ethical approvals are in place.

- *What are the hardware requirements?*

Hardware requirements are a minimum of 2 GB RAM with a recommendation for > 4 GB, disc space should be a minimum of 160 GB with a rule of thumb calculation of 10GB for the operating system and 4 GB per 10,000 participants and the CPU should be a recent server- grade or high-end consumer grade processor. Further information will be provided in the IT SOP. The cost will depend on the study institution's preferred hardware setup and the size of the study.

- *How long will it take?*

Setting up the Opal server (including R server installation), initial infrastructure testing and installing and testing DataSHIELD should take around 12 – 16 hours in total (i.e. not counting waiting time, such as waiting for someone to open the external firewall) assuming carried out by an IT professional following the SOP provided. Preparing and importing the dataset into Opal and setting up the server ready for harmonisation and analysis is likely to take a further 8 hours. Studies should also make some allowance for on-going work and configuration changes, estimated at 4 hours per month.

About network capabilities and research:

- *What analyses are possible through the federated analysis tool, DataSHIELD?*

The current analytical functions that are available can be found on the DataSHIELD wiki; these include the glm function for fitting generalised linear models. At the time of writing, work is in progress on providing functionality for time to event analysis using Cox regression and Kaplan-Meier plots. 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.

- *Will Mendelian Randomisation analyses be possible in InterConnect?*

Yes, it will be possible to perform Mendelian Randomisation analyses using summary estimates (e.g. beta coefficients, log odds ratios) and their standard errors obtained from regression models fit to data from InterConnect. Recently developed models based on individual level data which require WinBUGS software for implementation are not currently possible.

- *Do I have to use R for analysis or can I use other software such as Stata, SAS or SPSS*

Simple analyses such as contingency tables, summary statistics and data filtering can be performed through the web application that hosts the study registry. For more detailed analysis, R (though R Studio – www.rstudio.com) has been selected because it is available free of charge so any user can access to it. It is relatively straightforward to perform statistical analyses using the R studio interface, and there are numerous resources freely available on the internet for those who have not used R before.

- *Are there limits to the amount of data that can be uploaded onto the server for harmonisation and analysis?*

As with any system, there is a limit on the amount of data that can be stored and analysed. This limit is a function of the underlying hardware and the exact nature of the processing required.

- *Can I use InterConnect to study quantitative metabolic traits, or is it limited to diabetes and obesity?*

Yes, InterConnect can be used to study any trait or disease for which information is available in the source datasets.

- *What happens when I add more variables or data to my study later on?*

When a new variable becomes available for your study, this can be added to the data on your local study server. If this variable will be used in federated analysis and requires harmonisation with other existing variables, then the standard harmonisation process can be used.

About network outputs:

- *Will there be authorship rules for publications based on analyses done using InterConnect? How will a specific cohort receive credit for contributing data?*

Publication rules defining authorship of papers and other outputs will be developed and agreed by each particular grouping that comes together to address a specific research question. Each collaborative grouping will be self-determining and so all those providing data will have a say in deciding the authorship rules for outputs relating to the analysis to which they are contributing.

- *If I join the network, when will my input be required and when am I likely to see results?*

Your input will be required when you express interest in participating in a specific analysis. You will need to contribute to the stages defined in the document above – making meta- data available, setting up a local server, uploading the required variables and contributing to the harmonisation process. The time it takes to generate results will depend on the number of network members participating and the extent to which they are already set up.

FURTHER INFORMATION

Relevant publications

Budin-Ljøsne I et al: DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics. 2015;18(2):87-96. doi: 10.1159/000368959. Epub 2014 Dec 13.

Gaye A et al: DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014 Dec;43(6):1929-44. doi: 10.1093/ije/dyu188. Epub 2014 Sep 26.

Wallace SE et al: Protecting personal data in epidemiological research: DataSHIELD and UK law. Public Health Genomics 2014, 17(3), 149-157. doi:10.1159/000360255

Doiron D et al: Data harmonisation and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013 Nov 21;10(1):12. doi: 10.1186/1742-7622-10-12.

Jones EM et al: Combined analysis of correlated data when data cannot be pooled. Stat 2013;2:72-85.

Murtagh MJ et al: Securing the data economy: translating privacy and enacting security in the development of DataSHIELD. Public Health Genomics. 2012;15(5):243- 53. doi: 10.1159/000336673. Epub 2012 Jun 20.

Jones EM et al: DataSHIELD – shared individual-level analysis without sharing the data: A biostatistical perspective. Norsk Epidemiologi 2012;21:231-239.

Fortier I et al: International Harmonization Initiative. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol. 2011 Oct;40(5):1314-28. doi: 10.1093/ije/dyr106. Epub 2011 Jul 30.

Van Vliet-Ostaptchouk et al: The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. BMC Endocrine Disorders 2014, 14:9.

Links

InterConnect (www.interconnect-diabetes.eu) is a global collaborative network for diabetes and obesity research

Maelstrom Research (https://www.maelstrom-research.org/) provides software and methods to support harmonisation and integration, and this site contains a catalogue of studies that have undergone harmonisation

OBiBa (www.obiba.org) provide high-quality open source software for biobanks, including Mica (web portal software for epidemiological studies) and Opal (software to manage study data and enable data harmonisation using processing algorithms)DataSHIELD (http://datashield.org/) is a tool for federated analysis without physical sharing of data.

DataSHaPER (www.datashaper.org) is both a scientific approach and a suite of tools to facilitate the prospective harmonization of biobanks.

R software (www.r-project.org) is a free, open source statistical software package.

BioSHaRE (www.bioshare.eu) is a consortium of leading biobanks and research from many domains, which aims to build tools and methods to enable researchers to use pooled data from different cohort and biobank studies.