Global data for diabetes and obesity research

MEETING REPORT

INTERCONNECT: A GLOBAL INITIATIVE ON GENE-ENVIRONMENT INTERACTION IN DIABETES AND OBESITY. *Funded by EU FP7 grant agreement 602068*

Monday 14 September 2015, European Association for the Study of Diabetes, Stockholm, Sweden

MEETING ABSTRACT

This Symposium was to engage researchers in a new approach to optimising the use of existing data that is secure, scalable and sustainable within a simple governance framework. The FP7-funded InterConnect project aims to change the way that data are used in population research into the causes of diabetes and obesity. It seeks to create the foundation to enable research to move from explaining differences in the risk of diabetes and obesity within populations to being able to explain differences in risk between populations. Cross-cohort analyses will be enabled by an online study registry to enhance data discovery, optimising data for re-use via harmonisation and new toolkits and by creating a network for federated meta-analysis where data stays at source and the analysis comes to the data.

SESSION 1: SCIENTIFIC OPPORTUNITY, CHALLENGE AND VISION

Understanding differences in risk of diabetes and obesity between populations (Professor Nick Wareham, *InterConnect Co-ordinator & Director, MRC Epidemiology Unit, University of Cambridge, UK)*

Professor Nick Wareham, Co-ordinator of InterConnect and Chair of the Symposium, welcomed the participants. He opened by explaining how the InterConnect initiative is taking forward the vision from the international conference hosted by the European Commission in February 2012 on *'Diabesity – A worldwide challenge: towards a global initiative on gene-environment interactions in diabetes/obesity in specific populations'.*[1] The Diabesity meeting indicated that diabetes is a world-wide challenge and rising to this challenge entails different aspects of research, namely: research into individual and societal approaches to the prevention of obesity, diabetes and related metabolic disorders and health systems' interventions to better treat diabetes.

The aim of the Symposium was to highlight the scientific opportunity that arises from enabling research to move from explaining differences in risk of diabetes and obesity *within* populations to being able to explain differences in risk *between* populations. Enabling cross-cohort analyses is key to achieving this transition and is the focus of InterConnect.

Prof Wareham first described research into the incidence of type 1 diabetes. Finland has one of the highest levels of incidence, as does Sardinia. This has led to ongoing cohort studies within these specific populations, investigating the complex interplay between probable genetic susceptibility and environmental trigger(s). A related but different pattern exists for type 2 diabetes, with very high prevalence in some counties but not in others.

Some studies have tried to take this further, looking for possible explanations for these major differences in risk. The investigation of these epidemiological observations has gone through several different phases. During the first, descriptive, phase, in 1962, an hypothesis based on 'thrifty genes' (which enable individuals to efficiently collect and process food to deposit fat during periods of food abundance[2)] was developed and then, in 1992, the 'thrifty

phenotype' hypothesis was proposed (in which reduced foetal growth due to adaptations made by the foetus in an environment limited in its supply of nutrients is strongly associated with a number of chronic conditions later in life[3]).

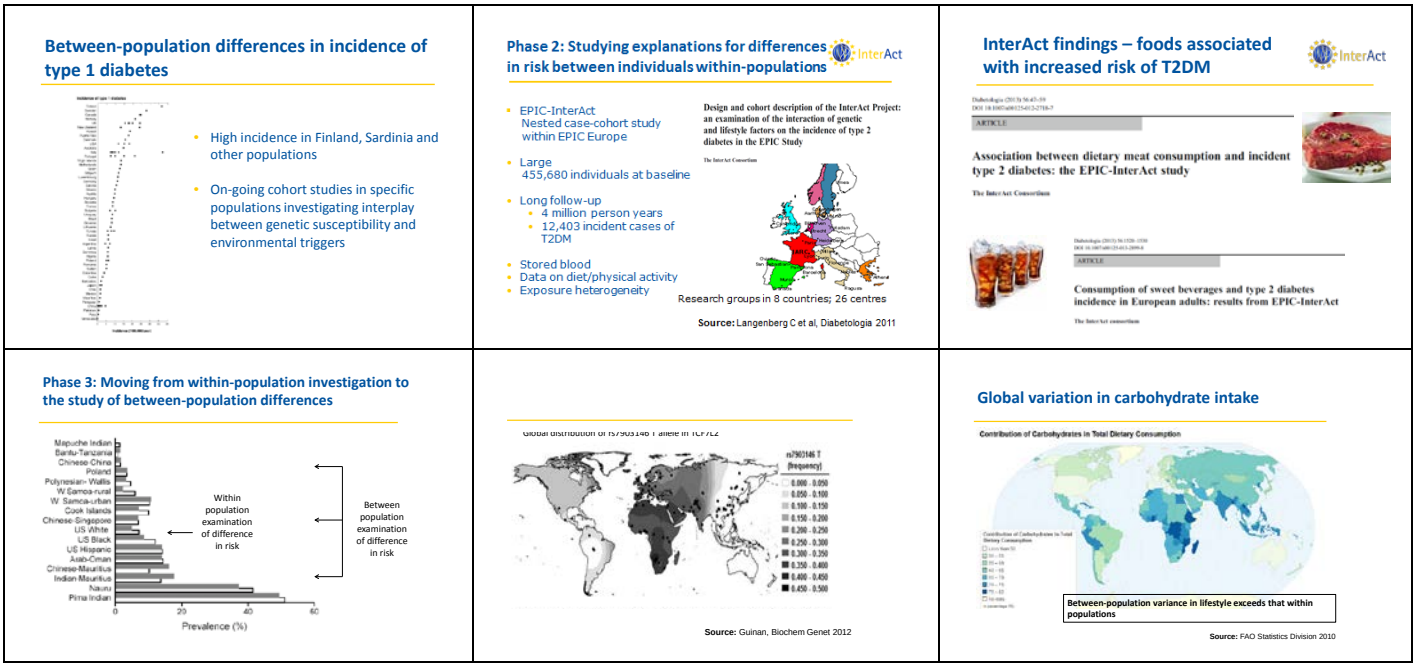
The second phase then looked within cohorts to try to explain differences and new, larger studies are now trying to understand individual risk *within* populations in more detail. The InterAct project[4] (www.inter-act.eu, funded under the EC Framework 6 Programme) is a study of half a million Europeans, including a cohort of 12,403 with type 2 diabetes. InterAct is analysing the effects of foods that might be associated with increased risk of type 2 diabetes, such as sugar-sweetened beverages and processed meat, alongside foods that might be protective, such as fish, fruit, vegetables and certain dairy products; it is also looking at physical activity and how this might be protective. Given its scale, InterAct is an important study in the field. However, even with a study of this size, no evidence of interaction of lifestyle factors with individual genes of known variants has been found. Consequently, global analyses across multiple studies are required if we are to further research on gene-environment interaction.

Prof Wareham went on to explain how the research community now needs to develop ways to study the large differences in risk that exist *between* populations - and how InterConnect is rising to this challenge. Globally, between-population differences in both genetics and also lifestyle or environmental factors are considerably larger than differences *within* populations. Unless we think globally we cannot tackle this problem. InterConnect attempts to address how we can bring individual level data together from around the world. To realise the vision of bringing data together to allow the study of between-population differences in risk, InterConnect aims to help researchers to:

- find relevant studies globally
- find out what data the studies have collected
- find an appropriate way of bringing data together
- find a way of interpreting different forms of data that are brought together – challenging but tractable.

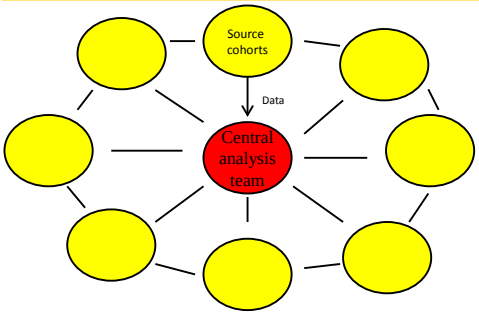
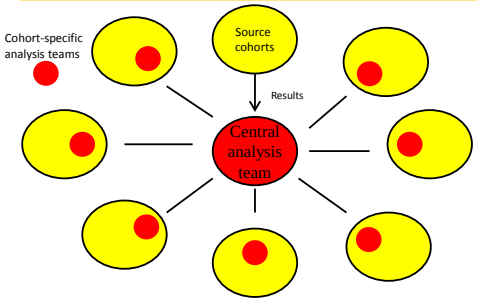
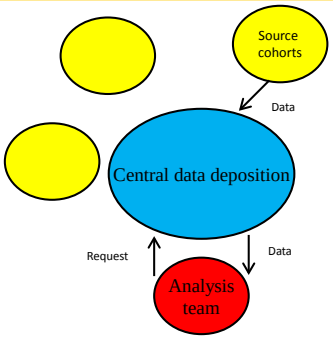
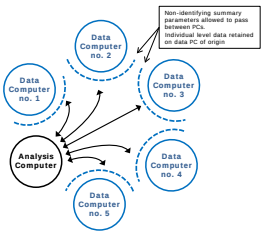
In this way, InterConnect aims to create the foundation for a sustainable, global network for diabetes and obesity population research that enables research to move from explaining differences in risk *within* populations to being able to explain the major variations in risk *between* populations.

Selected slides:



Challenges of data-sharing models (Dr Nita Forouhi, *InterConnect WP4 Leader & MRC Epidemiology Unit, University of Cambridge, UK)*

In this interactive session, Dr Forouhi invited participants to think through different data-sharing models, and the possible benefits and difficulties of each from different stakeholder perspectives. Ultimately, the aim was to try to imagine a future in which we are trying to connect multiple studies at a global level.

Model 1: Sharing of data between cohorts using traditional collaboration/consortia agreements 	This traditional model consists of physically sharing data, typically involving collaboration between a number of cohorts and a central analytical team. Almost half of the participants had used this model. While it enables physical sharing of individual level data and in-depth individual level meta-analysis, it also poses numerous challenges. In the discussion that followed, it was noted that this model is problematic in terms of regulatory issues relating to cross-border transfer, requires well established collaborative networks and trust between partners and if centralised around a sole analytical centre, resentment could potentially arise about imbalance of opportunities to lead analyses as opposed to contributing data.
Model 2: Ad hoc consortia - sharing of results 	This model has been shown to work well for genetic data, as the results of analysis are sent, not individual level data so avoiding administrative and organizational complexity. Typically, each cohort has its own analytical team and does its own analysis; results are then submitted and collated centrally to complete the analysis. Some of the ethical issues are eased here as the data remains in the control of the researcher who collected it. However, challenges do remain. The quality of the research could be variable as the central analysis team cannot control or standardise other cohorts' work within the project. Also, even if an organization is well resourced, there are large numbers of requests and investigators can find themselves servicing other people's research and not their own.
Model 3: Central deposition of data 	In this model, data are deposited into a central repository. The analytical team puts in a request, receives data and can then undertake analyses. Some participants had experience of using this more democratic model, which provides greater opportunity to a wide range of researchers to access the data. This model has advantages if a cohort does not have analytical capability, but governance issues still exist around who legally holds and owns data and how the data is centralised; there are also issues around potential duplication of work and confidentiality. This model is not likely to be sustainable. Access decisions can require delegated authority which would be a significant challenge on a global scale. Central deposition is also very difficult to mandate for historical data collections.
The future: Federated meta-analysis  • Data stays within governance structure of source cohort • Cohorts focus efforts on preparation of data and IT infrastructure for sharing • Analytical effort more focused on the scientific –led questions	Dr Forouhi went on to explain the InterConnect model is based on a 'federated meta-analysis' approach to enable data access for cross-cohort analyses. Here, the data stays behind the firewall of the server in the host institution, where the data was collected, and the data does not move. It stays within the governance structure of the source cohort. Cohorts focus their effort on preparing the data and providing access for analyses that are conducted remotely, with only results passing between computers. This can facilitate a democratic system where all partners can equally drive analyses. It also avoids the governance, ethical and legal challenges of conventional approaches to data sharing. Cohorts maintain control of their own data while enabled to access results generated from analysis of a wide spectrum of data, so providing access to greater heterogeneity for research.

InterConnect: Vision of a changed paradigm (Prof Nick Wareham)

In this session, Prof Wareham went on to explore the InterConnect vision which focuses on optimising the use of existing data to enable cross-cohort analyses, which is distinct from data sharing per se, more fully.

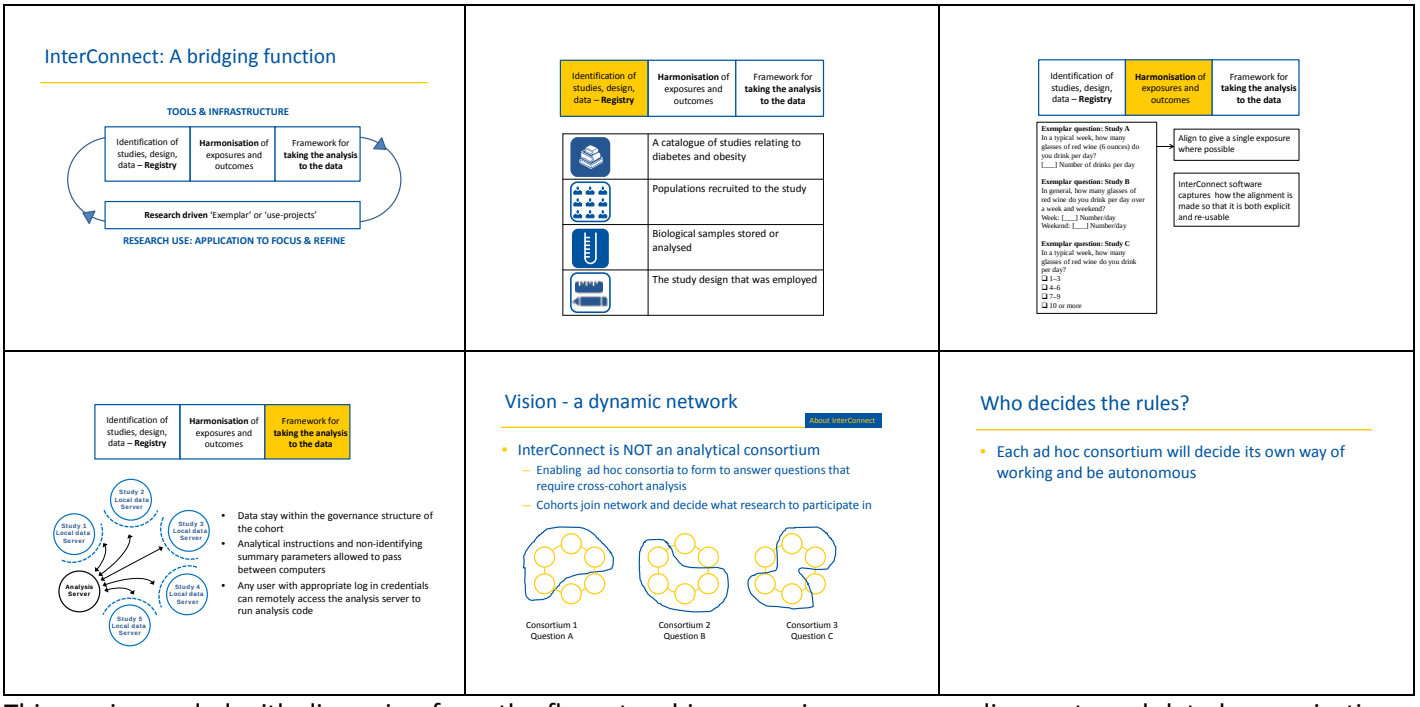
Data pooling and deposition agreements are complex and can be a barrier to cross-cohort analyses, given collaborators' fears around loss of ownership as well as the associated governance, ethical and legal issues. Traditional results sharing processes place a huge burden on collaborators in terms of preparing and analysing data, with funders paying for this work without clarity around how funding is being used.

Individual participant data meta-analysis using pooled data from specific cohorts is analytically desirable, as it is based on individual level data. InterConnect is enabling individual level data analyses without the necessity of the physical pooling of data by 'taking the analysis' to the data, rather than vice versa. While there is a place for new global studies, these are expensive and time consuming, and there is therefore merit in fully utilising existing data sets from around the world.

InterConnect builds on the work of a number of research groups, particularly Maelstrom Research and the EU-FP7 funded BioSHaRE project, that have developed infrastructures and open source tools to catalogue studies, support data harmonisation and enable federated meta-analysis.[5] InterConnect provides a bridge between the development of these tools and implementation by the diabetes research community. Researchers need to engage with this initiative for it to be successful. InterConnect aims to illustrate the effectiveness of the approach through exemplar projects. These will illustrate three main pillars:

- improving findability and identification of studies through the **Registry**
- facilitating **Harmonisation** of exposures and outcomes between studies efficiently and explicitly
- creating a **Framework** for taking the analysis to the data.

Selected slides:



This session ended with discussion from the floor, touching upon issues surrounding costs and data harmonisation. Prof Wareham noted that while there will be set-up costs involved, these are almost negligible compared with the expense of undertaking new studies. Participants were also keen to discuss data security, which is described more fully in the next section. Stakeholders from patient organisations were also reassured to learn that patients would be regarded as key stakeholders in this process, and indeed generally want all data to be utilised as widely as possible.

In order to drive this process forward, Dr Ken Ong went on to provide an insight into how the registry is developing. Professor Matthias Schulze, the investigator leading the registry topic as a whole was unfortunately unable to present in this session as originally planned due to illness.

SESSION 2: DELIVERING THE INTERCONNECT VISION

Data discovery: the registry (Dr Ken Ong, *InterConnect WP3 Leader & MRC Epidemiology Unit, University of Cambridge, UK)*

Dr Ong explained that a major part of InterConnect is the development of the study registry. This is necessary for researchers to be able to find out:

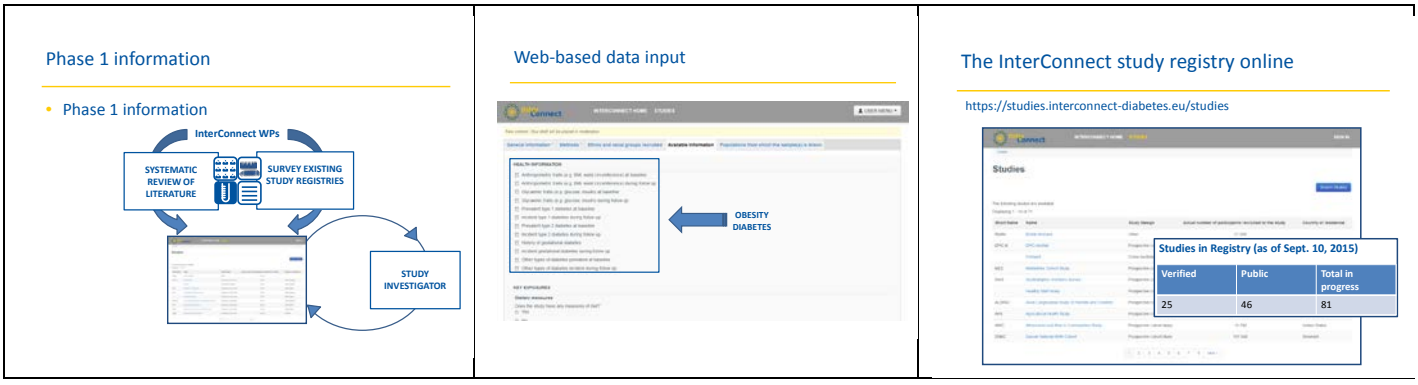
- what resources are available globally
- what study design was employed
- what populations were recruited
- whether samples were stored
- what data are available.

The registry work packages are responsible for setting up a standardised database and web-based procedures, so that external researchers across the world can contribute to it electronically, and to prepare a website which hosts the visualisation of the registry.

The registry has two phases of development. The first phase of the registry takes a 'broad and shallow approach' and is focused on gathering simple but useful information that can largely be collected from information already in public domain such as general information (study name, contact persons, web link), study design, ethnicity and race, the sampling frames, recruitment information and basic participant characteristics. This approach creates little burden for individual studies while enabling sign-posting of large numbers of useful studies. The registry is being populated by the InterConnect team and study investigators are being asked simply to verify the information. There are currently 71 studies in the registry, 25 have been verified by the study project Investigators and a further 46 entries are derived from publicly-available information.

In addition to this 'broad and shallow' phase one in registry development, there will also be a second phase built around the exemplar projects. This more detailed phase of the registry development will involve information to be collected directly from studies and will incorporate metadata about available data such as data sources and categories of available data (e.g. health, socio-demographic, lifestyle, physiological, biochemical, genotype information).

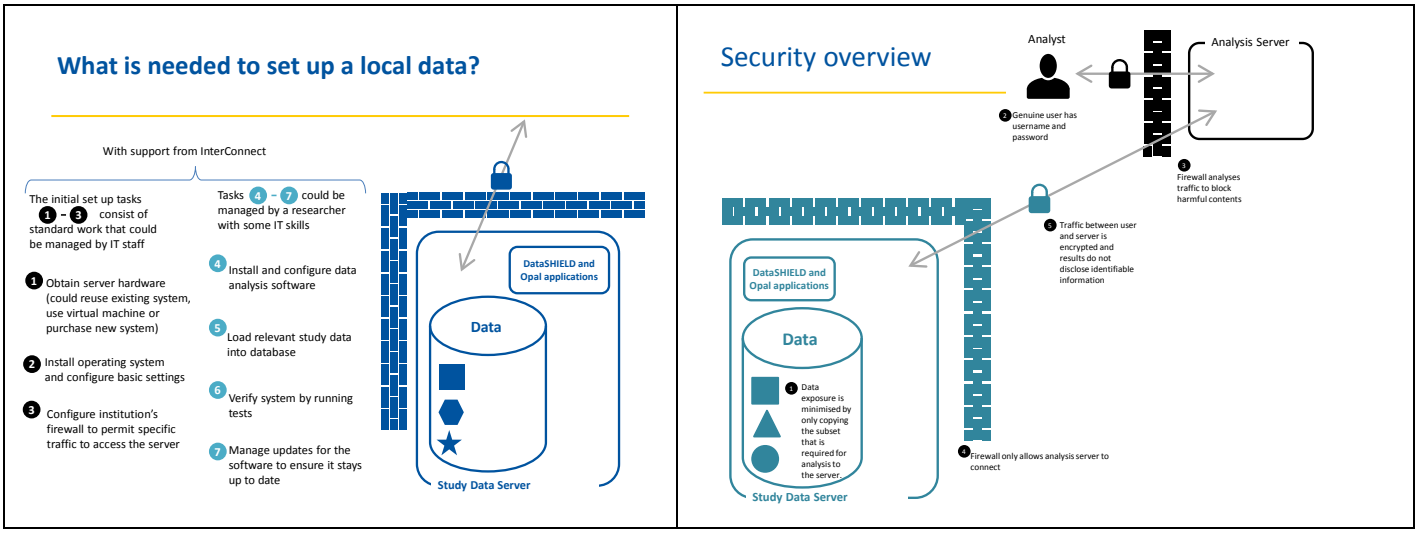
Selected slides:



Bringing the analysis to the data: proof of concept (Mr Tom Bishop, *Senior Data Scientist, InterConnect, MRC Epidemiology Unit, University of Cambridge, UK*)

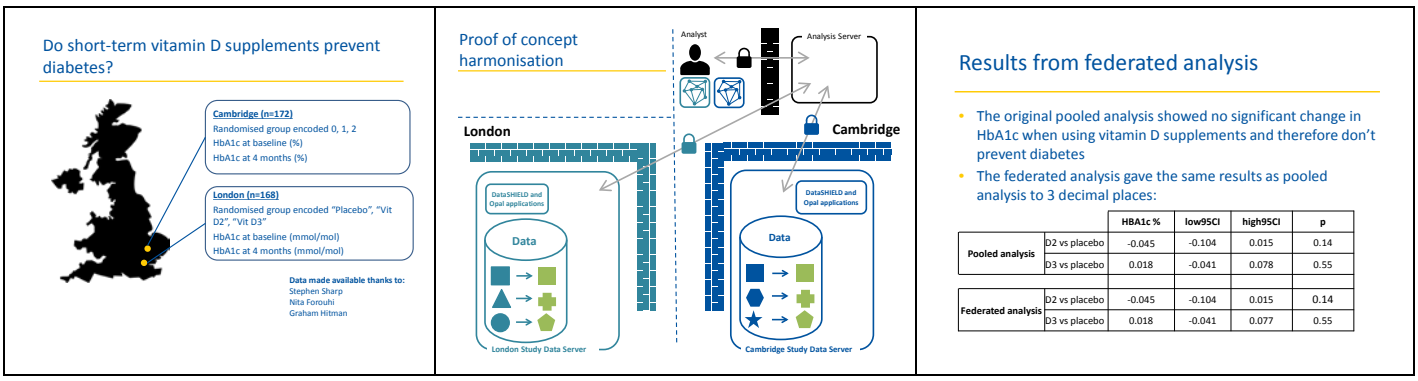
In this session, Mr Tom Bishop focused on the technical details involved in federated meta-analysis and how this can be of benefit to diabetes research. The InterConnect team has completed a proof of concept to demonstrate the technologies and methods work. This proof of concept also provides a way to illustrate the technical work required to set up a server and participate in research and the security features that are incorporated for data protection. These aspects are illustrated in the slides below.

Selected slides:



Proof of concept was based around the question of whether short-term intake of vitamin D supplements prevent diabetes, using data previously collected in London and Cambridge that had been physically pooled for analysis. The InterConnect team essentially repeated the analysis but kept the two data sets separate, on two different servers. A harmonization step was needed, and involved generating harmonization algorithms that were loaded onto respective servers, in Cambridge and London, resulting in a common set of variables. The original pooled analysis showed no significant change in HbA_{1c} when using vitamin D supplements and therefore they do not prevent diabetes. Crucially, the federated analysis gave the same results as the pooled analysis to 2 decimal places.

Selected slides:



		HBA1c %	low95CI	high95CI	p
Pooled analysis	D2 vs placebo	-0.045	-0.104	0.015	0.14
	D3 vs placebo	0.018	-0.041	0.078	0.55
Federated analysis	D2 vs placebo	-0.045	-0.104	0.015	0.14
	D3 vs placebo	0.018	-0.041	0.077	0.55

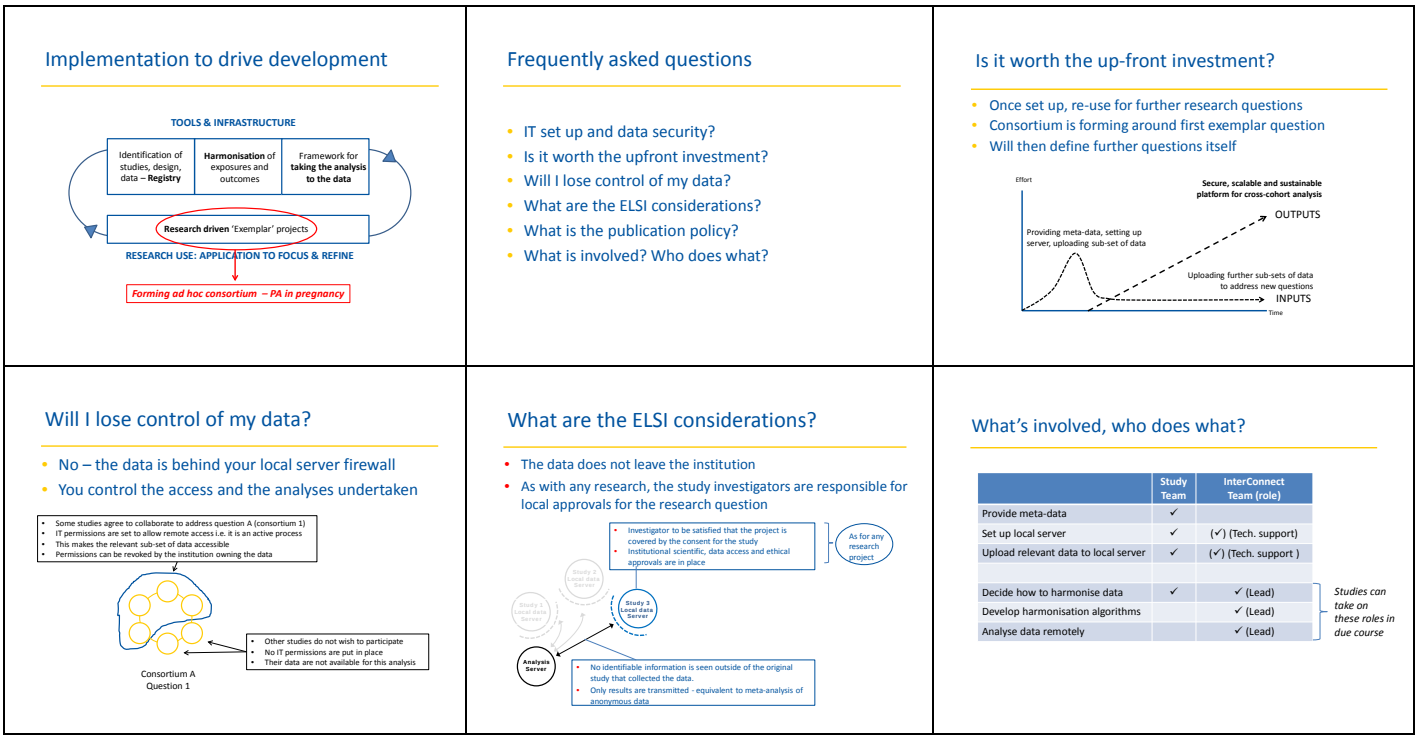
This successful pilot project demonstrates the effectiveness of the InterConnect approach. The next step is to work on new, original research questions (exemplars) and develop new analysis functionality. At this point in the session, participants were keen to engage with technical issues surrounding the protection of individual data and the encryption process and these aspects were addressed in the discussion. Prof Wareham introduced the next session, in which Dr Ong outlined how the exemplar research projects are successfully engaging researchers in the InterConnect approach.

Developing the vision via exemplar research questions (Ken Ong)

Dr Ong described how the vision was being developed through real-life exemplar research questions. The first question being taken forward focused on "is a mother's higher level of physical activity during pregnancy associated with lower offspring adiposity at birth?" This is an important research question, given the short-term risks of a large baby for the mother and her newborn and also the hypothesised long-term programming of metabolism in the offspring. Existing reviews show variable evidence and the effect of physical activity also differs in different subgroups. The exemplar questions are not simply procedural tests to develop and fine tune processes, but simultaneously seek to address important, unresolved research questions.

Dr Ong outlined the process through which he and Prof Gernot Desoye (InterConnect Research Network Lead for Pregnancy and Childhood) brought a number of cohorts together to engage with the InterConnect vision. They

identified relevant studies by contacting known investigators and undertaking literature searches. In the future, this process would be aided by the InterConnect registry. An *ad hoc* consortium was subsequently formed, after discussion through a WebEx meeting to talk through practical issues and address FAQs. Ethical, legal and social considerations were also discussed, but as the data do not leave the institution, these considerations are the same for any other research question, with the study investigators responsible for local approvals. Key concepts and discussion topics are illustrated by the slides below.



Dr Ong was pleased to conclude that a collaborative grouping had now successfully formed to address this research question and the work was now underway. An additional exemplar research question is also being developed around fish intake and type 2 diabetes, and others using genetics and GIS will also drive the future utility of the tools.

OPEN DISCUSSION AND INVOLVEMENT (*Prof Nick Wareham*)

To conclude the meeting, Prof Wareham welcomed questions and responses from the floor. A number of participants commented on the potential benefits of the InterConnect approach in enabling them to use existing global research data to the full. Prof Wareham made the point that while this kind of approach has also been initiated in other disease areas, diabetes might lead the way in practical implementation to address important research questions. It was noted that participants in the first exemplar question are limited to Western Europe and the USA; other exemplar questions will attract a wider spectrum of studies over time making the network more international. Cohort size was also discussed, with participants asking whether this is an issue. Prof Wareham commented that there is a current trend in epidemiology to prioritise large studies. However, variation is also essential, and smaller studies can be highly informative and so are also very much encouraged to participate. Prof Wareham also explained that InterConnect is developing a consortium management tool to enable collaborating cohorts to co-ordinate and make their own choices as to how they operate. To conclude, Prof Wareham thanked everyone for attending, encouraged participants to get involved with InterConnect, and to explore the available papers that provide the statistical basis for federated meta-data.

SPEAKER BIOGRAPHIES

Mr Tom Bishop, Senior Data Scientist, InterConnect, MRC Epidemiology Unit, University of Cambridge, UK. Tom is a Senior Data Scientist in the Aetiology of Diabetes and Related Metabolic Disorders programme, and is responsible for the technical delivery of InterConnect. Tom has a Master's degree in mechanical engineering from the University of Cambridge.

Dr Nita Forouhi, Group Leader, Nutritional Epidemiology Programme, MRC Epidemiology Unit, University of Cambridge, UK. Dr Forouhi leads the Nutritional Epidemiology Programme, which aims to understand the relationship between diet, nutrition and the risk of diabetes, obesity and related disorders. Dr Forouhi is also an Honorary Consultant Public Health Physician with Public Health England. She is the Chair of the area multidisciplinary Managed Care Network for Diabetes, a member of Diabetes UK Research Committee and Associate Editor of *Diabetic Medicine*.

Dr Ken Ong, Group Leader, Growth and Development Programme, MRC Epidemiology Unit, University of Cambridge, UK. Dr Ong leads the Unit's Growth and Development programme. His research identified rapid postnatal growth, weight gain and reproductive timing as determinants of, and also potential targets for the prevention of, childhood obesity, type 2 diabetes and related disorders. Dr Ong obtained his PhD in Paediatrics following research at the Universities of Oxford and Cambridge, using large population-based studies to explore gene-environment interactions in fetal and early childhood growth.

Professor Nick Wareham, Director MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, UK. Prof Wareham is an Honorary Consultant at Addenbrooke's Hospital, Cambridge and Co-Director of the Institute of Metabolic Science. He was formerly a Wellcome Trust Senior Fellow in Clinical Science in Cambridge. Professor Wareham was the coordinator of the EU FP6-funded InterAct project, which investigated how genes and lifestyle factors interact to lead to type 2 diabetes. His work on gene-environment interaction is based on quantitative trait studies and large scale population-based cohort studies. Prof Wareham is co-lead of the ADDITION study, a trial of screening for diabetes and intensive cardiovascular risk reduction undertaken in three European Countries.

References

1. 'Diabesity – A worldwide challenge: towards a global initiative on gene-environment interactions in diabetes/obesity in specific populations' http://ec.europa.eu/research/health/pdf/diabesity-conference-report-022012_en.pdf.
2. Neel, *American Journal of Human Genetics*, 1962.
3. Hales and Barker, *Diabetologia*, 1992.
4. The InterAct project funded under the EC Framework 6 Programme www.inter-act.eu.
5. BioSHaRE aims to facilitate data harmonisation and standardisation, data sharing and pooling across multiple biobanks and databases. It is a consortium of leading population-based cohort studies, with international researchers from diverse domains of biobanking science, including epidemiologists, statisticians, software developers and ELSI experts www.bioshare.eu.