PublicHealth@Cambridge

Round table discussion
Public Health and Big Data
Tuesday 2nd July 2013

Actions

Three areas of research relating to public health and big data were identified for action. Researchers at the meeting are exploring potential for collaboration around these three topics:

- Data collection around health behaviours;
- Additional methodology training for graduate students/early career researchers;
- Opportunities for networking through PublicHealth@Cambridge to aid methodological development to solve data challenges in public health research.

These opportunities will also be advertised across the wider Network and researchers should contact Paula Frampton for more information should they wish to become involved in any of these topics.

A round table discussion around the theme of 'Public Health and Big Data' was held on July 2nd 2013. The discussion was part of the development of the initial working themes of the PublicHealth@Cambridge strategic research network, with the aim of providing a forum to begin scoping the expertise and research interests available across Cambridge as well as the potential research opportunities on the horizon. The overall goal of the meeting was to provide an environment for knowledge exchange, identifying unifying factors and new collaborative partnerships to be explored and taken forward into the future.

Summary of discussion

This note summarises key points made during the discussion, for reference.

The Big Data Forum currently under development at the University as a potential future strategic initiative was introduced and it was noted that there was scope for useful overlap between the Forum and PublicHealth@Cambridge. Good communication had been established between the two initiatives and would continue as the Forum developed.

A number of current large-scale data projects where there could be opportunities for development were highlighted including the e-hospital project at Addenbrooke's, where there were technical considerations to explore, and large genomics data repositories, where it was proving challenging to handle the data in a robust, reliable and transferable form.

Privacy and data security

Cloud computing was discussed in relation to the need for privacy where health and personal data was involved. Within the Computer Laboratory at the University of Cambridge, a long term venture

is ongoing to build new secure systems for large scale computations, providing isolation between operators and users of data. A number of other issues were noted:

- The importance of considering issues of security and privacy at the start of developing a system to avoid costly correction work later.
- Moving large amounts of e.g. genomics data out to the cloud was technically difficult and expensive.
- In the UK, it is not possible to transfer your data protection liability if data is transferred to a different holder, which currently means a requirement for a local computing resource for data storage and handling. A key challenge is then how to allow collaborators and others to analyse data from that location in a secure manner.

To circumvent the data protection issues, a decentralised system where individuals held their own health data on portable devices could be envisaged in the future, partly driven by the potential for personalised medicine. However, for research use, some kind of federated system would still be needed which presented a challenge.

Shifts in cultural attitudes to personal data are likely to lessen the concern about issues of privacy in the future, particularly as the demonstrated utility of large scale data research becomes increasingly apparent. However, the current intermediate situation with a mix of freely available and restricted data presented significant challenges. The European Bioinformatics Institute has robust systems in place for operating within this environment (particularly with genomics data) although it was acknowledged to be difficult and there were problems with scalability.

It was agreed to be important to avoid undermining public trust and confidence in data research. Despite anonymisation of data, linking datasets for research always carried the risk of identifying individuals. The ideal situation would be to obtain full consent for data sharing at the outset of data collection then where this was not possible, to allow people the right to opt out and have their data removed.

The National Cancer Registration Service was a successful example of operating an opt-out system for patients and also provides 'patient portals' where individuals may access their own data. It was highlighted that an extremely low number of opt-outs had ever been requested amongst cancer patients.

For a system of broad consent for future data sharing, including essentially unrestricted, unspecified future uses in public health research, it was agreed that it would be important for the limits of data usage to be explicit and for the responsibilities of data collectors and users to be clearly defined. A (named) role-based access control system could define what groups of people may be able to use the data in the future and a system of discipline for misuse of data would also be required. It was noted that European regulators were currently looking at establishing a requirement for informed consent for every use of data to be required, which would effectively prohibit a broad consent model. Lobbying was underway from a number of UK bodies, including the Wellcome Trust and Academy of Medical Sciences to try to prevent such regulations being established.

Data collection

Restricted ability to access to health data is a major problem for effective public health practice. Data on health behaviours, such as tobacco and alcohol use, is particularly difficult to obtain, as well as data on exposure to particular environments and quantification of those environments. Without appropriate measurements of behaviour, it is not possible to determine how successful particular interventions have been.

Through the Data4Health project in the Cambridge Biomedical Research Centre, work is underway aiming to establish a primary care data resource. Data from the primary care system covers the vast majority of the country's population and includes data on health behaviours such as smoking status. Current systems to extract data from the primary care system are patchy and difficult to use, in part due to privacy/security issues and the lack of suitable IT systems. Public Health England is aiming to try to overcome some of this fragmentation by providing a unified national approach to public health service. Potential links to the MRC e-health centres, such as the one established at UCL, should be explored given their focus on health records linkage.

For health behavioural data, other routes to obtaining data could be useful such as the child measurement programme for obesity research or supermarket shopping data for alcohol-related behaviours. There was interest in sharing methodology between the social and behavioural sciences in this area.

Social network sites such as Facebook could provide large amounts of data but this was not necessarily useful as a freely available source of social data. Because of the nature of users' interaction with the site and its apps, the data would be biased and skewed in unknown and unpredictable ways, rendering the data difficult to use for research.

Work from the Computer Laboratory has taken place using mobile phones to examine the influence of public health advice on people's behaviour during the last 'flu epidemic. This was potentially a very powerful model for future studies given that a great deal of behaviour and movement can be monitored using mobile phone apps. There had been challenges in obtaining medical ethics approval for the work and there were ongoing challenges with including children in the research. There were limitations to consider including the fact that the lowest socio-economic groups were known to be less likely to have access to internet through mobile phones and that by asking people to opt-in to using an app, you bring in further sample bias which can be difficult to control for appropriately.

A bid from Cambridge was currently under preparation for the MRC Medical Bioinformatics call which aimed to address capability gaps in informatics research. The Cambridge bid, linking the Schools of Biological and Clinical Science, had two main components. Firstly to build capability, including hardware as well as improving computational training across graduate and undergraduate students, including pre-clinical medical students. The second piece focusses around more hospital-centred data and the recent £200m investment by Addenbrooke's hospital to make their records entirely electronic. This will be the first entirely electronic system across an Academic Health Science Centre, and is a 10 year project due to start shortly. Given the current very poor interaction with GP records systems, the project would also be seeking low cost ways to integrate defined sections of primary care data with the e-hospital data.

Methodology for data analysis

A number of methodology related issues were discussed.

It was agreed that what was needed for research was not necessarily generation of larger and larger datasets, and an interesting question would be to establish what power was necessary for answering particular research questions to define the scale, richness and complexity of data set required.

As datasets become larger and more complex, there is a need for new methodological developments to extract and integrate information from them. This includes more sophisticated statistical modelling tools and scalable algorithms, and involves work at the interface between statistics, engineering and computer science. For areas such as personalised medicine, there is a need to understand and quantify individual heterogeneity from within a population dataset and to investigate how you can then generalise from that. Improved statistical tools are also important for addressing biases within the data.

To aid the development of data handling tools for public health research, it was agreed that the Network should explore ways of bringing together researchers with data related questions and those seeking to develop methodological tools. Given the multidisciplinary nature of the field, there were 'language' challenges to overcome and examples of networking in this area such as the Cambridge Network of Networks and the stats clinic run by the Statistics Laboratory were suggested as potentially useful models to explore.

Another area where the PublicHealth@Cambridge research network could help in this area was to consider establishing an information repository describing relevant data and tools that were available across the network.

Training/Capability

A need for additional training or short courses covering statistics and programming tools for graduate students across areas relating to public health was proposed. It was evident that although there are a number of examples of training courses around the University, awareness of these across different groups was low and there was little sharing of relevant modules across departments. There was also a lack of high quality training facilities/space. Members of the group agreed to forward examples of existing training courses in the area to Paula Frampton to collate, and then a further working group would be brought together to discuss how best to take this area forward. Training facilities at the EBI were excellent and given their remit across a very broad interpretation of 'life sciences', it was possible to explore working with them to deliver any new training package developed. Others offered the possibility of laboratory placements for practical training or use of data for generic methodology training and development.

It was agreed to be important to aim for joined up thinking across efforts to build big data resources to allow sharing of best practice and prevent repeating any unsuccessful methods. The Big Data initiative being developed across the University may provide a forum to enable knowledge exchange and collective learning but a change in general philosophy and working practices would be required to truly share learning between resources.

Chairs:

Sylvia Richardson	MRC Biostatistics Unit
Simon Tavaré	Cancer Research UK Cambridge Institute and Department of Applied Mathematics and Theoretical Physics

Attendees:

Ross Anderson	Computer Laboratory
Tiffany Bergin	Department of Sociology
James Brenton	Cancer Research UK Cambridge Institute
Goylette Chami	Land Economy
Jon Crowcroft	Computer Laboratory
Daniela DeAngelis	MRC Biostatistics Unit
Emanuele Di Angelantonio	Department of Public Health and Primary Care
Paul Flicek	European Bioinformatics Institute (EBI)
Julian Flowers	Public Health England Knowledge and Intelligence Team East
Paula Frampton	PublicHealth@Cambridge
Simon Frost	Department of Veterinary Medicine
Zoubin Ghahramani	Department of Engineering
Alison Hall	PHG Foundation
Pietro Lio'	Computer Laboratory
Peter McCallum	Cancer Research UK Cambridge Institute
Theresa Marteau	Department of Public Health and Primary Care, Behaviour and Health Research Unit
Rupert Payne	Department of Public Health and Primary Care, Primary Care Unit
Paul Pharoah	Department of Public Health and Primary Care, Cancer Epidemiology
Jem Rashbass	Public Health England National Director for Disease Registration
Augusto Rendon	Haematology
Stefan Scholtes	Judge Business School
Alex Sutherland	Institute of Criminology
Becky Turner	MRC Biostatistics Unit

Apologies:

Afzal Chaudhry	Cambridge University Hospitals NHS Foundation Trust
Robert Glen	Chemistry
Gos Miklem	Genetics
Richard Samworth	Statistical Laboratory
Jackie Scott	Department of Sociology
Michael Simmons	Physics/Cambridge Big Data Forum
David Stillwell	Psychometrics Centre
Anna Vignoles	Faculty of Education
Eiko Yoneki	Computer Laboratory