ROYAL
STATISTICAL
SOCIETY

A Scotland-wide Data Linkage Framework for Statistics & Research consultation

Response of the Royal Statistical Society, June 2012

Introduction

The data linkage concept is simply one manifestation of a growing recognition that ever increasing volumes of data, if ethically and competently linked and analysed, have the potential to bring many benefits. The Royal Statistical Society congratulates the Scottish Executive on this innovative approach to tackling the issues.

The Royal Statistical Society made three recommendations in response to the Royal Society of London's recent consultation on Science as a Public Enterprise that are relevant to this consultation. These were:-

a) Standards of data management need to be sufficiently high that research data **can be** shared for the public good – such as to create new discovery-potential;

b) For transparency, national databases should have a publicly-available protocol which describes the data held, their regular analysis, and any approved record-linkages; and

c) Better public understanding is needed about databases, their linkage, and value-added analyses.

We suggest that the Scottish Executive keeps in touch with the results of the project through Professor Graham Laurie.

Question 1

Are there any benefits of data linkage from statistical and research purposes not sufficiently described?

This is a fairly comprehensive list of benefits based on what has been achieved in Scotland and around the world over the last 30 years. However, this list does not include examples from more contentious areas such as Counter Fraud and Policing. The boundary between administrative use of linked data (for example managing service provision across health and care) and statistical and/or research purposes can be quite fussy, it would benefit from additional examples from sectors other than Health, for example Crime and Neighbourhood Profiles.

The examples given in Benefit 5 are again all focused on Health. Data on costs tend to be a weaker area of public sector data and again examples from other sectors would add to the impact of the Framework, e.g. Geographical Information Systems giving travel times linked to DVLA data. Perhaps other examples could be found from Scandinavian Countries, where aggregate reports from linked information are readily available since a unique social services number is used for all government transactions.

It would also be worth including links to current work at the European Commission. Under the new cross border collaboration directive, the Commission is seeking to create standardisation of registries to ensure linkage across Europe specifically in the area of medical devices to ensure safety for patients, given some high profile failures for implants.

Question 2: challenges and barriers

a) Uncertainty about legalities and public acceptability

Technology has greatly changed since the 2002 NHS Report on Confidentiality and Security of Administrative Data was published. This report addressed the issues of public acceptability of NHS data for research purposes, and was a model for developing the current linkage system between GROs and ISD for longitudinal data.

A market for commercial data has developed over the last decade based on digital and web technologies. News headlines on the legalities of Google photographing every street to produce their mapping tool, produced a ripple of protest at the time, but Google mapping tools remains one of the most used applications on the internet. Mobile phone signals can be used to track movements, as can CCTV, and the recent Suzanne Pilley murder trial demonstrated how digital footprints can be followed to reconstruct a person's movements – unless the individual takes precautions to counter digital surveillance. Public opinion at the time of the trials appeared supportive of such applications. Current generations that rely on mobile technology for all their communications are likely to have a very different approach to the legality and acceptability of use of personal data than their parents and grandparents did, and this needs to be recognised in the framework.

The market for data will continue to grow, indeed the Westminster government is encouraging its development through the Open Data initiative. The Framework needs to be clear on what steps are taken to ensure that individuals cannot be accidently identified from outputs from linked data that were provided for statistical and/or research purposes.

b) Incomplete data or data that cannot be linked

There are always issues regarding incorrect matching, even with unique identifiers (that can be mistyped), and data quality is potentially highly variable. Mismatching and the effects this may have on any inferences are important and as more data sets are linked, will become even more challenging.

Currently users of linked data can specify acceptable levels of sensitivity and specificity. For example if you are linking for operational purposes, then you may want to set a zero probability of wrongly identifying an individual as dead, and this will increase the number of dead people who you still think are alive in the data set. In a research setting, since you will not be contacting individuals then you are likely to set higher tolerance limits for mismatching. When many data sets are linked, further research will be needed to understanding the impact of these errors.

One additional challenge concerns handling the potentially large volumes of data, and the software needed- both for matching and for visualisation. Close collaboration between the statistical and data mining communities may offer some solutions to these challenges.

In large complex data sets, it is essential that inconsistencies can be spotted and checks made. The success of the original work on Medical Record Linkage in Scotland was due to a dedicated team working on the project and building up a level of knowledge and competency that was recognised across the globe.

3. Principles

A further challenge concerns prevention of inadvertent disclosure of information about an identifiable individual of individual information and security. Statistical disclosure control and security are not new challenges but are ones that need to be adequately addressed. These concepts need to be outlined and promoted across the technology sector. There is a need for modernisation of current approaches, for example, blanket cell suppression for non-zero counts under five, should not be the standard technique.

Data must be accompanied by clear and succinct information about its provenance, context, purpose and reliability (metadata). Without this, users will not be able to use the data effectively and are liable to misinterpret or over-interpret. Data must also be provided in a form that is clear and easy to understand and easy to re-use; this applies whether it is in an Excel or csv spreadsheet or in some more sophisticated format. Technical sophistication must not be bought at the price of clarity. It must be explicitly recognised that providing good metadata for linked sources and supplying clear presentation require time and effort.

4. Privacy Advisory Service

The RSS welcomes the idea of a National Privacy Advisory Service especially for tackling situation where “Data custodians may often be unsure whether they can legally and appropriately make data available for linkages and so, to be on the safe side, turn down requests for access to data” – working well it should encourage better secondary use of public data for research purposes.

The nuances of research ethics and privacy are complex. The Scottish Executive has recently supported the four NHS Scotland Teaching Boards to set up their own services. How would a single central committee harness the expertise which is already delegated through SAREC’s and NHS Boards?

The feasibility of a single service need to be carefully considered, since there is a danger that it is too bureaucratic and slow to meet the needs of users. Modern methods of working would need to be adopted, for example establishing a self-help electronic knowledge repository with web support – a WIKI for Privacy Advice. It’s already an area of expertise for Scotland and it would be beneficial to share that knowledge across the UK.

We would wish to be consulted on firmer proposals.

5. National Data Linkage Centre

Recognising that the proposed model is based on what was established over 15 years ago, we would suggest that thought is given on how processes and procedures can be modernised. This was an underlying concept for the SHIP when funded over 5 years ago, and hopefully some of the results from this research might be quickly implemented.

Replacing the Census is an essential outcome of this work, and we are pleased to see links with the ONS Beyond 2011 team and the development of a population spine. Public sector technology has tended to be slow to use web applications because of security concerns and we'd like to see some innovative and forward thinking approaches used by the proposed centre to deliver a timely service.

We would be happy to respond to proposals for a Scottish Data Linkage Centre.