Towards a Health Research Remote Data Laboratory (HRRDL) - some implementation details

For the immense value of NHS patients' data to research to be realised in ways that will preserve and enhance public trust demands the highest protections. Not only physical and operational security measures, but uncompromisingly robust and transparent processes - including audit and governance - must be engineered into the system. This document focuses on the frameworks that must interact to support and meet the governance and operational requirements of the public and researchers.

There are many examples, within both the UK Government and internationally, of managed facilities for research access to sensitive data by researchers: the ONS Virtual Microdata Laboratory for business level data and individual level census data; the HMRC Data Lab; the National Opinion Research Center at the University of Chicago; the US Census data laboratory; or the Biobanks. Models from a number of organisations might be replicated and adapted for HSCIC purposes but, for reasons of organisational similarity, we have based this document in the main part on the ONS Virtual Microdata and Business Data Laboratories.

In this document we make reference to the US standard for a "Secure Compartmented Information Facility"[1], a standard for facilities in which sensitive intelligence data can be discussed. In a similar vein in the UK, the Administrative Data Liaison Service (ADLS) is funding the installation of "safe pods" in universities for remote access to sensitive administrative data over a remote network which meet various technical and physical standards for access[2].

The framework outlined below suggests changes over and above that which is done by the ONS Virtual Microdata and Business Data Laboratories, which is to date the most well-established similar facility incorporating remote access. ONS VML / BDL has solved a number of scalability issues balancing the demand for data accessibility, whilst preserving the highest levels of both technological and non-technological protection.

There are 7 areas that the HRRDL must consider and integrate; the six ONS areas, plus 'safe hosts' for remote access:

- The nature of the data

[1] https://www.fas.org/irp/dni/icd/ics-705-1.pdf
[2] http://www.adls.ac.uk/adls-resources/esrc-sensitive-data-secure-rooms/

- Safe Statistical Outputs
- Safe Research / Analysis
- Safe Researchers
- Safe Organisations
- Safe Settings
- Safe Hosts

N.B. Technical measures used to ensure security commitments are met and specific software choices are out of scope for this document, but none of these are unsolved problems.

The nature of the data

The data within the HRRDL will be mostly unperturbed data on all individuals who have not withdrawn informed consent for their data to be used for secondary purposes, including commissioning. Patients whose data are used must have full transparency on those projects which have access to it. The definition and operation of the 9Nu4 dissent code will need reworking as HRRDL data does not leave the HSCIC, though it may still be used for secondary purposes by those outside the HSCIC.

As the data are unperturbed and potentially highly sensitive, all the other security measures are designed to ensure that only authorised researchers on authorised projects in authorised institutions get access, in a way which generates public confidence and trust in *bona fide* research. Once research has been conducted, only safe outputs can be withdrawn from the system via HSCIC approval and assessment process.

The data itself, however, cannot be “made safe”. The data in the HRRDL is fundamentally identifiable, due to the richness of linked patient data necessary for research. Given that the precautions required to deal with this will be applied to all of the data, there need not necessarily be systemic prohibitions on data analysis of so-called “sensitive conditions” - e.g. mental health, STIs, HIV/AIDS - so long as the extra governance processes are in place to ensure that use for those projects is appropriate. (Such processes are outside the scope of HRRDL itself, but HRRDL would help facilitate a safer solution in which patients can have greater confidence.). The HRRDL should only provide access to the data requested for analysis.

Data minimisation processes should be applied in every case, so that each safe researcher only gets access to the data required for their safe research. If a researcher doesn’t need psychiatric data, it shouldn’t have access to it.

All data are held within a safe “remote” environment, to which safe researchers have no physical access, and which has no access to the internet. Data input and output is via drives managed by HSCIC. All use of such systems should be audited, to ensure that the powerful tools used for research are not abused.

Safe Statistical Outputs

The primary purpose of the HRRDL in the service of research is the production of safe statistical outputs, which can be removed from the HRRDL environment and used for a wide range of purposes and benefits.

To ensure public benefit and confidence, in line with Government mandates for research using public resources, all publications containing material (safe statistical outputs) removed from the HRRDL must be made fully Open Access.[3]

To remove statistical information from the HRRDL, it must be extracted and approved as being non-disclosive by HSCIC staff following defined Statistical Disclosure Control policies. Existing processes have evolved over a decade of experience in corner cases and potential issues as detailed in the academic literature[4] [5], and those experiences must be built upon. Those techniques will not be unique to the HRRDL, but the nature of the HRRDL data may lead to new Statistical Disclosure Control requirements.

Safe Research

Safe Research projects are the fundamental operational block of the HRRDL.

All research projects must be approved by the Confidentiality Advisory Group (CAG), be from safe researchers, be done in safe settings, with any other restrictions required by CAG. This area otherwise follows the ONS VML model, with CAG replacing the Microdata Release Panel.

---

[3] https://www.gov.uk/government/speeches/open-access-research

[4] e.g. Ritchie F. (2013) Output-based disclosure control for regressions. http://www2.uwe.ac.uk/faculties/BBS/BUS/Research/economics2012/1209.pdf

[5] Brandt M., Franconi L., Guerke C., Hundepool A., Lucarelli M., Mol J., Ritchie F., Seri G. and Welpton R. (2010),Guidelines for the checking of output based on microdata research, Final report of ESSnet sub-group on output SDC http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

All research must also have passed through an ethics review policy at the researcher's host organisation, which meets the NIHR ethics requirements. All projects must appear in a public register, including the project title, plain English description, data requested and details of the safe organisation. Once projects have produced outputs, these should also be linked from the register - in fact, Open Access compliance would make this a requirement.

Safe research must be in the public interest, and all results be published in fully Open Access publications. No patient should have to pay to read the outputs of research done on their data.

Safe Analysis

As HRRDL class facilities move beyond research use only, and into other, radically different areas, such as invoice reconciliation. "Safe Analysis" is the more generic term for "safe research" which covers the specific type of analysis done in any particular implementation of a HRRDL class facility. While most "safe analysis" should be capable of being conducted as "safe research", only the form of "safe analysis" that an implementation of a HRRDL facility was designed for should be conducted in that facility.

In short, "safe analysis" is answering repeated targeted questions; "safe research" is finding those questions, and potentially developing data products for them.

Safe Researchers[6]

Safe researchers have the skills and knowledge to use the data, to work on questions for which no other less sensitive data are suitable, and to operate within a safe setting. This section follows the ONS VML and ESRC's ADRN model for "approved researchers". Such a model ensures that researchers are incentivised to work with the HRRDL rather than seeing it as a burden to evade where possible.

Researchers may use any safe host willing to grant access to their safe setting.

Safe Organisations (of safe researchers)

[6] Desai T. and Ritchie F. (2010) "Effective researcher management", in Work session on statistical data confidentiality 2009; Eurostathttp://www.unece.org/stats/documents/ece/ces/ge.46/2009/wp.15.e.pdf

An organisation (generally the employer, or education institution) that acts as the legal entity that is responsible and legally accountable for a safe researcher. Such organisations must confirm that the individual has had suitable training to be able to handle the data safely and knowledgably, and to be responsible for their own actions.

A Board-level member of the organisation must counter-sign every application from one of its researchers to confirm that they will accept this responsibility, and the consequences that come from misuse. Due to the hierarchical nature of organisations, and for very large organisations, it may be that a single legal entity may contain multiple "safe organisations". A large University, for example, may have nominal "safe organisations" for each of its constituent faculties or divisions, i.e. larger than a single department. Such signatories need not necessarily be as high up the 'parent' hierarchy as to require the signature of the Vice-Chancellor of a University, or CEO of a large company in a commercial context, but the counter-signatory must hold a sufficiently senior post.

To qualify as a safe researcher, a person must have an organisation willing to be the "safe organisation" legally responsible for them.

Safe Settings

Primarily at HSCIC, but potentially located in other organisations, safe settings are physical spaces within HSCIC or secure spaces in other locations, both of which would have thin client connections over protected network links to HSCIC's server facility. Data never leaves HSCIC's server facility other than through the safe statistical output process, or via display through thin clients.

Utilising N3, safe settings (modelled on the ESRC "Safe Pods"[7]) need not be solely within the HSCIC physical environment, if a "safe organisation" is able to meet every standard and requirement set by relevant bodies.

Every safe setting will have a safe host that is accountable for every action taken within its safe setting - as distinct from safe organisations, which are responsible for the actions of their safe researchers within any safe setting those researchers may use.

Safe Hosts (of safe settings)

[7] http://www.adls.ac.uk/adls-resources/esrc-sensitive-data-secure-rooms/

Safe hosts are safe organisations which in addition operate their own safe settings with remote access to the HRRDL, on an equivalent basis to the HSCIC safe setting.

In order to receive approval to host its own safe setting, an organisation must demonstrate it can meet every requirement for running a safe setting equivalent to the HSCIC safe setting. In addition to technical measures, the safe host would for example be responsible for nominating or appointing a named individual to maintain an audit log and keep records of every access to their safe setting.

Rules should be determined by Independent Information Governance Oversight and CAG.
“Safe hosts” are responsible for “safe researchers” in good standing from their own organisation (i.e. where the approval would be countersigned by the same Board-level member), but there is no requirement for a safe host to accept safe researchers from outside their safe organisation, unless they wish to.

Organisations with “safe host” status will routinely act as host for their own “safe researchers”. Where an organisation is not willing to act as a host for one of its own researchers, this should call into question the relevant “safe researcher” or “safe host” status. That is not to say that safe organisations/safe hosts may not make reciprocal legally transparent agreements regarding access. In effect, HSCIC will be forming a version of such an agreement each time it approves a safe organisation to become a safe host.

Subject to booking and fair use, HSCIC must act as the safe host for any safe researcher who wishes to use the facilities at HSCIC.

The ONS VML does not have facilities outside of National Statistics Offices, due to the complexities of the Government Secure Intranet. With the increased reach due to the flexibility and protections available over N3, the HRRDL could treat remote users with secured links over N3 as ‘first class citizens’, subject to them meeting all standards and requirements.

Sam Smith
medConfidential
April 2014
The Safe Analysis section was added in July 2014