How can the Secure Data Facility support Risk Stratification?

Risk stratification is roughly separable into two highly distinct activities. Firstly, design of strata - the creation of models, interactions, treatment effects; and secondly the application of those models to a specific area and population, which is far more prevalent than their creation.

Neither of these are direct care, and attempting to use identifiable data with a direct care justification is questionable for the reasons NHS England explain in the past attempts to create Accredited Safe Havens[1].

Deconflating the different aspects of risk stratification, to separate out the various data needs for the different users at different steps in the process, the differing purposes, and hence the required different data products, allows for much clearer data flows, audit and consent.

This can be safe, consensual and transparent; the choice is whether or not it will be.

Generating Risk stratification models

The generation of models, the testing of hypotheses to form them, the development of publications justifying them, is research. The process for a model developed by MRC is almost indistinguishable from a model for the Mayor of London's office. The models should be subject to publication, peer-review, etc. As such, that development should take place within a data laboratory, on the rich, detailed patient data required to distinguish between (or at least, potentially investigate) correlation and causality. As development works on models of a population, it can use fully consented data

As part of creating a risk stratification model within a safe setting, part of the requirement can be to design the data specification required for the model to be run. Those outputs should be aggregated, and area (CCG) specific, and can then be made approved by HSCIC before being copied to a different facility for access by users of a risk stratification model.

Where a model needs to include data from other sources, and many of them will, the HSCIC's safe setting, when operational, should be welcomed into the club of equivalent government safe settings (ONS, MoJ, HMRC, DWP) with cross-departmental acceptance of standards, such that data of lower sensitivity could be available in the HSCIC safe setting for that particular project. This is established process, and while it can not be confirmed before such a laboratory is operational, it would be a cause for public concern if HSCIC’s design for the Facility did not exceed the standards of the other facilities available in Government.

[1] such as in the 2014 DH ASH consultation.

Using Risk stratification models

Most organisations will not generate their own models from scratch, but apply models developed elsewhere to their own areas. Even organisations which build their own models will likely draw these distinctions internally, with separate research and commissioning functions which interact but are operationally distinct.

Models developed for stratification can include the specification of the data they need for any particular area of interest. Such specifications will be aggregated statistical outputs, with up to some number of dimensions of data tabulated. While those tables will include small numbers, and so can not be used in uncontrolled environments, they are aggregated figures about an area, and should not be individual level data.

As a result, there will still need to be a controlled environment for an organisation to look at the risk stratification data that they are interested in, consider different iterations and boundaries, based on the specification of the model using constraints designed by the model's designer. No model allows an unlimited amount of variation, which is why different models are required for different purposes.

However, models can be used within a safe setting appropriate to the data input to the model (ie the custom data product designed by the model creator), which will be of much lower sensitivity than individual level patient data. The primary output of the application of a model to an area is not a list of patients, but a set of types and thresholds of treatments and counts, to enter into other documents - it will likely be an aggregated dataset in several dimensions without small number suppression having been applied.

Where appropriate, after agreement, those types and thresholds can then be run by GPs against their full patient dataset, which will include individuals who have opted out of secondary uses for data sharing. They are not included in the generation of the model, but by careful and correct data design.

Such a design has been discussed as part of the emerging “big data” approach to health data. Whereas the question of general research requires access to individual-level data[2], specific research questions can be answered using particular data products designed for the purpose.

Developers of models which are heavily used in this framework, may wish to enhance their tools with the full integration offered by the providers, with the additional efforts that requires.

medConfidential, January 2015

[2] DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. http://ije.oxfordjournals.org/content/39/5/1372.full.pdf+html