Document filename: Data Linkage & Data Quality Sub Group ToR v1.1			
Programme	HSCIC Data Pseudonymisation Review	**Project**	HSCIC Data Pseudonymisation Review
Document Reference			
Project Manager	Matt Spencer	**Status**	Draft
Owner	Chris Roebuck	**Version**	1.1
Author	Chris Roebuck	**Version issue date**	24/11//2014

HSCIC Data Pseudonymisation Review - Data Linkage & Data Quality Sub-group

Terms of Reference

Contents

1. Purpose of the Sub-group

The Data Linkage & Data Quality Sub-group will advise the Steering Group on the impact of different approaches to pseudonymisation on data linkage and data quality. It will look to provide quantitative evidence in all possible instances, such as linkage rates. It will consider the impact both to HSCIC current approaches and proposed future approaches to data linkage and data quality analysis.

2. Context of Data Linkage and Data Quality for Pseudonymisation

There is a need to link datasets for research and other purposes that will lead to benefits for patients. The Data Linkage & Data Quality sub-group will consider how this would be affected by different pseudonymisation approaches.

Given the large range of different datasets that could in theory be linked together, there is a need for the group to consider the technical feasibility of how this can be achieved under different pseudonymisation approaches. The group should consider different sets of data with different ranges and numbers of data submitters. It should prioritise investigations for those datasets that are used most frequently for linkage and have the greatest level of interest around them, such as Hospital Episode Statistics and primary care data. There also may be cases in which linkage is approved between health data and data from sources such as social care that do not have a reliable NHS number – the sub-group should consider the impact of different pseudonymisation approaches on these.

The sub group should also look at how potential enhancements to the HSCIC's methodology would be affected by the different pseudonymisation approaches. For example, the HSCIC currently performs deterministic matching, but the impact of probabilistic matching should be explored. Equally the potential for linkage to a patient index either pseudonymised before or after linkage should also be explored.

The sub-group should look at a set of studies to determine the impact of pseudonymisation before linkage and pseudonymisation after linkage on the quality of that data linkage. It should look at several linkages between sets of data of varying degrees of quality and should also look at the impact on probabilistic and deterministic matching, essentially covering all elements in the following quadrant: It should explore recent approaches to incorporating record similarity measures within the pseudonymisation procedure, such as Bloom filtering and similarity score prediction methods.

	Deterministic matching	Probabilistic matching
Pseudonymisation before linkage		
Pseudonymisation after linkage		

The subgroup will also assess the extent to which the different approaches to pseudonymisation enable the HSCIC to deliver its statutory functions around reporting on data quality and how they could enable the following activities:

- Assess the quality of the data used for pseudonymisation against mandated standards from the NHS data dictionary.

- Make the results of those assessments available to all prospective users of the pseudonymised data to inform their view of its fitness for their purposes: If pseudonymisation were carried out at source, each submitting organisation would be responsible for:
    - Carrying out the data quality assessments on the data items used in the pseudonymisation process
    - Producing the data quality reports from those assessments
- These assessments would need to be produced in a standard form and collated centrally and published on the HSCIC website for onward use by customers to enable them to assess whether the quality of their data is suitable for their purposes

3. Work and associated outputs from the Sub-group

The Sub-group will:

- Deliver studies to the steering group to review the impact of pseudonymisation before and after linkage on deterministic and probabilistic approaches to matching and the resultant levels of linkages and quality. This will be delivered in priority order by:
    - working with the Clinical Practice Research Datalink (CPRD)and HES datasets to evaluate the impact of different pseudonymisation approaches on deterministic and probabilistic linkage and any other that IG approvals could provide agreement to make available for the sub-group evaluations and produce summary report for steering group
    - Evaluate the availability and data quality of identifiers on a range of other datasets that could be linked for the benefit of care in England to gain an understanding of linkage potential under different methods and produce summary report for steering group, for example mortality and cancer registration data from ONS.
    - Perform any other data linkage studies needed to gain an understanding of impact of pseudonymisation on linkage and produce summary report for steering group
    - Review any technical developments that would affect linkage, pseudonymisation and their interaction, including Bloom filtering and similarity scoring and summarise these to steering group

- Perform desktop exercise to gain understanding of how HSCIC's data quality reporting requirements could be delivered under the pseudonymisation models, coordinating with the pseudonymisation at source group as necessary to understand this for pseudonymisation at source
- Devise a study to test ability of different types of organisation from across the Health and Social Care system to pseudonymise consistently where required in order to enable data linkage, coordinating with the pseudonymisation at source group as necessary
- Respond to requests/direction from the Steering Group on investigating topics arising from the work of the Steering Group
- Report to the Steering Group on the above topics and activities as necessary.

The Sub-group will not itself

- Look at approvals process or legal basis for linking data, as this is covered outside the pseudonymisation review, for example through Confidential Advisory Group approval process
- Weigh up the impact on data linkage and quality with other factors such as impact on data security and cost of different pseudonymisation approaches. These other factors will be examined by the pseudonymisation at source sub group, with the steering group weighing them up

4. Membership of the Sub-group

The Sub-group shall comprise members of the Steering Group and relevant experts that the Sub-group feels it needs to co-opt to fulfil its remit. This is subject to the Review's resource constraints. The membership of the Sub-group and the need for co-opted members has already been agreed at its first meeting on 3rd July 2014.

The members of the Sub-group are

• Harvey Goldstein	University College London and University of Bristol
• Xanthe Hannah	NHS England
• Julia Hippisley-Cox	Nottingham University
• Chris Roebuck (Chair)	HSCIC
• Ralph Sullivan	RCGP
• Tim Williams	CPRD
• Peter Jones	ONS
• Emma Gordon	ONS
• Sean McPhail	PHE
• Anthony Chuter	Patient Representative
• John Sharp	HSCIC

5. Method of Operation of the Sub-group

The Sub-group will have an initial face-to-face meeting to agree an outline 'programme of work' of tasks to undertake the activities outlined above.

This meeting will be followed by work undertaken as necessary to fit into timetables agreed with the Project Manager for reporting to the Steering Group within its schedule of meetings.

Subsequent meetings are expected to be mainly electronically based, except where significant issues would benefit from resolution through face-to-face meetings.

The HSCIC Pseudonymisation Review SharePoint site should be used as the repository for documents generated in the work.

The modus operandi on reports and decision-making (for recommendations to the Steering Group) will follow the rules set out in the Terms of Reference for the Steering Group.

The sub-group is expected to follow the Steering Group ToR in regards to Section 3 'Standards of conduct for Steering Group members'.