HSCIC Data Pseudonymisation Review – Final Report

Version 1.0 | 24 March 2017

Document Management

Revision History

Version	Date	Summary of Changes
0.1	March 2015	Draft version created
0.2	March 2015	High level structure presented to March Steering group
0.3-0.5	March 2015	Internal team review of structure against Pseudo Review 2014 Interim Report
0.6	July 2015	High level structure outlined to July Steering Group
0.7-0.9	July- Sept 2015	Draft Final Report prepared for Steering group review
0.10-0.13	Jan-April 2016	Draft incorporating comments from Steering Group members
0.14	May 2016	Rewrite incorporating comments from Steering Group members
0.15	July 2016	Rewrite incorporating comments from final Steering Group review
0.16	February 2017	Changes applied following comments received from Steering group member
1.0	March 2017	Changes applied following Steering group review and up issued to approved

Reviewers

This document must be reviewed by the following people: author to indicate reviewers

Reviewer name	Title / Responsibility	Date	Version
HSCIC Pseudonymisation Review Steering Group		03/03/2017	V0.16

Approved by

This document must be approved by the following people: author to indicate approvers

Name	Signature	Title	Date	Version
HSCIC Pseudonymisation Review Steering Group			24/03/2017	1.0

Glossary of Terms

Term / Abbreviation	What it stands for
ASH	Accredited Safe Haven
CAG	Confidentiality Advisory Group
CCG	Clinical Commissioning Group
CDS	Commissioning Data Sets
CP	Central Pseudonymisation
CPRD	Clinical Practice Research Datalink
CQC	Care Quality Commission
CSU	Commissioning Support Unit
DAAG	Data Access Advisory Group
DARS	Data Access Request Service
DoB	Date of Birth

DPA	Data Protection Act 1998
DSCRO	Data Service for Commissioners Regional Office
DSfC	Data Services for Commissioners
DSS	Data Stewards Service
EMT	Executive Management Team
EU	European Union
GP	General Practitioner
GPES	GP Extraction Service
HES	Hospital Episodes Statistics
HRA	Health Research Authority
HSCIC	Health and Social Care Information Centre
ICO	Information Commissioner's Office
IG	Information Governance
IGARD	Independent Group Advising on the Release of Data
IIGOP	Independent Information Governance Oversight Panel
MRIS	Medical Research Information Service
NHS	National Health Service
NHSE	NHS England
ONS	Office for National Statistics
P@S	Pseudonymisation at Source
PAF	Postcode Address File
PAS	Patient Administration System
PbR	Payment by Results
PHE	Public Health England
QOF	Quality Outcomes Framework
SAIL	Secure Anonymised Information Linkage
SAR	Subject Access Request
SIRI	Serious Incident Requiring Investigation
SUS	Secondary Uses Service
UCL	University College London
UKAN	UK Anonymisation Network

Contents

Executive Summary

This report summarises the evidence and perspectives presented to the Reviews Steering Group by a wide range of stakeholders during the HSCIC Data Pseudonymisation Review.

It includes recommendations to the Health and Social Care Information Centre (HSCIC) Executive Management Team (EMT) in relation to the HSCIC approach to Pseudonymisation.

An initial review was undertaken from November 2013. This reviewed the use of pseudonymisation in respect of data in transmission to, received, held and disseminated by the HSCIC. The output of that review was an interim report published in July 2014[1].

The interim report set out the need for a next stage review, to look in particular at three broad options for pseudonymisation of data collected by the HSCIC:

- pseudonymisation of data centrally (after receipt by the HSCIC)[2];
- pseudonymisation of data at source (before disclosure to the HSCIC);
- a mixture of pseudonymisation at source and pseudonymisation centrally

This next stage review commenced in August 2014 with the aim of reporting recommendations on approaches to pseudonymisation by the HSCIC for consideration by the HSCIC EMT.

The HSCIC Data Pseudonymisation Review has been overseen by an Independent Steering Group, with membership made up of experts representing a wide spectrum of views and interests on the subject area, the group also included a patient representative.

The Steering Group considered evidence from a wide range of stakeholders including data providers, customers who receive data from the HSCIC, Arms-Length Bodies and suppliers of systems.

It is clear that pseudonymisation is a very complex and emotive subject area, often with highly polarised views depending upon the standpoint of that stakeholder. This report summarises these different perspectives.

Whilst the original intention from the Interim Report was to consider the three broad options for pseudonymisation it became clear that a one size fits all assessment would not be feasible. As a result the recommendations presented in this Final Report instead focus on the options on a per data flow basis whilst taking into account the broad options for pseudonymisation.

This review has taken place over a period of considerable change to the data sharing, privacy and confidentiality landscape including:

[1] https://www.gov.uk/government/publications/data-pseudonymisation-review

[2] Where the HSCIC currently employs pseudonymisation, it is performed centrally, typically after data quality and data linkage work.

- the HSCIC commitment to honouring patient objections,
- adoption of the recommendations of the Partridge Review into data releases by the NHS Information Centre,
- the Care Act 2014,
- formalising the role of the Confidentiality Advisory Group (CAG) and
- the National Data Guardian review consultation on guidelines for data security of patients' confidential data including a new Consent/Opt-out model.
- increased awareness of European legislation including the development of a new General Data Protection Regulation (GDPR) and the potential impact of this on pseudonymisation.

This report should be read in the context of the changes to the wider data sharing, privacy and confidentiality landscape.

The review has found that pseudonymisation can be an effective technique for reducing the risks associated with the sharing a patients confidential data, and whilst not fully eliminating the risk, when used in combination with other safeguards can render data effectively anonymised.

It also considers the role of the HSCIC to collect, process, analyse and publish or otherwise disseminate health and social care data, including identifiable data, using it's statutory powers under the Health and Social Care Act 2012. This role includes recognition as the Safe Haven for health and social care information.

During the review substantial evidence has been submitted by a diverse range of stakeholders and considered by the Review Steering Group and Sub-Groups. This evidence has been captured in a series of sub-group deliverables from which a number of key assumptions (see Appendix 2) and key findings (see Appendix 3) have emerged.

Based upon this evidence, the Review Steering Group makes the following recommendations to the HSCIC Executive Management Team.

No.	Topic	Recommendation
1	Public Confidence	*The HSCIC need to build public confidence by continuing to address public and professional concerns through a two-way dialogue. This will include being transparent about the data that it collects and processes, how it is kept securely, and whom the data is shared with, for what purpose and on what legal basis and how their confidential data is protected. It should also inform patients how they can express preferences to how their data is used.* *The HSCIC should list all the data sets it collects and processes, as well as all data it releases, on its website in an easily searchable form. This will help the public easily determine what data the HSCIC is likely to hold on them as an individual and who has access to that data in identifiable form.*

No.	Topic	Recommendation
1a	Communicating Benefits of data sharing	*The HSCIC should identify and communicate the benefits to the patients and the wider health and care associated with the collection, analysis, publication and other dissemination of health and care data, as well as the risk and means used to minimise it, including personal, sensitive and confidential data*
2	Ensuring use of pseudonymised data is appropriate	*Pseudonymisation on its own is often insufficient to protect the confidentiality of patient data. The HSCIC should provide training to HSCIC staff, the wider NHS and customers which covers the organisational, legal and technical implications of using pseudonymised, data, including the risks involved and legal penalties, prior to the sharing of data*
3	Irreversible Pseudonymisation	*The HSCIC should apply pseudonymisation which is irreversible by the recipient unless there is a legitimate health related reason and appropriate organisational, technical and legal measures in place for the data to be re-identified. The HSCIC should by default own and control the pseudonymisation keys or lookup tables in cases where it disseminates pseudonymised data.* *The HSCIC should develop a policy around other types of key management requested in its dissemination of pseudonymised data and the circumstances under which it would consider such disseminations to be identifiable, in liaison with CAG.* *There should be transparency around the type of pseudonymisation applied, for example, in data release register.* http://www.hscic.gov.uk/dataregister
4	Establish Centre of Expertise and Capabilities	*The HSCIC should develop an internal centre of expertise, which can provide best practice advice and guidance in relation to the* ***de-identification*** *of data, including pseudonymisation for itself and the wider NHS. This would include the development of relevant standards* *As a priority it should:* • *Develop specific criteria against which individual data collections by the HSCIC can be evaluated for the optimum usage of pseudonymisation in terms of the purpose of the data collection and respecting privacy.* • *Develop existing techniques for anonymisation to increase the utility of the data once its disseminated* • *Communicate to the public the results of this activity in understandable terms.*
5	Developments in privacy enhancing technique and technologies	*The HSCIC should consider how best to review and appraise developments in privacy enhancing and data security techniques and technologies on an ongoing basis to ensure that it adopts them at the earliest opportunity where appropriate. This includes technologies to reduce the flow of identifiable data to the minimum required for specific purposes, in line with requirements of the Data Protection Act.*

No.	Topic	Recommendation
6	Existing National data flows to HSCIC	*Existing National flows of identifiable data into the HSCIC should be subject to a rolling programme of regular review against specified criteria to ensure data flows in the least identifiable form necessary to meet the purpose.* *Each data flow should be reviewed in the light of legislative changes or significant technical developments, or if the requirements around individual flows change.* *It is accepted that there are specific purposes for which the HSCIC needs to collect and process identifiable data for example to perform probabilistic data linkage or when patients have consented to specific research e.g. BioBank.*
7	Segregation of patient identifiers from activity within HSCIC	*Where present on inbound data Patient identifiers should be segregated from remaining data upon landing within the HSCIC. Access to Patient Identifiers should be restricted to the minimum number of staff that absolutely requires access to these items for specific discrete purposes, with non-identifiable alternatives derived for analysis purposes e.g. Age rather than Date of Birth.* *Individuals should not routinely be able to access both patient identifiers and activity data. In the exceptional circumstances where access to both are required strict protocols must be adhered to including Senior Level approval* *Access to data should be fully controlled, logged, audited and monitored on a continuous basis to assure compliance.*
8	New National data flows to HSCIC	*Any new national data flow should be subject to IG review, through a Privacy Impact Assessment, and would involve groups of the relevant data subjects and controllers where required. This should consider whether aggregate, fully anonymised or data pseudonymised at source or identifiable data could be used to meet the business objectives and realise the benefits to health and care, using data with the minimum risk of re-identification, to meet that purpose.*
9	Pseudonymisation at Source Proof of Concept	*At the point that a new national data flow into the HSCIC is identified where the benefits could be fully met under a pseudonymisation at source model a Proof of Concept should be initiated to prove the efficacy of this approach in relation to the HSCIC operating model.*
10	Improving support to privacy of patient data	*The HSCIC should provide standards and tools to support the self-assessment and audit of the techniques to create and use pseudonymised and de-identified data across the health and social care system.* *The HSCIC should provide advice on local flows that do not currently involve the HSCIC when requested to do so.*

1 Introduction

This report summarises the evidence and perspectives presented to the Reviews Steering Group by a wide range of stakeholders during the HSCIC Data Pseudonymisation Review.

It includes recommendations to the Health and Social Care Information Centre (HSCIC) Executive Management Team (EMT) in relation to the HSCIC approach to Pseudonymisation.

An initial review was undertaken from November 2013. This reviewed the use of pseudonymisation in respect of data in transmission to, received, held and disseminated by the HSCIC. The output of that review was an interim report published in July 2014[3].

https://www.gov.uk/government/publications/data-pseudonymisation-review

The interim report set out the need for a next stage review, to look in particular at three broad options for pseudonymisation of data collected by the HSCIC:

- pseudonymisation of data centrally (after receipt by the HSCIC)[4];
- pseudonymisation of data at source (before disclosure to the HSCIC);
- a mixture of pseudonymisation at source and pseudonymisation centrally

This next stage review commenced in August 2014 with the aim of reporting recommendations on approaches to pseudonymisation by the HSCIC for consideration by the HSCIC EMT.

The HSCIC Data Pseudonymisation Review has been overseen by an Independent Steering Group, with membership made up of experts representing a wide spectrum of views and interests on the subject area, the group also included a patient representative.

The Steering Group considered evidence from a wide range of stakeholders including data providers, customers who receive data from the HSCIC, Arms-Length Bodies and suppliers of systems.

All members of the Steering Group were committed to achieving the aims of protecting the privacy and rights of patients in the use of their personal data, respecting the ethics of clinicians who are the data controllers and ensuring public good by maximising the benefits to health and care derived from the use of such data.

In some cases members presented points and counter-points which challenged the group to achieve a balance between these dual aims. Despite this the Steering Group did reach a common understanding in respect to many areas discussed within the report, and whilst this

[3] https://www.gov.uk/government/publications/data-pseudonymisation-review

[4] Where the HSCIC currently employs pseudonymisation, it is performed centrally, typically after data quality and data linkage work.

does not represent consensus on all issues, sufficient common ground was found to agree recommendations for the HSCIC EMT to consider.

It is clear that pseudonymisation, in its wider context, is a very complex and emotive subject area, often with highly polarised views depending upon the standpoint of that stakeholder. This report summarises these different perspectives.

Whilst the original intention from the Interim Report was to consider the three broad options for pseudonymisation it became clear that a one size fits all assessment would not be feasible. As a result the recommendations presented in this Final Report instead focus on the options on a per data flow basis whilst taking into account the broad options for pseudonymisation.

This review has taken place over a period of considerable change to the data sharing, privacy and confidentiality landscape. In many cases these changes have impacted the wider context for the Review or, in some cases, have directly impinged on the Review.

These changes include:

- the public and parliamentary interest in care.data and the developments in plans around its implementation
- formalising the right of patients to object to the use of their data and the subsequent commitment by the HSCIC to honour patient objections from early 2016
- strengthening of the HSCIC's data management and release practices including adoption of recommendations from the Partridge Review of data releases by the NHS Information Centre (the predecessor to the HSCIC)
- the need for detailed patient level data to support the local and national care service commissioning processes. This involved the implementation for Stage 1 Accredited Safe Havens (ASHs), the expectation for regulations to be developed to support further ASHs and developments around Data Services for Commissioning (DSfC) by NHS England
- systems developments have been taking place within the HSCIC in order for it to meet its obligations; some of which involve the need for pseudonymisation of data
- the Care Quality Commission (CQC) review of standards of data security for patients' confidential data across the NHS including clear guidelines for the protection of personal data as set out by the National Data Guardian
- Care Act 2014 restrictions on dissemination of data by the HSCIC other than for the provision of health care or adult social care, or the promotion of health
- regulations defining the role of the Health Research Authority's (HRA) Confidentiality Advisory Group (CAG)

This report should be read in the context of the changes to the wider data sharing, privacy and confidentiality landscape.

Context

Pseudonymisation

Pseudonymisation is only one of many techniques for reducing the risks associated with personal data,

The ICO Code of Practice on Anonymisation provides guidance on best practice techniques for anonymising data.

https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Pseudonymisation is a technique used to replace direct patient identifiers within data, such as NHS Number, Date of Birth or Postcode, with a pseudonym which does not reveal the individuals real world identity. This reduces the risk of re-identification of a patient's data and protects their privacy and confidentiality whilst still enabling the data to be used for a variety of purposes including data linkage.

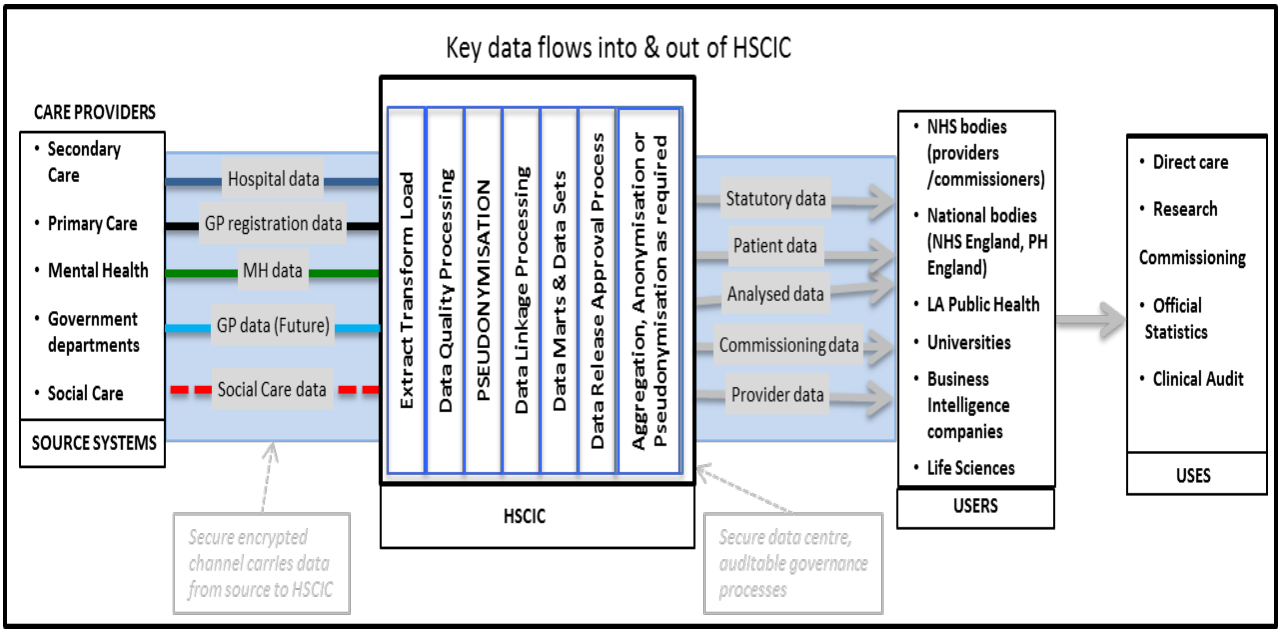
The Data Protection Act 1998 (DPA) only applies to personal identifiable data. It does not apply to data that offers an insignificant risk of identifying the data subject(s) involved. If pseudonymisation as an anonymisation technique does not sufficiently reduce the risk of re-identification, the data concerned must be treated as personal identifiable data.

Chapter 7 of the ICO's Anonymisation Code of Practice outlines limited access safeguards which can be used to reduce the risk of re-identification, such as conditions for the use of data and penalties where conditions have been breached. Where these additional safeguards are in place pseudonymised data may be considered to be de-identified in line with the ICO code of Practice on Anonymisation.

Remit of the Health and Social Care Information Centre

The Health and Social Care Information Centre has powers under the Health and Social Care Act 2012 to collect, process, publish and disseminate data including patient identifiable data. It has been established as the Accredited Safe Haven for the NHS.

Figure 1 below describes the flows of patient data into and out of the HSCIC.



N.B. The HSCIC is also required to disclose information where there is a statutory and mandatory legal basis to do so e.g. requests to the police under section 29 of the Data Protection Act or as a result of a Court Order.

Figure. 1: Key data flows into and out of the HSCIC

Typically national flows of data into the Health and Social Care Information Centre (HSCIC) are in identifiable form, principally containing direct patient identifiers such as NHS number, Date of Birth and Postcode. In a small number of instances data collected by the HSCIC may be aggregated where it is not available as patient level data or this is sufficient to meet the purpose, or as anonymised patient-level data, for example where the data is highly sensitive.

The HSCIC collect and process data for a wide range of purposes including secondary use purposes and also for direct patient care purposes e.g. diabetic retinopathy screening and Summary Care Record (SCR) and the Electronic Prescription Service (EPS). The HSCIC will only collect data where it has a legal basis to do so such as a Direction under s254 of the Health and Social Care Act 2012, S251 of the Health Act 2006 or with patient consent[5].

In many cases data from a variety of different sources is linked in order to gain a deeper understanding of the different elements of an individual's care, analysed and either published or disseminated to customers with the ultimate goal of improving health and social care. This is principally known as secondary uses of data.

[5] In some instances the HSCIC may act as a data processor under instruction from another organisation (as data controller).

Data released by the HSCIC is usually in the form of published aggregate statistics or de-identified data in line with the ICO Code of Practice on Anonymisation[6]. Identifiable data is only released where the Common Law Duty of Confidentiality can be set aside, i.e. it is for direct care purposes, where the patient has consented, with support under S251 of the NHS Act 2006 or where another legal basis exists.

Figure 2 below describes the controls which are in place within the HSCIC to ensure the confidentiality and security of patient's personal data.

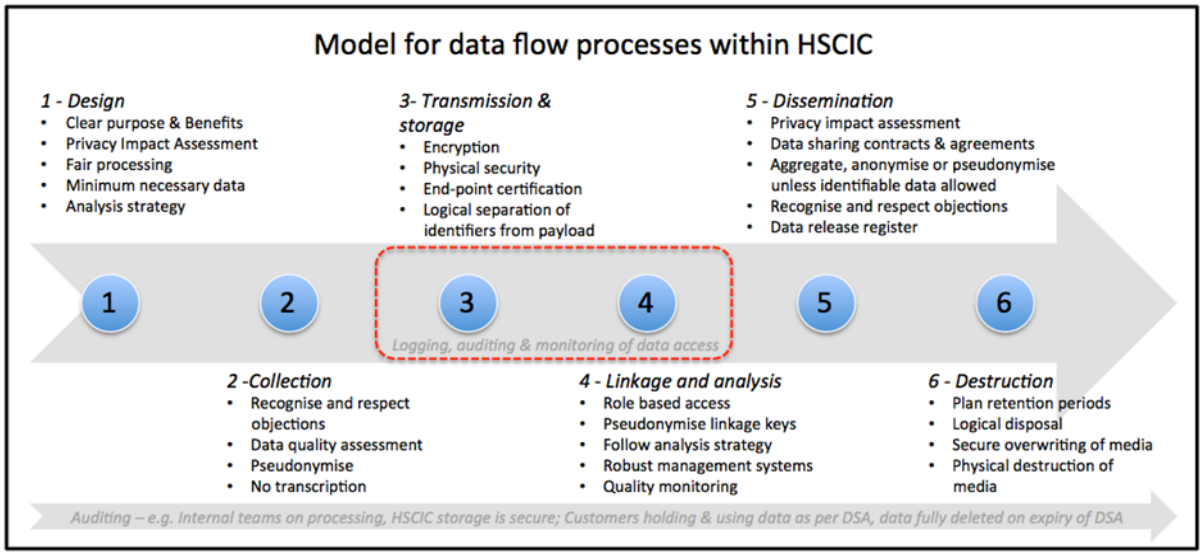


Figure 2. Model for data flow process within the HSCIC

The controls range from collecting only the minimum data necessary for the purpose; security controls such as encryption, role based access controls and auditing of access; treatment of data prior to dissemination to rendering it less- or non-identifying e.g. pseudonymisation, anonymisation or aggregation; to the use of additional safeguards such as Data Sharing Contracts and Agreements, and the eventual secure destruction of the data. The HSCIC also undertakes audits to ensure that these controls are adhered to.

[6] This includes pseudonymised data where additional safeguards are in place as set out in the ICO code of Practice on Anonymisation.

2 Pseudonymisation Review Approach

2.1 The Steering Group

The Interim Report, in proposing a further review into Pseudonymisation, also proposed that contributors should be drawn from those constituencies with an interest in the subject and who could provide the necessary rigour, intellectual analysis and independence to ensure that the Review's work reflects the wide a range of views that exist.

The members co-opted to the Steering group are listed in Appendix 2, Section 8.4.

The Steering Group's Terms of Reference, available at the Review's website [7]covered its role, responsibilities, scope and membership.

The Steering Group's role was as an advisory group to provide recommendations to the HSCIC on its pseudonymisation approach.

2.2 Overview of work and processes

The Steering Group identified three areas of work and sub-groups were set up accordingly to cover the subjects below -

- standards for pseudonymisation including glossary and terminology
- linkage and data quality
- considerations around pseudonymisation at source.

Each sub-group had a chair, which was responsible for recruiting its members, including from the Steering Group, to contribute to the work of the Review. The sub-groups developed a list of deliverables on relevant topics; these were duly produced, signed off by the relevant sub-group and considered and ratified by the Steering Group. This process was effective in ensuring that the coverage of the subject was appropriate, that the output was of a suitable standard to contribute to the work of the Review and enabled the Steering Group to be aware of all developments and findings during the Review.

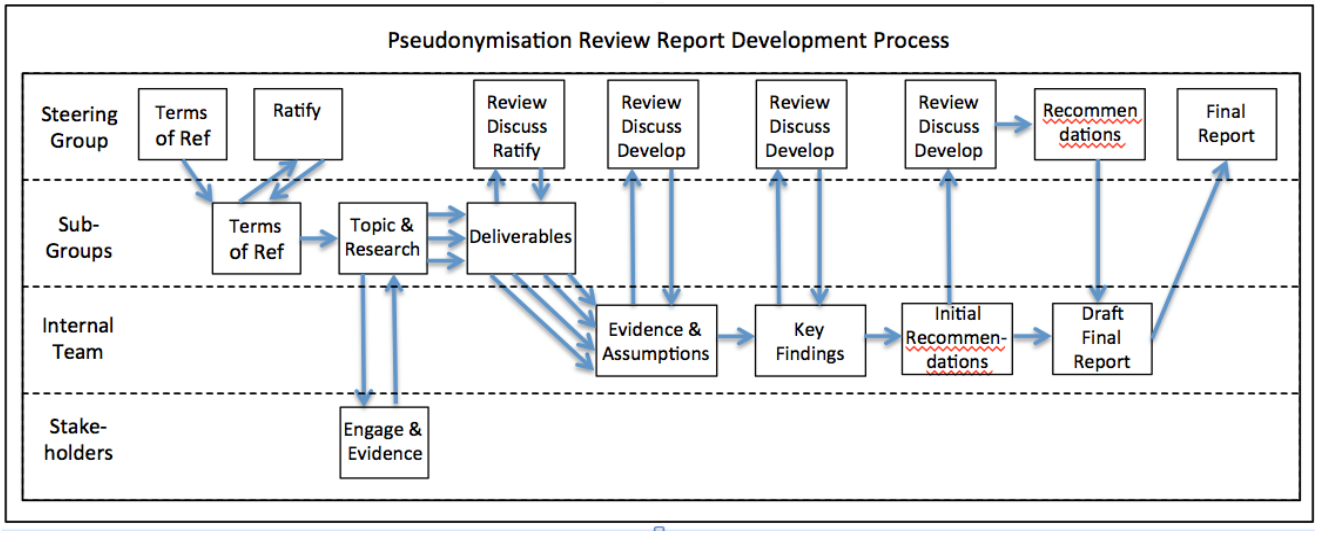
The reports (see Section 6.1 for more detail) from the Sub-groups enabled a set of Assumptions (see Sections 4.2 and 6.2 for more detail) to be developed setting out the factual and assumed basis about pseudonymisation and the HSCIC's purpose and role. This enabled a set of Key Findings (see Sections 4.3 and 6.3 for more detail) to be developed and agreed by the Steering Group.

Recommendations were derived from the Key Findings through meetings, workshops and review of electronic documents and signed off at a meeting of the Steering Group.

The overall process is summarised in Figure 3.

Figure 1 Review Report Development Process

[7] https://www.gov.uk/government/publications/data-pseudonymisation-review



2.3 Sub-groups

Three sub-groups were established with the remit to consider specific areas of the review as described below.

Sub-Group	Remit
Standards & Terminology	To set out the context in which pseudonymisation is useful and where de-identification is applicable, the legal context, the standards which apply to the use of pseudonymisation and to produce a meaningful list of the words and terms used to describe pseudonymisation and associated matters.
Data Linkage and Data Quality	To assess the impact that various forms of pseudonymisation have on the ability to link datasets, and the resulting utility of the linked data for secondary use purposes, such as research.
Pseudo at Source	To consider the three broad models, as detailed in the Interim Report, for pseudonymising the data collected by the HSCIC that should be considered further as part of next stages of the review.

The Terms of Reference for each Sub-Group, including membership of each group, are available on the HSCIC Data Pseudonymisation Review website:

https://www.gov.uk/government/publications/data-pseudonymisation-review

A list of the deliverables produced by each Sub-Group is available in **Appendix 1**.

2.4 Stakeholders

In addition to contributions by members of the HSCIC Data Pseudonymisation Review Steering Group and Sub-Groups, evidence has been submitted to the review from a wide range of stakeholder organisations including the following:

- Data Providers e.g. Secondary Care, Mental Health Trusts, GPs
- Health and Social Care Information Centre (HSCIC) Operational Teams (responsible for collection, processing, analysis and dissemination of information)
- Data Recipients e.g. research and commissioners
- Local authorities
- Arms-Length bodies e.g. Office for National Statistics (ONS) and Care Quality Commission (CQC)
- Suppliers of systems including pseudonymisation products, as represented by industry body TechUK
- Data subjects, e.g. patients

In addition to the above relevant research, was reviewed by the sub-groups during the evidence gathering phase of the Review. This included research papers from Gareth Hagger-Johnson, University of Central London (UCL) and Mark Elliot, University of Manchester and UK Anonymisation Network (UKAN) who presented on, and provided draft versions of, The Anonymisation Decision-Making Framework, to the Independent Steering Group.

3 Evidence Base & Key Findings

3.1 Evidence Base

The deliverables from each Sub-group formed an evidence base, together with material from other sources, (e.g. Anonymisation Decision Making Framework produced by the UK Anonymisation Network), plus the extensive knowledge and experience provided directly by members of the Steering Group and the Sub-groups. The deliverables are not being published by the Review however the assumptions and key findings, derived from the deliverables, are available in Appendix1 of this Report.

3.2 Assumptions

The statement of assumptions was developed to bring together material from the Pseudonymisation Reviews Sub-group deliverables into one place as a stepping-stone to facilitate the generation of Key Findings, which in turn were used to support the Steering Group in deriving recommendations from the Review.

The Assumptions paper also included some background information, not covered in Sub-group reports, to respond to issues raised during discussion in Steering Group meetings or workshops.

The Assumptions derived from the evidence captured during the review are set out in **Appendix 2** and are ordered by the following topics:

- Pseudonymisation and Managing Risk
- HSCIC role
- Central Pseudonymisation
- Pseudonymisation at source
- Hybrid model
- Other assumptions

Each assumption is categorised as one of the following:

- Context – essentially relevant background facts
- Assumption - where it is needed to make discussion and analysis possible; e.g. for Central Pseudonymisation *Assumption* - No data are pseudonymised prior to submission to the HSCIC (highly sensitive records are and will continue to be anonymised prior to submission to the HSCIC)
- Requirement – for items for which the HSCIC are required to meet goals or standards, NB a requirement may also be a supporting statement in a Context or Assumption
- Assertion - where detailed evidence is not available to back up the statement.

The following additional evidence has also been incorporated within Appendix 2:

- Relevant HSCIC Functions, Users and Data Usage, Data Flows
- Illustrations of definitions of Pseudonymisation and De-identification
- Single versus Multiple Keys for Pseudonymisation at Source (P@S)

A selection of key assumptions which emerged from the review are:

- Under a pseudonymisation at source model every data flow to the HSCIC will be pseudonymised prior to receipt by the HSCIC. The only exceptions to this would be where identifiable data is required and there is a legal basis for such a flow, such as information to support direct patient care or where the patient has consented to their data being used for that purpose
- The HSCIC will not be able to de-pseudonymised/re-identify individuals within any flow pseudonymised at source, although the submitter could allow the HSCIC to re-identify data in exceptional circumstances
- A single common pseudonymisation key (or a single key per purpose) would need to be used by all organisations submitting data to the HSCIC to ensure that data from disparate organisations can be successfully linked to provide a full picture of an individual's care
- Each direct patient identifier e.g. NHS Number, Postcode, Date of Birth would be pseudonymised separately (where it is needed e.g. for data linkage) – it will not be practical to generate a compound pseudonym for an individual e.g. based upon the combination of NHS Number, Postcode and Date of Birth, due to data quality issues if any of these differ between data sets, or split and pseudonymise components of an item separately such as day, month and year due to the risk of re-identification
- Pseudonymised data will often be deemed to be personal data under the Data Protection Act, however additional safeguards can be used to reduce the risk of re-identification

Full details of all assumptions captured during the Review are available in **Appendix 2**.

3.3 Key Findings

Key Findings in relation to Pseudonymisation and the HSCIC's role and functions, were derived from the Sub-group deliverables, the contents of the Assumptions document, discussions within the Steering Group and Sub-groups, reviews by Review members of draft Key Findings documents developed by the Review's internal team and from a workshop on 18th June 2015.

The Key Findings derived from the evidence captured during the review are set out in **Appendix 3** and are ordered by the following topics:

- General Findings
- Central Pseudonymisation
- Pseudonymisation at Source
- Hybrid model

Each key finding is categorised as follows:

- Context – essentially relevant facts
- Finding - where a conclusion has been drawn based on agreement and evidence based
- Implication – where the Finding implies possible consequential action
- Assertion - where detailed evidence is not available to back up the statement.

Some of the key findings highlighted within the evidence collected during the review are listed below:

- More needs to be done to address concerns of the public and healthcare professionals about what information is processed by the HSCIC, and how and in what form the information is released, to whom and what for

- A variety of different pseudonymisation products are available on the open market, in addition to bespoke solutions and functions in common tools
- Linkage of data using deterministic techniques should not be impacted by using pseudonymised data. However data linkage requiring the use of "fuzzy" or probabilistic techniques may not be successful using current pseudonymisation techniques depending upon the level of linkage required and resulting bias introduced. In some cases, it may not even be possible to probabilistically link such pseudonymised data.
- The cost of implementing a full pseudonymisation at source model for all flows into the HSCIC is likely to be significant, although it may be more cost effective to implement for certain areas
- There are not considered to be any barriers to interoperability associated with any of the models for pseudonymisation
- Pseudonymisation at source may impact the HSCIC's ability to accurately uphold patient objections which rely on NHS Number, or respond to Subject Access Requests (SARS) to identify the information held about an individual or S10 under the Data Protection Act (DPA) which require processing of an individual's data to cease as both will require records for the individual to be identified

A full list of the key findings can be found in **Appendix 3**.

4 Recommendations

The table below outlines the key recommendations from the Pseudonymisation Review Steering Group. The Steering Group, comprising both internal HSCIC and external members, is acting in an advisory capacity to the HSCIC EMT and as such offers the below set of final recommendations for consideration by EMT at its earliest convenience.

These recommendations have been derived from the evidence submitted during the review. This evidence has been incorporated into the sub-group deliverables listed in **Appendix 1**, which in turn have been distilled into the key assumptions and key findings outlined in **Appendix 2** and **Appendix 3** respectively.

No.	Topic	Recommendation
1	Public Confidence	*The HSCIC need to build public confidence by continuing to address public and professional concerns through a two-way dialogue. This will include being transparent about the data that it collects and processes, how it is kept securely, and whom the data is shared with, for what purpose and on what legal basis and how their confidential data is protected. It should also inform patients how they can express preferences to how their data is used.* *The HSCIC should list all the data sets it collects and processes, as well as all data it releases, on its website in an easily searchable form. This will help the public easily determine what data the HSCIC is likely to hold on them as an individual and who has access to that data in identifiable form.*
1a	Communicating Benefits of data sharing	*The HSCIC should identify and communicate the benefits to the patients and the wider health and care community associated with the collection, analysis, publication and other dissemination of health and care data, as well as the risk and means used to minimise it, including personal, sensitive and confidential data*
2	Ensuring use of pseudonymised data is appropriate	*Pseudonymisation on its own is often insufficient to protect the confidentiality of patient data. The HSCIC should provide training to HSCIC staff, the wider NHS and customers which covers the organisational, legal and technical implications of using pseudonymised, data, including the risks involved and legal penalties, prior to the sharing of data*

No.	Topic	Recommendation
3	Irreversible Pseudonymisation	*The HSCIC should apply pseudonymisation which is irreversible by the recipient unless there is a legitimate health related reason and appropriate organisational, technical and legal measures in place for the data to be re-identified. The HSCIC should by default own and control the pseudonymisation keys or lookup tables in cases where it disseminates pseudonymised data.* *The HSCIC should develop a policy around other types of key management requested in its dissemination of pseudonymised data and the circumstances under which it would consider such disseminations to be identifiable, in liaison with CAG.* *There should be transparency around the type of pseudonymisation applied, for example, in the data release register.* http://www.hscic.gov.uk/dataregister
4	Establish Centre of Expertise and Capabilities	*The HSCIC should develop an internal centre of expertise, which can provide best practice advice and guidance in relation to the **de-identification** of data, including pseudonymisation for itself and the wider NHS. This would include the development of relevant standards* *As a priority it should:* • *Develop specific criteria against which individual data collections by the HSCIC can be evaluated for the optimum usage of pseudonymisation in terms of the purpose of the data collection and respecting privacy.* • *Develop existing techniques for anonymisation to increase the utility of the data once its disseminated* • *Communicate to the public the results of this activity in understandable terms.*
5	Developments in privacy enhancing technique and technologies	*The HSCIC should consider how best to review and appraise developments in privacy enhancing and data security techniques and technologies on an ongoing basis to ensure that it adopts them at the earliest opportunity where appropriate. This includes technologies to reduce the flow of identifiable data to the minimum required for specific purposes, in line with requirements of the Data Protection Act.*

No.	Topic	Recommendation
6	Existing National data flows to HSCIC	*Existing National flows of identifiable data into the HSCIC should be subject to a rolling programme of regular review against specified criteria to ensure data flows in the least identifiable form necessary to meet the purpose.* *Each data flow should be reviewed in the light of legislative changes or significant technical developments, or if the requirements around individual flows change.* *It is accepted that there are specific purposes for which the HSCIC needs to collect and process identifiable data for example to perform probabilistic data linkage or when patients have consented to specific research e.g. BioBank.*
7	Segregation of patient identifiers from activity within HSCIC	*Where present on inbound data Patient identifiers should be segregated from remaining data upon landing within the HSCIC. Access to Patient Identifiers should be restricted to the minimum number of staff that absolutely requires access to these items for specific discrete purposes, with non-identifiable alternatives derived for analysis purposes e.g. Age rather than Date of Birth.* *Individuals should not routinely be able to access both patient identifiers and activity data. In the exceptional circumstances where access to both are required strict protocols must be adhered to including Senior Level approval* *Access to data should be fully controlled, audited and monitored on a continuous basis to assure compliance.*
8	New National data flows to HSCIC	*Any new national data flow should be subject to IG review, through a Privacy Impact Assessment, and would involve groups of the relevant data subjects and controllers where required. This should consider whether aggregate, fully anonymised or data pseudonymised at source or identifiable data could be used to meet the business objectives and realise the benefits to health and care, using data with the minimum risk of re-identification, to meet that purpose.*
9	Pseudonymisation at Source Proof of Concept	*At the point that a new national data flow into the HSCIC is identified where the benefits could be fully met under a pseudonymisation at source model a Proof of Concept should be initiated to prove the efficacy of this approach in relation to the HSCIC operating model.*

No.	Topic	Recommendation
10	Improving support to privacy of patient data	*The HSCIC should provide standards and tools to support the self-assessment and audit of the techniques to create and use pseudonymised and de-identified data across the health and social care system.* *The HSCIC should provide advice on local flows that do not currently involve the HSCIC when requested to do so.*

Note on Pseudonymisation at Source Costs

Whilst evidence presented during the review indicated that the costs of implementing a full pseudonymisation at source model (i.e. all data flows into the HSCIC) are likely to be considerable, the Steering Group are keen to emphasise that cost alone should not be seen as the basis for preventing the use of pseudonymisation at source for individual national data flows where there are no other barriers to using this method. The Steering Group are keen that is considered in the development of any criteria to assess the suitability of pseudonymisation at source for a particular national data flow.

Appendix 1 - Sub-Group Deliverables

The table below outlines the deliverables produced by each Sub-Group based upon the evidence gathered. These have been ratified by the Steering Group.

Data Linkage and Data Quality Sub-Group Deliverables		
Ref No.	**Title**	**Description**
DLDQ03	Data Quality	The Steering Group requested the Data Linkage & Data Quality sub-group to consider the prevalence and quality of a number of identifiers available in current datasets received by the HSCIC.
DLDQ04	Analysis on Impact of Pseudonymisation on Data Linkage	The aim of this paper is to set out the high level impact analysis of pseudonymisation , both at source and centrally, on linkages undertaken by the HSCIC
Pseudonymisation at Source Sub-Group Deliverables		
Ref No.	**Title**	**Description**
PS03	Report on open market of Pseudonymisation products	The Steering Group requested the sub-group to obtain a market level view of suppliers of pseudonymisation products that could be requested to provide the detailed specifications to meet a number of requirements and technical specifications.
PS04Q	Identify current capabilities of Pseudo @ Source products	The aim of this paper is to set the sub-group to obtain a market view of pseudonymisation products by using the identified suppliers from PS03. The report considered the capabilities of pseudonymisation products against a range of criteria covering technical, standards and implementation requirements. These were used to advise the Steering group as to the capability of such products to meet the stated criteria. The Report does not seek to identify individual supplier or product capabilities as being suitable for the HSCIC, but the capability of the market as a whole to meet requirements.
PS04A	Assess current capabilities of Pseudo @ Source market	The aim of this paper was to provide an assessment of the open market's level of capabilities in the pseudonymisation products identified in PS03. The assessment provided a whole of market view for the specific capabilities listed n PS04Q in order to advise the Steering group that the market penetration of those capabilities existed at a level that could potentially meet future HSCIC requirements for pseudonymisation.
PS05	Barriers of implementation of Pseudo @ Source products	The aim of this paper was to elicit the pros, cons and barriers for each of the three pseudonymisation models being considered by the Review. The responses from a range of stakeholders provide a range of perspectives, from different constituencies, as to the potential benefits and dis benefits for each of the models.

PS06	Impact of Pseudo products on Government Interoperability	This report aims to consider the impact of pseudonymisation at source on current government interoperability standards.
PS07	Report on implications of P@S on HSCIC DSA, Patient Consent and transparency requirements	This report considered the impact of 'pseudonymisation at source' on areas of HSCIC operations involving Information Governance, transparency requirements and it statutory obligations.
PS11	Relative security benefits and risks of different pseudo models	This report aims to consider the impact of pseudonymisation at source on current security standards and operations within HSCIC
Standards and Terminology Sub-group Deliverables		
Ref No.	**Title**	**Description**
ST01	Vocabulary & Glossary	The Steering Group requested S&T to provide a glossary of terms to support the review; a long list was reduced to a shorter list with elaboration on a set of key terms to ensure consistency and coherence; the key terms being de-identification, anonymisation and pseudonymisation.
ST02	Context of Pseudonymisation	The aim of this paper is to set the context for enabling both the use of the process of pseudonymisation and the resulting pseudonymised data on a sound legal basis by setting out the necessary associated technical, organisational and legal measures. This provides the overall context in which individual person level can be legitimately used.
ST03	Standards	This paper covers the standards applicable to pseudonymisation and the wider de-identification requirements and the need for good practice guidance within the NHS.
ST04	Legislation	The aim of this paper is to set out a list of the legislation that affects the use of pseudonymisation and de-identified data, the legal framework in which the HSCIC operates and to provide information on potential changes in legal matters relating to pseudonymisation and de-identified data

Appendix 2 – Review Assumptions

The Assumptions derived and developed during the Review are set out in Table 5. For background information on Assumptions, see Section 4.2.

Please note shading of rows in Table 5 is used to distinguish different groups of topic

Table 1 Pseudonymisation Review Assumptions

No	Topic	Assumption	Rationale
1	Pseudonymisation and managing risk	*Context* - The purpose of the Review, in general, is concerned with trying to balance the privacy and rights of patients with the needs of the public good when healthcare data are used. Specifically, for the HSCIC in collecting, processing and disseminating patient data, this means the Review should outline how Pseudonymisation, as an enabler, can help the NHS to operate securely, efficiently and effectively as well as providing other legitimate users of healthcare data, such as managers, funders and researchers, to gain useful information to benefit people in a safe and effective way, whilst maintaining the ability for people to object about the use of their data. *Requirement* - It is assumed that the public & professionals expect HSCIC to take all steps to maintain patient data confidentiality, in line with the legislation which established it. Therefore, there is a need to ensure that the HSCIC operates on a sound legal basis and is seen to do so in order to regain and maintain the public trust. There is inevitably an element of risk in the use of the personal data, such as hacking, inappropriate disclosure or inaccuracy of linkage, which will each require their own mitigation. Such risk cannot be totally removed, but it does need to be managed and suitably mitigated and minimised in order to achieve a mutually satisfactory outcome.	*Purpose of review and S&T Context paper*
2	Pseudonymisation and managing risk	*Context* - Pseudonymisation is one of many ways to reduce the risks associated with the inappropriate disclosure of personal health and social care data for secondary purposes. Other techniques include obfuscation, perturbation, data masking and anonymisation. Other controls include physical and organisational measures such as the use of data sharing contracts and agreements. *Requirement* - A requirement is the handling of patient consents and patient objections.	*S&T Context paper*
3	Pseudonymisation and managing risk	*Context* - Pseudonymisation on its own rarely produces data that would be regarded as de-identified, and therefore outside the remit of the Data Protection Act 1998. Such data must be in *de-identified form*, with or without pseudonyms, and can only be disclosed where the recipient organisation has undertaken relevant technical, organisational and legal measures that are agreed to by the data controller releasing the data.	*S&T Context paper*
4	Pseudonymisation and managing risk	*Context* - How Patient Objections will be implemented is not known yet (early July 2015). It is not known for instances whether Patient Objections apply to both identifiable and pseudonymised data. It is necessary to assume that there will be an impact arising from Patient Objections on data flows, data management and	*Awareness from internal discussions and public statements on Patient*

No	Topic	Assumption	Rationale
		potentially the use of pseudonymisation, but that that impact is unknown. (see Finding ***3A(4))***	*Objections*
5	Pseudonymisation and managing risk - Data Controller	*Context* - The accountability and liability for any security breach that may occur for the data that the organisation holds always lies with the Data Controller of the data, who needs to manage the associated risk.	*DPA*
6	HSCIC role – legislated role	*Context* - The Health and Social Care Act 2012 empowers the HSCIC to process identifiable data and has been confirmed as the Safe Haven for identifiable health and social care data for secondary use purposes. There are other bodies that hold personal data, such as ONS (for births and deaths) and PHE for infectious diseases etc. but only the HSCIC acts as a safe haven for processing and data linkage of healthcare episode records.	*S&T Legislation paper*
7	HSCIC role – identifiable data	*Context* - The Health and Social Care Act 2012 also allows relevant bodies to direct the HSCIC to collect and process identifiable data. Researchers with patient consent or S251 approval can request the HSCIC to collect and /or provided identifiable data. *Assumption* - In addition, the long-term model for data to support the commissioning process is based on the flow of identifiable data into the HSCIC in order to produce suitable pseudonymised data output. See Tables 2 and 3	*S&T Legislation paper*
8	HSCIC role – status of pseudonymised data flowing in	*Assumption* - Any pseudonymised data, (N.B. not de-identified data) flowing into the HSCIC is likely to be deemed personal data under the Data Protection Act (DPA) due to the richness of data required to meet the wide range of purposes that the data needs to support and the large volumes of different data sets processed. This leads to the potential that individuals could be identified by the combination of non-direct patient identifiers within a data set or across different data sets.	*PS07*
9	HSCIC role – Data Controller	*Context* - The HSCIC acts as Data Controller, either solely, joint or in common, for the majority of data sets that it processes and disseminates. As a Data Controller the HSCIC needs to comply with the DPA and related legislation.	*Statement of current practice*
10	HSCIC role – Data Processor	*Context* - The HSCIC acts as a Data Processor in some instances collecting, processing, linking and disseminating data under instruction from another organisation acting as Data Controller. The HSCIC would still need to comply with the Data Protection Act and related legislation and conditions within any approvals that it is operating under.	*Statement of current practice*
11	HSCIC role – data stored and risk	*Context* - The HSCIC processes large volumes of identifiable data and as a result there will always be the potential that a security breach could occur; the HSCIC owns this risk.	*Statement of current practice*
12	HSCIC role – data management to reduce risk	*Context* - The HSCIC splits identifiers from payload data and uses different physical locations for storing different information together with role based access controls for authorised users of data. This is in order to reduce risk of inappropriate access to data. It restricts and closely monitors those with access to both datasets. In effect (e.g. for SUS) the HSCIC operates 'pseudonymisation on landing' in using a 'root pseudonym' that	*Statement of current practice*

No	Topic	Assumption	Rationale
		acts as system identifier in place of the NHS Number. Any identifiable data that is output has to be created by joining relevant subsets of payload data with the same NHS Number via the root pseudonym. *Assumption* - It is a working assumption that the HSCIC will continue to split types of data and hold in different physical locations and closely restrict and monitor anyone with access to both identifiers and clinical data.	
13	HSCIC role – data quality	*Requirement* - The HSCIC has a statutory role concerning data quality. Personal confidential data are currently used to highlight records where data quality issues arise.	*Interim Report*
14	Central data flows involving HSCIC	*Requirement* - All existing national flows of data into the HSCIC and releases of data by the HSCIC meet minimum NHS standards for the transfer of data securely. Standards required are IG Toolkit Level 2, use of AES256, etc.	*Statement of current practice*
15	Central pseudonymisation (CP) - definition	**Definition for the Review** - the pseudonymisation of identifiable data after collection from its sources (e.g.by the HSCIC), perhaps after further processing, (e.g. linking), is known as Central Pseudonymisation. Its definition complements the definition of Pseudonymisation at Source, (see No 20 below)	*S&T Review Glossary*
16	CP – receipt of identifiable data	*Context* - No data are currently pseudonymised prior to submission to the HSCIC (apart from sensitive records which will continue to be anonymised as currently).	*Statement of current practice*
17	CP - pseudonymisation	*Context* - CP is already operational within the HSCIC with the risks understood and mitigation in place. There have been no security breaches or Serious Incidents Requiring Investigating (SIRI) relating to use of the central pseudonymisation model within the HSCIC and its data processing.	*Statement of current practice*
18	CP – data disclosures	*Context* - Separately, there have been errors in data disclosures, which have been reported to the ICO by the HSCIC and were subject of the Partridge Report. Revised procedures have been put into place to avoid similar problems in future	*Statement of current situation*
19	CP – reversible & irreversible pseudonymisation	*Context* - Pseudonymisation is used in irreversible and reversible ways; the latter means that records can be re-identified by pseudonymiser (or at their behest) where there is an overriding need to do so.	*Statement of current practice*
20	Pseudonymisation at Source (P@S) - definition	**Definition for the Review** - P@S has been defined in the Review Glossary, as 'The *pseudonymisation* of identifiable data by the data controller (i.e. organisation) that created the identifiable data, such as a patient's GP. Although it may be undertaken by a data processor on behalf of the original data controller, it must be done where only EU data protection legislation and practice apply (the latter refers to organisations meeting EU legislation requirements, but may be physically outside the EU). In effect, for this Review it means - the pseudonymisation of data prior to submission to the HSCIC or any other recipient of patient-level data from, for example, a practice.	*S&T Review Glossary*
21	P@S – not for direct care or explicit consent data or legal	*Assumption* – under a pseudonymisation @ source model the majority of national flows into the HSCIC would be pseudonymised prior to submission to the HSCIC. Some flows may continue to flow as identifiable data e.g. where they have the benefit of patient consent, where it is required to support direct patient care	*PS05/P@S Sub-Group*

No	Topic	Assumption	Rationale
	central functions	or for certain National Back Office functions such as the Patient Demographic Service where identifiable data are a prerequisite for the function (providing there is a legal basis for this to continue).	
22	P@S - HSCIC not able to de-pseudonymise	*Assumption* - The HSCIC will not be able to routinely de-pseudonymised/re-identify individuals within any flow P@S, unless there is an agreement specifying how HSCIC is to gain access to the pseudonymisation keys.	*Statements from P@S Sub-group*
23	P@S – HSCIC onward disclosures pseudonymised	*Assumption* - All data released by HSCIC to customers will be pseudonymised irrespective of the legal basis of the customer's requirements as the HSCIC will not hold any clear/identifiable data, except for situations outlined in Assumption 21.	*Statements from P@S Sub-group*
24	P@S – single pseudonymisation key	*Assumption* - A single pseudo key will be applied to ALL relevant national flows (i.e. excluding those in No 21). As a result a single pseudonymised data flow would replace the existing identifiable data flow - it will consequently not result in an increase in the number of data flows (i.e. no multiple separate data flows using different pseudonymisation keys for different purposes). (For rationale for this Assumption, see Part 4 of this paper)	*Statements from P@S Sub-group*
25	P@S – data items to be pseudonymised	*Requirement* - The data items to be pseudonymised for P@S are NHS Number, Date of Birth and Postcode. These will be pseudonymised separately as individual data items. Pseudonymising these 'direct identifiers' fits with the definition of Pseudonymisation and pseudonym in the Review Glossary (see also Table 5).	*Statements from P@S Sub-group and S&T Review Glossary*
26	P@S – pseudonym format	*Requirement* - Pseudonymised version of data items will be a different length and format to their non-pseudonymised source equivalents e.g. NHS Number (n10) compared with the Pseudonymised NHS Number (an40). Pseudonymised equivalents that are the same data type and format as their non-pseudonymised counterparts would not be sufficiently secure.	*Statements from P@S Sub-group*
27	P@S – standards are to be applied to P@S	*Assumption* - The use of a single universal tool for P@S would not be enforced. Providers would have an opportunity to procure the pseudonymisation tool that most meets their specific needs OR ties into their existing contracts/technology stack. These will all apply the same standards so will ensure interoperability in outputting the same pseudonymisation key for the same input item.	*Statements from P@S Sub-group*
28	P@S - providers data management	*Requirement* - To achieve a consistent pseudonymised identifier required to support data linkage, it is important that input data are of high quality using the appropriate standards and with relevant data cleansing routines applied. The HSCIC has powers to specify, design, validate, monitor and enforce such data standards to ensure that the processing of identifiers in source systems is undertaken in a consistent way. The HSCIC would also need to monitor data linkage rates to identify potential problems such as pseudonymisation keys not being correctly applied thus preventing linkage.	*Statements from P@S Sub-group*
29	P@S – linkage process	*Assumption* – under a pseudonymisation @ source model linkage of data would either need to be undertaken on identifiable data prior to pseudonymisation and submission to the HSCIC OR on pseudonymised data within the HSCIC.	*Statements from P@S Sub-group*

No	Topic	Assumption	Rationale
		To support the latter a single pseudonymisation key would need to be applied consistently by all organisations submitting the data to be linked. The proliferation of such a common pseudo key across the health and care system could increase the risk of individuals being re-identified. An alternative approach to pseudonymise for each purpose is considered to be challenging to implement and would result in additional burden to providers, an increased number of flows of data and complex key management arrangements.	
30	Hybrid Pseudonymisation - model	*Assumption* - For the purposes of the Review, the hybrid pseudonymisation model is considered as some flows being pseudonymised at source and others centrally.	*Statements from P@S Sub-group*
31	Hybrid Pseudonymisation - model	*Assumption* - For the purposes of making comparisons between models, hybrid pseudonymisation is considered as the mid-point of hybridisation with half of all flows into the HSCIC being pseudonymised at source, and the other half of all flows continuing to be in identifiable form. However, if either of the extreme models (CP or P@S) is not possible, a hybrid solution may be the best way forward.	*Statements from P@S Sub-group*
32	Hybrid Pseudonymisation – Key shared with HSCIC	*Assumption* -The HSCIC would require the use of the pseudonymisation keys applied to data pseudonymised at source to enable the HSCIC to centrally pseudonymise data to the same key for the same purpose.	*Statements from P@S Sub-group*
33	Need to future – proof person level data processing	*Assumption* - It is assumed that the use of patient level data for additional secondary purposes and in different contexts will continue to develop, (e.g. as health and social care processes are more closely linked, and different authorities exchange information to meet their statutory duties for safeguarding, mental health etc), together with public expectations. It is important that the direction of travel set out in the Pseudonymisation Review can legally support relevant processing of patient data for the foreseeable future with suitable techniques and tools.	*NHS England approach on secondary use of data for key health & social care purposes*
34	Limitation to the Review	*Scope* - Whilst the scope of the review focuses upon national flows into the HSCIC, pseudonymisation at source cannot be considered on this basis in isolation as there will be knock on impacts upon local flows through the health and social care system which have not been assessed as part of this review.	*From P@S report PS05*
35	Penalties for breaches or	*Context* - Staff employed by the NHS and its contractors have contracts that conform to the requirements of the HSCIC Code of Practice on Confidential Information[8]. Any organisation that meets the following criteria	*Review Workshop & HSCIC Code of Practice*

[8] *See http://systems.hscic.gov.uk/infogov/codes/cop/code.pdf*

No	Topic	Assumption	Rationale
	mishandling confidential information	**must** have regard to this code of practice • *health or social care bodies that collect, analyse, publish or otherwise disseminate confidential information concerning, or connected with, the provision of health services or of adult social care in England, and* • *persons other than public bodies who provide health services or adult social care in England pursuant to arrangements made with a public body exercising functions in connection with the provision of such services or care.* Section 23 of the Code requires that organisations • *Adopt formal contractual arrangements with all contractors and support organisations that include compliance with requirements for the handling of confidential information.* • *Adopt employment contracts with all staff handling confidential information on behalf of the organisation. These contracts should include compliance with requirements for handling confidential information.* Such contracts will include relevant sanctions in the case of breaches etc. Any additional action will depend on the event and severity of the transgression.	*on Confidential Information*
36	Costing Recommendations	*Assumption* – It is expected that any recommendations from the Review will need to be costed.	*From Steering Group Members*

Assumptions Part 2 - Relevant HSCIC Functions, Users and Data Usage, Data Flows

Table 2 List of HSCIC Specific Functions

A. Supporting Functions

Initials	Description	Purpose
DARS	Data Access Request Service	Manages and requests for access to HSCIC held data, principally HES
DAAG	Data Access Advisory Group	An independent group, hosted by the Health and Social Care Information Centre (HSCIC), which considers applications for sensitive data made to the HSCIC's Data Access Request Service
DSS	Data Steward Service	Maintains reference data (e.g. organisation codes, Post code address files (PAF)) and meta data in data libraries for HSCIC systems and operations; liaises with organisations which provide reference data; supports data quality by providing standards based reference data for comparative assessment of submitted data.
IGARD	Independent Group Advising on the Release of Data (IGARD)	An independent group with an independent chair and membership and an expanded remit to enable improvements in decision making in the respect of data releases. IGARD will succeed the Data Access Advisory Group (DAAG) during 2016.

B. Some patient data related Functions and data types

Initials	Description	Purpose	Data Types & access
NBO	National Back Office	Provision of a service for clinicians by identifying and linking each NHS patient in England, Wales and the Isle of Man to the care records uniquely associated with that person, and correcting confusions, duplications and inaccuracies. This is a demographics based service and does not include clinical data. (Source Nick Partridge Report)	Demographic data only
PDS	Patient Demographic Service	PDS is the national electronic database of NHS patient demographic details such as name, address, date of birth and NHS Number and forms the basis for the NBO service	Demographic data only
SUS	Secondary Use Service	SUS is the single, comprehensive repository for healthcare data in England which enables a range of reporting and analyses to support the NHS in the delivery of healthcare services, i.e. for 'secondary uses'; purposes other than primary clinical care. SUS acts as a data collection and collation point for all secondary care activity for use and analysis for multiple purposes. These include an activity Extract Mart service for commissioners and providers and a Payment by Results (PbR) Mart after case-mix and tariffs have been applied. See Table 4 for summary of uses and data type. SUS provides identifiable activity data to HES.	Identifiable data provided by secondary care providers Clinical data held against a SUS Root Pseudonym Analysis and processing undertaken internally with Root Pseudonym instead of NHS Number Output at patient level provided with NHS Number for authorised users and either user specific pseudonyms or in anonymised form.

Initials	Description	Purpose	Data Types & access
HES	Hospital Episode Statistics	Hospital Episodes Statistics (HES) is an analysis and publication service based on a data warehouse containing records of all patients admitted to NHS hospitals in England. It contains details of inpatient care, outpatient appointments and A&E attendance records fed from SUS. Data held includes • clinical information about diagnoses and operations, • information about the patient, such as age group, gender and ethnicity, • administrative information, such as time waited, and dates and methods of admission and discharge • geographical information such as where patients are treated and the area where they live.	Identifiable data feed from SUS Clinical data held against HESID (a pseudonym); Analysis undertaken internally with HESID instead of NHS Number; Output at patient level provided with user/purpose specific pseudonyms; NHS Number linked to HESID of external data set for linkage without staff involvement.
DLS	Data Linkage Service	Facilitates record linkage or combining and matching data sets at an individual record level in a secure environment, e.g. HES and external data sources, if approved by DARS & DAAG, mainly for researchers.	Demographic for identification of records; clinical data to be linked but not seen by service provider
MRIS	Medical Research Information Service	MRIS provides a service to researchers undertaking longitudinal studies. MRIS helps organisations, such as universities, to track cohorts of patients which typically range from around 1,000, with the biggest one being 1.3million. (Source Nick Partridge Report)	
PROMs	Patient Reported Outcome Measures	PROMs measures health gain in patients undergoing hip replacement, knee replacement, varicose vein and groin hernia surgery in England, based on responses to questionnaires before and after surgery. National figures are regularly published, together with analysis tools and reusable data packs – aggregated data only.	Identifiable data with consent
	Clinical Indicators Service	Generates and provides health indicators through an Indicator Portal as a health information resource. The indicators include CCG Outcomes, Population Health, Inequalities Indicators, GP Practices, Social Care, Quality Accounts, NHS Outcomes Framework, Summary Hospital level Mortality Indicator.	Indicators, e.g. percentages, ratios, graph plots
	Adoption Registration Service	The Adoption Registration Service can check for a record of a civil death registration in England, Wales and the Isle of Man on behalf of adoptees and birth relatives to establish if an adoptee or birth relative is recorded as deceased and can assist with forwarding information about hereditary medical conditions to an adoptee's or birth relative's GP.	Demographic only

Initials	Description	Purpose	Data Types & access
	1939 Registration Service	The 1939 Register Service answers requests for data held on the 1939 Register for England and Wales, as recorded on 29 September 1939 – demographic data only..	Demographic only
GPES	General Practice Extract Service	GPES collects information on behalf of specific and approved organisations that have Department of Health or NHS England sponsorship. Customers include NHS England, the Learning Disabilities Observatory and the Public Health England Diabetic Retinopathy programme known as GP2DRS. GPES extracts data to calculate individual practices' Quality and Outcomes Framework (QOF) achievement. QOF rewards practices for how well they care for patients rather than simply how many they treat, based on performance against indicators.	De-identified clinical data with HSCIC acting as an intermediary in the supply of data to relevant authorised bodies Extracted data deleted from HSCIC as soon as it has been passed on to the authorised customer

Assumptions Part 2 - Relevant HSCIC Functions, Users and Data Usage, Data Flows

Table 3 HSCIC Current Functions in relation to national multi-purpose data flows

Function heading	Function Detail (in relation to NHS data)	Users of Function
Safe Haven	To provide single safe haven facilities for collection and collation of identifiable and de-identified patient level activity data on behalf of a range of users of data	Providers, regulators, public health bodies, registries, commissioners, NHS England
Data Collection	Collect data sets from secondary and tertiary care. (NB data are also collected from primary care, mental health, community and social care)	Providers, regulators, public health bodies, registries, commissioners, NHS England
Patient Objections to use of their data	The HSCIC is committed to honouring patients wishes by upholding type 2 patient objections for data released outside of the HSCIC from April 2016 as Directed by the Secretary of State for health and the Department of Health. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/517522/type2objections.pdf	HSCIC on behalf of patients
Data quality and additional data derivations	Check basic data quality (e.g. fields present, suitable values); undertake national standard derivations to add fields; apply case-mix and national tariffs; meet statutory requirements	Providers, regulators, public health bodies, registries, commissioners, NHS England
Clearing House services of Record Collation and Dissemination	Collate and link patient level activity data from multiple sources for dissemination to authorised users and users who contract with the HSCIC for legitimate usage, effectively many-to-many relationships. Specifically, individual patient activity data in Commissioning Minimum Data Set (CMDS	Regulators, public health bodies, registries, commissioners, NHS England, information intermediaries. For commissioners, the data are used for multiple

Function heading	Function Detail (in relation to NHS data)	Users of Function
	or CDS) formats is sent from providers to the HSCIC to be collated into data sets for the relevant commissioner (specialist or CCG) and relevant geographic area (CCG) via CSUs. CMDS are sent on bulk and net change basis.	purposes, such as monitoring contracts and patient pathways, activity analysis and challenges, service planning and invoice validation and basis for payment.

An underlying requirement to meet the safe haven and clearing house functions in the simplest and most effective manner is that data landed in the HSCIC for the same person from different sources (i.e. different health care providers) must be linkable for HSCIC to carry out its statutory functions. This requires that data for the same person from all sources have common identifiers.

Assumptions Part 2 - Relevant HSCIC Functions, Users and Data Usage, Data Flows

Table 4 Users and stated needs of patient and clinical data provided by HSCIC[9] from national multi-purpose data flows

User (arising from Functions in Table 6)	Data type required (related to legal basis for use of data)	Purpose of use of data
HSCIC – data quality[10]	Identifiable	To support identification validation & data quality to avoid false matches in linkage and bias in data and as part of overall data quality standards
Researchers	Pseudonymised or identifiable legally enabled	To support clinical and non-clinical health and social care research
CCGs[11] & NHS England (NHSE)	Pseudonymised	Supporting NHS commissioning processes, e.g. contract management, service planning (e.g. location of services such as clinics and pharmacies)
CCGs & NHSE as commissioners	Identifiable data permitted by decisions under S251 in short term to pass to CSUs[12] Identifiable data to HSCIC in longer term with pseudonymised data for CSUs	To enable payment by commissioners for activity undertaken by providers of secondary care, particularly for non-contracted activity or out of area treatments.
HSCIC, NHSE & CCGs	Pseudonymised	For detailed in-depth analysis for enquiries at local and national levels (e.g. Mid Staffs), queries, (e.g. impact of junior doctor rotation), impact of policy changes (e.g. 7 day hospital working), etc.
CCGs	Pseudonymised	Assessing health risks at population level – risk stratification
Primary and community care	Pseudonymised & re-identifiable (where	Assessing health risks at individual person level – case finding (NB case finding

[9] *Based on Figure 1 in Context paper*

[10] *Requirement of Data Services for Commissioners (DSfC) programme arising from their Data Quality Impact on Linkage paper*

[11] *CCGs only receive patient level data relating to their registered and resident populations*

[12] *See http://www.england.nhs.uk/wp-content/uploads/2014/08/who-pays-advice-22-08-14*

User (arising from Functions in Table 6)	Data type required (related to legal basis for use of data)	Purpose of use of data
organisations	legally enabled)	involves authorised patient re-identification for clinicians with legitimate relationships)
Public Health England and in Local Authorities	Pseudonymised or identifiable where legally enabled	Public health surveillance
NHS Regulators	Pseudonymised or identifiable where legally enabled (eg CQC)	To support monitoring of NHS operations
Information intermediaries	De-identified with pseudonyms	To provide benchmarking products for NHS organisations

Assumptions Part 2 - Relevant HSCIC Functions, Users and Data Usage, Data Flows

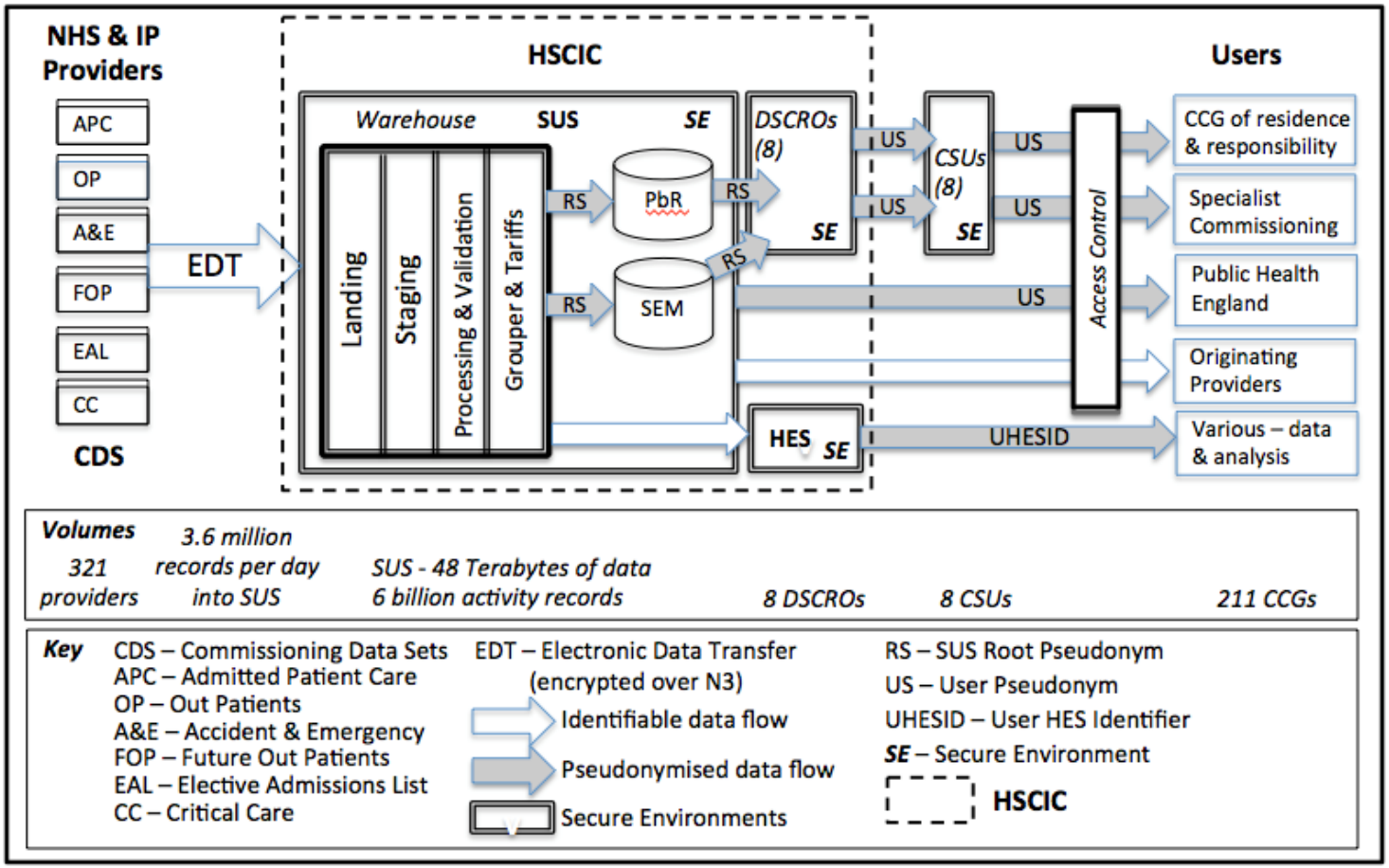
Table 5 Overview of existing patient and clinical data flows in the NHS in England relating to commissioning and performance management

Types	Description	Purpose
National	Flow of standardised datasets of patient activity for specified purposes from providers to HSCIC (e.g. commissioning minimum data sets (CMDS or CDS) for A&E, OP, Admitted Patient Care etc.) for SUS and HES.	To enable collation of data for multiple purposes as outlined in Table 3 by enabling the HSCIC to (i) act as a Clearing House for directing data from providers to local and national commissioners (see Table 2) (ii) act as national data repository for subsequent use on a national basis (iii) act as national data repository to support regulators and PHE operating on national, local and topic basis (iv) act as a national data repository to provide relevant subsets of data to authorised researchers (via HES) (v) act as service to provide de-identified data for Information Intermediaries (via HES or SUS)
Local	Flow from providers to local commissioners or others (e.g. researchers) of locally agreed data sets	To enable data to be provided on a point to point basis for use by organisations or researchers, etc. for specifically agreed local purposes (e.g. to supplement CDS to support local initiatives)

Assumptions Part 2 - Relevant HSCIC Functions, Users and Data Usage, Data Flows

Figure 4 illustrates a high-level view of CDS data flows from providers to SUS within the HSCIC and onward to commissioners and other users. Linkages of records between two data sources are ***not*** part of this diagram, but are mainly associated with HES. Within SUS there are internal 'joins' of data from activity-based tables (such as A&E and APC) for an individual patient's root pseudonym identifier or the use of Spell Identifiers and Pathway identifiers to join episodes of care together.

Figure 2 Illustrating data flows into and out of HSCIC



Assumptions Part 3

Single versus Multiple Keys for Pseudonymisation at Source (P@S)

The table sets out the issues for the use of single and multiple keys for P@S for the multi-purpose national flows illustrated in

Single and multiple keys

The main requirement is that data landed in the HSCIC for the same person from different sources (i.e. different health care providers) will have the same pseudonymised identifier to enable linkage; this may possibly be for a particular purpose or for multiple purposes. In effect the linkage requires a single key to enable different organisations, which have provided care to a particular person to assign the same pseudonym to that person. This will be particularly important where linkage of data collected over long periods time and across different providers is required.

Multiple keys then arise from using different single keys for different reasons, such as different purposes or different organisations, however linkage for the same person cannot be achieved across organisations for the latter approach.

Multiple keys will also arise if the HSCIC apply another level pseudonymisation to data for dissemination in order to offset the security risk arising from the P@S single key leading to users, e.g. commissioners, having the same keys as data providers, e.g. secondary or primary care providers. This is explored in Model 5 in Table 11.

Context of use

The consideration of the use of single key or multiple keys related to purpose has to be undertaken in the context of the data flows to the HSCIC, as opposed to local, point to point flows. The flows to the HSCIC are on a huge scale in terms of the number of sources, the range of data sets, the volumes of data due to the sizes of the CMDS and numbers of episodes of care, which are compounded by the need to manage net change updates on patients' stays in hospitals. The context is further complicated by the multiplicity of purposes for which the data are sent, as indicated in Tables 3 and 4 above.

Specific examples of use of data in commissioning where a consistent single pseudonymous identifier needs to be used across the system include

- to enable the commissioner to contact the relevant provider to discuss activity for a particular patient
- to identify duplicated activity i.e. provider 1 (main contract) and provider 2 (sub-contractor) both submitting and getting paid for activity.
- to be able to link activity across commissioning organisations e.g. to make sure patients don't fall between gaps.

Consideration of Single and Multiple Key Models

The consideration of different models of single and multiple keys is set out in Table 11 below in the context of use of multiple purposes for the flows of data set out above. The table indicates that there are strengths and weaknesses to each model.

The conclusion is that Models 3 and 4 with specific keys for each organisation for any purpose will not enable data for individual patients to be linked across organisations, a fundamental requirement. Whilst Model 2 based on a generic key being used across all organisations for a specific purpose is technically feasible, it is complex and difficult to manage where multiple purposes are involved. Model 5, where the HSCIC applies another layer of pseudonymisation to the submitted pseudonymised data for data dissemination means that it is not possible to track patients back from commissioners

to providers as providers and commissioners will have 2 different pseudonyms for an individual. This prevents communication to enable the example activities, listed in the previous section, from being undertaken, unless there is a system for the holders of the pseudonymisation keys to collaborate.

This leaves Model 1 of a single key across all organisations for all purposes as the only viable option for the large-scale multi purpose national flows. However, the benefit of replacing the NHS Number, in secure flows with another common pseudonymised identifier used across the NHS has been questioned. This is because NHS Number is deemed to be a direct patient identifier despite originally being intended as a pseudonym.

The table also implies that where there are multiple single purpose flows of data to the HSCIC, multiple keys, e.g. one per purpose as in Model 2 is applicable and could be considered for use

Table 6 Single and Multiple Key Models – pros & cons

No	Model	Description	Pros	Cons
1	Generic key across all Organisations for all purposes (single key for any purpose)	Each organisation uses the same key irrespective of the purpose	Data can be linked across organisations Simpler to implement	Security Risk: risk of key being exposed/compromised is increased as it is held in many organisations, particularly if used repeatedly or over a prolonged period – if cracked for one it is cracked for all. Data can be linked for purposes it is not meant to be used for unless secondary central pseudonymisation applied
2	Generic key across all Organisations for a specified purpose only (single key for specified purpose)	Each organisation uses the same but for a specific purpose only. Different keys are used for different purposes.	Data can be linked across organisations	More complex to implement and manage if many different purposes. Security Risk: risk of key being exposed/compromised is increased as it is held in many organisations. May result in the need for multiple flows of each data set e.g. risk stratification, analysis, commissioning, invoice validation etc. and additional costs processing these data Multiple keys imply more pseudonyms in use, so there will be a need for longer pseudonyms to be generated to avoid collisions. If same data are to be sent for multiple purposes, the sending organisation has to prepare and manage the data multiple times for each recipient/use rather than just the same set being sent to each
3	Specific key for each organisation for any purpose (multiple keys for any purpose)	Organisations use a pseudo key specific to them irrespective of the purpose	Reduces risk impact if pseudonymisation key is breached as restricted to a single organisation Relatively simple to implement	Impossible to link data received from different organisations for patient e.g. research, patient pathway/patient centric analysis, duplicate payments
4	Specific key for each	Organisations use a pseudo	Reduces risk impact if	Impossible to link data received from different

No	Model	Description	Pros	Cons
	organisation and purpose (multiple keys for specified purpose)	key specific to them and the purpose	pseudonymisation key is breached as restricted to a single organisation and purpose	organisations for patient e.g. research, patient pathway/patient centric analysis, duplicate payments May result in the need for multiple flows of each data set e.g. risk stratification, analysis, commissioning, invoice validation etc. and additional costs processing these data
5	Generic key across all Organisations for all purposes (single key for any purpose) for data submission; different key applied by HSCIC to submitted pseudonymised data for data processing and dissemination (single or multiple keys for dissemination)	Each organisation submitting data uses the same key irrespective of the purpose at source. HSCIC applies second pseudonymisation to the submitted pseudonymised data for data dissemination	Data can be linked across organisations Cracking input key does not impact on disseminated data	Security Risk: risk of key being exposed/compromised is increased as it is held in many organisations. Data subjects not identifiable (even if legal reason to do so) as applications of different input and output keys prevent this.

Appendix 3 - Key Findings

The Key Findings from the Review are set out in Table 12. For background information on Key Findings, see Section 4.3.

Please note shading of rows is used to distinguish different groups of topic

Table 7 Key Findings in relation Pseudonymisation and the HSCIC's role and functions

No	Topic	Finding	Evidence
1	General findings – working with patients	*Context* - The HSCIC does not routinely collect names and addresses when receiving data from secondary care organisations except for databases that support direct care. *Finding* - It is evident from the contributions of those involved in the Review, especially the lay representatives, that the importance, purposes and ranges of use of patient data are not widely understood (e.g. there appears to be a belief that names and addresses are generally stored and available). *Implication* - There is a need for greater simplicity and better communications in providing the public with the rationale for the collection and use(s) of the data, together with explaining the benefits and techniques for managing the risks. This need arises from enabling the common good to be achieved, but also from the need to maximise the amount of data that can continue to be available in order to minimise inadvertent bias caused by withholding of records. This suggests that greater and real transparency is required, preferably through the involvement of patients for them to understand the benefits.	*Derived from observations of non-NHS based members and some NHS based members during the Review*
2	General findings – HSCIC & public concerns	*Finding* - The HSCIC need to continue to address public and professional concerns through other means including being transparent about the data that it processes; how it is kept securely and whom the data are shared with and for what purpose; meeting Fair Processing requirements and data deletion requests	*Discussions at meetings of Steering Group and sub-groups*
3	General findings – HSCIC principles	*Finding* - HSCIC should collect the minimum data in the least identifiable form possible and apply the greatest level of de-identification to releases of data to meet the stated purpose of the customer and only ever with a legal basis to do so, supported by relevant data retention and destruction policies.	*Discussions at meetings of Steering Group and sub-groups*
4	General findings – Patient Objections	*Context* - As indicated in Assumption 4, the impact of the emerging policy about Patient Objections is unknown. It is clear however, that the flows of patient data into and out of the HSCIC will be affected. This means that due allowance for the impact of Patient Objections must be made in future	*Discussions at meetings of Steering Group and sub-groups*
5	General findings - pseudonymisation	*Finding* – Pseudonymisation, in its wider context, is a highly complex subject area with different understanding of key concepts, opinions of the best way to address the problems and highly polarised views amongst expert stakeholders.	*Discussions at meetings of Steering Group and sub-groups*
6	General findings – pseudonymisation	*Finding* - There are significant complexities and various permutations of how the various models could be implemented, each with different issues and costs associated, and as a result it is very difficult to assess	*Discussions at meetings of Steering Group and sub-*

No	Topic	Finding	Evidence
	models	the potential impact accurately	*groups*
7	Central pseudonymisation (CP) – Risk ownership	*Context* - As in Assumption 5, as a Data Controller, the HSCIC owns the risks for the identifiable data it receives and processes in its overall role and a provider of central pseudonymisation services.	*Statement of current situation*
8	CP - local flows	*Context* - The model of central pseudonymisation as operated currently by the HSCIC does not support local data flows, e.g between providers and commissioners.	*Statement of current practice*
9	CP – providers & commissioners preference	*Assertion* - The CP model is preferred by secondary care providers and commissioners, as this is most similar to the as-is situation and would not impose significant additional costs and burden upon them.	*PS05*
10	CP – primary care preference	*Assertion* - The CP model is not preferred for primary care data by practices on the grounds of data security and the breach in doctor patient confidentiality which would result and has not been the case before.	*PS05*
11	CP – Pseudonymisation Experience	*Context* - The HSCIC has experience of pseudonymisation, when there is a legal basis for releasing identifiable data, and routinely applies pseudonymisation centrally to data that it releases to customers. Pseudonymisation is applied on a per customer per purpose basis which guards against unauthorised linkage of data that is provided to different customers or to the same customer but for a different purpose.	*PS05*
12	CP – data management to reduce risk	*Implication* - As the HSCIC is able to segregate data into payload and internal system identifiers (see Assumption 12), the potential exists to avoid the use of S251 for some data flows in two or more ways. First, data from within the HSCIC systems could be linked for a particular purpose within HSCIC systems and then disclosed with pseudonyms without external users having access to the identifiable data; second through disclosing pseudonymised data with the same pseudonyms as data from another source, e.g. pseudonymised at source. See also Finding 49.	*Development of current practice*
13	CP – meeting HSCIC functional requirements	*Assertion* - The range of functions set out in Assumptions Table 3 Functions can be met, including the legal disclosure of identifiable data where appropriate, with implementation of Type Two objections yet to be completed	*Statement of current situation*
14	CP – meeting users' stated needs	*Assertion* - The range of needs set out in Assumptions Table 4 Users' Needs can be met.	*Statement of current situation*
15	CP - costs	*Assertion* - The costs of the central pseudonymisation model are met by providers sending securely encrypted identifiable data and by the HSCIC being responsible for the costs of pseudonymising data on landing prior to the data's subsequent processing and use.	*Statement of current practice*
16	Pseudonymisation	*Context* - As in Assumption 5, as a Data Controller, the organisation undertaking P@S (e.g. a practice)	*Statement of current*

No	Topic	Finding	Evidence
	at Source (P@S) – Risk ownership	owns the risks for the identifiable data it processes and pseudonymised data it produces.	*situation*
17	P@S – local flows	*Finding* - There is evidence that P@S has been successfully implemented to support some local flows of data between providers and commissioners as well for research & working with councils	*Evidence to P@S Sub-group*
18	P@S – primary care preference	*Assertion* - The P@S model is preferred for primary care data by practices on the grounds of data security and because it avoids breaching patient confidentiality.	*PS05*
19	P@S – providers & commissioners preference	*Assertion* - The P@S model is not preferred for secondary care data by providers as this would impose significant additional burden and costs upon them. The P@S model is also not preferred by commissioners because of the need for a single consistent pseudonymous identifier for patients across the system to enable them to fulfil their role.	*PS05*
20	P@S – pseudonymisation products for general use	*Finding* - A variety of different pseudonymisation products are available on the open market, including through Procurement Frameworks such as G-Cloud6. In addition bespoke solutions have been implemented within the NHS, and pseudonymisation functionality is also available in some GP and Patient Administration Systems (PAS). Furthermore bespoke tools can utilise functionality available within widely used database products such as Oracle, SQL Server and SAS.	*PS03*
21	P@S – pseudonymisation products	The capabilities, for pseudonymisation products, from market responses received would indicate the open market is sufficiently mature, robust and compliant, with several standards, to be considered for pseudonymisation, of patient data, in any of the three operating models under consideration. The Review should consider the market has the capabilities to support any of the models but should also consider risk (key management), deployment (distribution model) and costs when reporting to the Review's Steering group.	*PS04*
22	P@S – Interoperability	*Finding* - P@S will not conflict with Government Interoperability standards and methods. However changes to mandatory NHS Information Standards will be required; these are likely to have a long lead-time and will involve additional costs. However where pseudonymising at source is mandated then the HSCIC should consider provide direction on both the use of pseudonymisation at source, a new standard is currently being developed, and that an impact assessment on interoperability standards, is undertaken.	*PS06, PS05A*
23	P@S – Risk ownership	*Context* - The accountability and liability for any security breach that may occur always lies with the Data Controller of the data, who needs to manage the associated risk. Under the P@S model the HSCIC will receive only pseudonymised data, but may not be in a position to manage aspects of risk if the common key is compromised.	*Implications of implementing P@S*
24	P@S – Common key risk	*Finding* - There is a security risk associated with employing a common pseudonymisation key, required to allow linkage of data, across all health and social care settings including independent sector	*Implications of implementing P@S*

No	Topic	Finding	Evidence
		organisations. If this is compromised once it could potentially enable all data to be re-identified.	
25	P@S – Common key risk	*Implication* - Over time, the use of the same common key may lead to the pseudonymised NHS Number effectively replacing the NHS Number as a patient identifier in its own right.	*Implications of implementing P@S*
26	P@S – Common key risk	*Finding* - There is a risk that pseudonymisation may not be applied correctly at source by some organisations at some time thus rendering relevant data items not sufficiently obscured. However, as indicated in Assumption 28, the HSCIC has powers to specify, design, validate, monitor and enforce such data standards to ensure that the processing of identifiers in source systems is undertaken in a consistent way.	*Implications of implementing P@S*
27	P@S – Reputational risk	*Implication* - Concerns were raised by some stakeholders that roll out of a mandated P@S mechanism may result in unintended consequences such as the creation of a "black market" of unofficial local data flows. The reputational (and potentially worse) risk arises from the continued flows of identifiable data on an unofficial basis.	*PS05*
28	P@S – IG Impact	*Implication* - P@S may impact the ability of the HSCIC to adhere to elements of the Data Protection Act (DPA) and Information Governance Policy including Subject Access Requests (SAR), S10, Patient Objections and Preventing Use	*PS07*
29	P@S – Operational Experience	*Finding* - There is evidence that P@S has been successfully implemented within a few organisations, although such examples are based upon simple relationships between a data provider and a data recipient. There are also some examples where pseudonymised data are shared via an intermediary organisation such as QResearch, Clinical Practice Research Datalink (CPRD) and Secure Anonymised Information Linkage Databank (SAIL). The HSCIC has a more complex operating model acting as a Safe Haven to facilitate the sharing of data appropriately through collating multiple disparate data sets from many providers, linking this data and disseminating subsets to many customers. The HSCIC would need to explore such instances of P@S and intermediaries if the P@S model is adopted.	*PS05*
30	P@S – Operational Experience	*Finding* - There is evidence that suitable pseudonymisation capabilities do exist within systems provided by most suppliers to GP Practices and within systems used by some Secondary Care Providers.	*PS05*
31	P@S – Operational Capability	*Finding* - Within many hospital trusts, there may not be a sufficient level of technical expertise to ensure that P@S is carried out effectively, consistently and to the appropriate standards. P@S is also seen to be challenging to implement by some trusts. There would be risks that pseudonymisation is not applied properly resulting in a risk of re-identification or the wrong pseudonymisation key applied preventing the linkage of data to meet the business need. The risks could be mitigated by use of standards and supported by contracts, as outlined in Assumption 28. The technical aspects could be automated.	*PS05*
33	P@S – meeting HSCIC functional	*Finding* - The range of functions set out in Assumptions Table 2 cannot all be met. This is because identifiable data are required, such as for the collection, collation and dissemination of identifiable data.	*Consequent on the definition of P@S and ST04*

No	Topic	Finding	Evidence
	requirements	Meeting this requirement would mean the pseudonym would need to be reversible. This also appears to prevent meeting the need to comply with Type Two Objections. However, as set out in Assumption 4, the requirements for patient objections are not yet known.	*Legislation paper section 3.4*
36	P@S – Data quality – HSCIC role	*Finding* - The HSCIC has a statutory duty to provide independent data quality reports, which are mostly based upon patient identifiers e.g. NHS Number, Date of Birth, postcode etc. The capability to provide such reports with P@S data, especially on NHS Number, will be removed. However P@S data can provide a mechanism to validate NHS number at source then this should be explored should P@S be considered as a future operating model.	*Consequent on the definition of P@S*
37	P@S – Data quality by providers	*Finding* - The coverage and validity of NHS number is in the high for large scale national datasets, as set out in DLDQ03 (ie high 90%s). It is more challenging to understand the accuracy and where data are inaccurate whether having the data in identifiable form would enable any improvement to linkage and further work is needed in this area, as also set out in the NHS England paper on Data Quality. The Review's DLDQ04 Data Linkage can be used to infer some of the inconsistencies in identifiers that may occur under a pseudonymisation at source or central pseudonymisation model. Under such models patient identifiers should be validated against the patient demographic service to ensure maximum accuracy, in the former case by data suppliers and in the latter case by the HSCIC.	*DLDQ03; NHSE DQ Report*
38	P@S - Deterministic Linkage	*Finding* - Providing pseudonymisation keys are managed and applied correctly no impact on the ability to link pseudonymised data using deterministic linkage methods is expected, though there would be an increased overhead in processing and complexity of the linkage process. Also partial deterministic matches (e.g. part DOB) would be impacted, depending how these data were pseudonymised.	*DLDQ04*
39	P@S - Probabilistic Linkage	*Finding* - The effectiveness of partial, fuzzy and probabilistic matching techniques is expected to be reduced if these are needed by the HSCIC in the future to maximise linkage of data for an increasingly diverse range of data sets with differing characteristics in terms of patient identifiers and with different levels of data quality.	*DLDQ04*
40	P@S – meeting users' stated needs	*Finding* - A need listed in Assumptions Table 4 cannot be met by P@S. This is the need by researchers for HSCIC to provide secondary care identifiable data subject to legal basis (e.g. S251). Significant concerns have been raised by research organisations about their ability to continue receiving the identifiable data that they require to undertake vital research and have a legal basis to receive under a P@S model, or the additional cost and effort of re-identifying data where this is a feasible alternative. Alternative ways would need to be found by customers to meet their needs, which do not appear feasible in the short-term. Any solution is expected to lead to duplication of flows, in both identifiable and pseudonymised forms and re-processing of pseudonymised data into identifiable data for restricted use.	*PS05*

No	Topic	Finding	Evidence
41	P@S – meeting users' stated needs	*Finding* - Some of the needs listed in Assumptions Table 4 cannot be met by P@S. These include the capability to undertake identity validation and assess bias arising in secondary care data. This does appear to rely on the availability of identifiable data to enable this function to be carried out.	*NHSE DQ Impact on Linkage paper*
42	P@S – meeting users' stated needs	*Finding* - Authorised re-identification, (e.g. for risk stratification) can be supported through P@S for originating data providers (i.e. at source) if single key used at source and same key used for dissemination.	*Consequent on the definition of P@S*
43	P@S – impact on existing stored data at HSCIC	*Implication* - The HSCIC holds large amounts of data covering many years of secondary care activity. Much of the data are in identifiable form (indirectly identifiable in SUS as activity data stored against root pseudonyms and not NHS Numbers) or in a pseudonymised form using the HESID in HES. These data are used for providing comparative statistical data over time, for linkage and pathway data over time. Changing to a P@S model raises issues about what happens to the existing data at the point of change in order that there can be continuity of service to end users. Solutions may be technically feasible, but will require design and implementation, taking time and adding costs.	*Implications of implementing P@S*
44	P@S – impact on existing stored data at CSUs	*Implication* - HSCIC supplies pseudonymised data to CSUs (CCG Support Units) via DSCROs. The same issue, the impact of change of the type of data being provided arises, but on a local basis and in a different form as the HSCIC will not be able to apply the same pseudonyms to existing patients to enable pathways to be continued etc. Again, solutions may be technically feasible, but will require design and implementation, taking time and adding costs.	*Implications of implementing P@S*
45	P@S – operational management	*Implication* - The operation of P@S across 315 Trusts and 8,000 practices (albeit through 4 main system suppliers and large scale data centre operations) will require administration, coordination and management to ensure synchronised delivery of data to ensure the timely processing of data by the HSCIC. The types of processes involved will include conformance to standards, timetabling of data flows, dealing with basic data quality issues (e.g. incorrect coding leading to rejection of interchange of data). *Assertion* - This operational management can be developed from the existing arrangements for the supply of data to HSCIC.	*Implications of implementing P@S*
46	P@S – new flows	When new data flows are envisaged, P@S should be considered as part of the Privacy Impact Assessment	*Discussions at meetings of Steering Group and sub-groups*
47	P@S - Costs	The implication of costs to implement a Pseudonymisation at Source is likely to be considerable when comparing to a Central pseudonymisation solution. The main costs are changes to National systems, operated by the HSCIC, and for Secondary Care Providers where there is a wide range of variance of implementation models possible. For GP / Primary Care costs for pseudonymisation are not considered to be significant, although costs for changes to GP systems have not been collated, however for data recipients, being largely under a central	*PS05A*

No	Topic	Finding	Evidence
		pseudonymisation model, then costs for GP / Primary Care would not present a barrier.	
48	P@S – application and usage	*Implication* - The implications of the consideration of single and multiple keys in the Assumptions Part 4 are that (i) use of P@S is best suited to single purpose record linkage using shared pseudonymisation keys, as exemplified in Findings for local flows, specific research projects, i.e. subsets of the overall population (ii) P@S is not well suited to support large-scale multi-purpose national flows of patient records into the HSCIC and the subsequent many-to-many linkages. Whilst use of P@S may be potentially technically possible with single or multiple keys, there would be layers of complexity of operation and management with additional costs. The single key approach forms the least complex and least costly, but the benefits of replacing the NHS Number, *deemed to be a direct patient identifier despite originally being intended as a pseudonym*, in secure flows with another common pseudonymised identifier used across the NHS have been questioned (iii) The use of a second key on data submitted via P@S operated with a single key can offset the security risk of a common pseudonym at both ends of data flows. However, there will then be basic requirements that will not be met, such as re-identification by data providers with risk stratification or the provision of identifiable data where there is a legal basis to do so. (iv) by implication of (i) above, where data are extracted from SUS or HES for specific purposes involving linkage with other data sources, the P@S technique is suitable to be used to via a relevant pseudonymisation tool for data from SUS or HES and other related source(s) to facilitate such linkage.	*Assumptions analysis and evidence of use of P@S*
49	Hybrid model – pseudonymised output to reduce risk through providing a pseudonymisation service	*Implication* - An implication of Finding 12 is that the potential exists to avoid the use of S251 for some data flows if the HSCIC is able to provide a facility or service to irreversibly pseudonymise output for relevant files of data on the same basis that the external user irreversibly pseudonymises their data. In effect this would be a P@S type operation working on data from HSCIC and from end users giving the same irreversible pseudonyms to enable linkage. In principle, this could be achieved by the HSCIC either (i) enabling any service vendors to provide their certified pseudonymisation tool for the HSCIC to use on specific data sets or (ii) the HSCIC and service vendors develop a tool, possibly open-source, to operate as a black box with key management on HSCIC sourced data and the service vendor's service users' data. Organisations covered by previous permissions under S251 would then have the opportunity to consider at annual review whether such a service could remove the need for them to continue using identifiable information.	*Implication from Finding 12*

Appendix 4 - Steering Group Terms of Reference & Membership

I. The Role of the Steering Group

The steering group is an advisory group that will provide recommendations to the HSCIC on its pseudonymisation approach. The group will agree a set of options around pseudonymisation and a set of criteria for evaluating them against. Where agreement cannot be reached then the divergent views will be noted, together with approximate numbers holding these views.

II. The Responsibilities of the Steering Group

Once the draft report has been completed, the group will be responsible for evaluating the agreed set of options to quantify the advantages and disadvantages of each one and produce recommendations. Where recommendations do not achieve consensus amongst the steering group membership then the divergent views will be noted against particular recommendations, together with approximate numbers holding these views.

In providing these recommendations, the group is expected to take a number of factors into consideration in its evaluation, including but not limited to: technical feasibility, impact on data security, timelines and cost and impact on benefits.
It is envisaged that the steering group will provide recommendations on a specific aspect of pseudonymisation once it has been considered, rather than produce all recommendations together at the end of a process.

The group will be able to request the HSCIC to perform background work to aid it in its evaluation of the different options and assist in the prioritisation of this work. The HSCIC will assess the resource requirement to deliver this background work and decide whether it can deliver it. The Steering Group can invite external experts to its meetings where their skills are pertinent to the particular subject matter being discussed.

The Steering Group can convene a number of ‘task and finish’ subgroups to look in more detail at specific aspects of pseudonymisation, drawing on a subset of steering group members and outside experts where appropriate.

The Steering Group will provide formal recommendations on the following areas:

- The ways in which pseudonymisation could or should feature in relation to current and planned data flows into and out of the HSCIC
- The risks, issues, opportunities and constraints pertaining to pseudonymisation.

The Steering Group will not provide formal recommendations on the following areas, but some of them will be of interest to the group and the group should be mindful of any implications on these areas in making its recommendations:

- The use of pseudonymisation in point-to-point contexts independent of the HSCIC.
- Assessment of the merits of central data warehouses or models for customers accessing HSCIC data, for example on-site access or delivery of extracts;
- Assessment of consent models, e.g. 'opt in' versus 'opt out';

Assessment of any Information Governance recommendations that may emerge from the IIGOP

- Any general ethical aspects of using identifiable or de-identified data.

III. The Scope of the Steering Group

The Steering Group will provide formal recommendations to the HSCIC Executive Management Team (EMT). The HSCIC EMT will respond to such recommendations.

The Steering Group will have some common membership with the Independent Information Governance Oversight Panel to enable appropriate links to be made. The group has an overarching role looking at all HSCIC current or future datasets so has no specific links to forums that consider individual programmes of data expansion, such as the Care.Data Independent Advisory Group.

The Steering Group can make recommendations on pseudonymisation of data currently or planned to be received, processed and disseminated by the HSCIC, its data processors or its data controllers in common.

IV. Membership

As indicated in Section 2.4, representatives of relevant organisations and experts in specific fields were invited to become members of the Pseudonymisation Review Steering Group with observers representing other interested parties or areas of expertise. Members had two different means of participation 'active members' and 'corresponding members'. Core Members attended the majority of Steering Group meetings on a face to face basis, dialled in or were represented by colleagues, whilst Corresponding Members' participated by providing comments on documents or through email dialogue.

Table 8 Pseudonymisation Review Steering Group Members

Name	Organisation	Specialism	Role
Max Jones	HSCIC - Director of Data and Information Services	Chair providing input and steering group direction	Chair (July to November 2014)
Chris Roebuck	HSCIC	Review Co-ordinator	Chair (from December 2014)
Chris Carrigan	Public Health England	User of HSCIC Data	Active Member (deputy - Sean McPhail)

Name	Organisation	Specialism	Role
Antony Chuter	Member of the public	Patient Representative	Active Member
Professor Harvey Goldstein	University College London & University of Bristol	Academic Expert (Data Linkage)	Active Member
Ian Herbert	BCS Primary Health Care IT Specialist Group	GPES Independent Advisory Group member	Active Member
Dr Julia Hippisley-Cox	GP and Nottingham University	Academic Expert on Data Linkage and EMIS National User Group	Active Member
Geraint Lewis	NHS England	Chief Data Officer	Active Member (deputy – Xanthe Hannah)
Tim Williams	MHRA, Director of CPRD	User of HSCIC Data	Active Member (replaced John Parkinson January 2015)
James Wood	HSCIC	Infrastructure Security Manager	Active Member
Dr Paul Cundy	General Practitioners Committee and BMA	Representing BMA in its entirety; GPC (a Sub Committee of BMA) and ex Chair of the RCGP	Corresponding Member
Alan Hassey	HSCIC	HSCIC IG lead and member of Dame Fiona Caldicott's IG panel	Corresponding Member (due clash of meetings with Chairing DAAG)
Dr Phil Koczan		Representative of the RCGP and member of the Health Informatics Group	Corresponding Member
Daniel Ray	University Hospital Birmingham	Head of NHS CIO Network	Corresponding Member
Dr Hashim Reza	Oxleas NHS Foundation Trust	Consultant Psychiatrist and Mental Health Information expert	Corresponding Member
Eve Roodhouse	HSCIC	Care.data Programme Director	Corresponding Member (supported by David Ibbotson)

Richard Pantlin of Oxfordshire Social Services attended 3 meetings, but declined to attend further meetings due to the lack of coverage of the use of Social Services' data.

Table 9 Pseudonymisation Review Steering Group Observers

Name	Organisation	Specialism	Role
Natasha Dunkley Kambiz Boomla	Confidentiality Advice Group	Provide input on patient confidentiality. N.B. One CAG member attended each Steering Group meeting	Active and corresponding members

Name	Organisation	Specialism	Role
C Marc Taylor			
Wally Gowing	HSCIC	Pseudonymisation Adviser	Active
Nicholas Oughtibridge	HSCIC	Leading on Code of Practice for Confidentiality	Active
Dawn Monaghan	Information Commissioners Office	Data Protection Act - Supported by Stacey Egerton	Corresponding

V. Steering Group Members' Interests

The stated interests of Steering Group meetings are recorded in Table 15 (interests are in addition to stated roles in Table 14)

Table 10 Steering Group Members' Interests

Member	Stated Interest
Harvey Goldstein	Working on record linkage project at UCL that has access to HSCIC data
Dr Phil Koczan	GP, Chief Clinical Information Officer UCLP Digital Clinical Champion London (NHS England)
Martin Staples	NHS England – Data Sharing & Privacy Advisor
Julia Hippisley-Cox	Medical Director ClinRisk Ltd Trustee EMIS National User Group (charity), Director QResearch (not for profit venture between University of Nottingham & EMIS) Member of Confidentiality Advisory Group Health Research Authority Expert witness to Health Select Committee
Kambiz Boomla	GP principal & partner – Chrisp Street Health Centre, London Senior Clinical Lectoruer at Clinical effectiveness Group, Queen Mary University, London IT clinical lead – Tower Hamlets CCG
Ian Herbert	Director of S I Herbert & Associates Ltd , providing informatics services mainly in the field of healthcare Technical appraiser of standards, NHS Information Standards Board Committee member of British Computer Society – Primary Healthcare Specialist Group Vice Chair (Partnerships) BCS Health , Board Member, UK Faculty of Health Informatics