Document filename: HSCIC Data Pseudonymisation Review – Interim Report			
Organisation	HSCIC	**Project**	Pseudonymisation Review
		Status	Interim Report
Owner	Chris Roebuck	**Version**	1.0
Author	Chris Roebuck	**Version issue date**	31 July 2014

HSCIC Data Pseudonymisation Review – Interim Report

Document Management

Revision History

Version	Date	Summary of Changes
0.1	10-03-2014	First draft
0.2	25-04-2014	Updated following comments from within HSCIC
0.3	16-06-2014	Updated following comment from meeting of steering group
0.4	03-07-2014	Updated following further comments from steering group
0.5	14-07-2014	Updated following further comments from steering group & HSCIC communications team
0.6	29-07-2014	Further amendments following steering group final review and final proof read
1.0	31-07-2014	Document published

Reviewers

This document must be reviewed by the following people:

Reviewer name	Date	Version
HSCIC data pseudonymisation review steering group	25-07-14	0.5

Approved by

This document must be approved by:

Name	Title	Date	Version
Chris Roebuck	Director of Benefits and Utilisation	31-07-14	1.0

Contents

1. Introduction

1.1. In November 2013 a review into the HSCIC's use of Data Pseudonymisation was commissioned by the HSCIC's Director of Data and Information Services.

1.2. The scope of the review was the use of pseudonymisation in respect of data in transmission to, received, held and disseminated by the HSCIC.

1.3. This interim report sets out recommendations and options from the review to date, which will form the basis of the work of the pseudonymisation review steering group.

1.4. The aim of applying pseudonymisation to the HSCIC's datasets is to help protect patient confidentiality (by reducing the potential identifiability of the data collected, held and disseminated by the HSCIC) whilst still enabling the datasets to be used for public benefits such as research and improvement of the health service. This review incorporated interviews, workshop and a webinar with subject matter experts, which have been summarised in section 6 of this report. It also included an examination of the HSCIC's business functions and discussions with leads for some of the areas. This document presents recommendations and options around pseudonymisation from the review for further consideration.

2. Summary of recommendations

i. The HSCIC should use pseudonymisation as an important safeguard to help protect patient confidentiality whilst still enabling the use of the data for health related public benefits. Identifiability of patient level data is a function of both data content and its context so pseudonymisation must be used in conjunction with other controls including policy, security, governance and transparency[1].

ii. Further work to review the use of pseudonymisation should be undertaken in coordination with HSCIC work streams covering other controls[2] and there should be wide public engagement undertaken around these controls. The use of pseudonymisation should be externally reviewed at regular intervals to ensure that its use is effective and appropriate and that best use is being made of technological advances.

iii. Where pseudonymisation is applied to data disseminated by the HSCIC, it should be applied each time data are disclosed across an organisational boundary using per customer, per purpose specific pseudonyms. This will help mitigate unauthorised data linkage by third parties

iv. Pseudonymisation is not a "one size fits all" solution. There is a residual risk of jigsaw re-identification even when data items regarded as person identifiable have been replaced. Providing samples rather than whole dataset outputs and removing or obfuscating more information from the data lowers this risk, but potentially decreases the utility of the data for some purposes. Therefore, the extent to which different components of a dataset need to be obfuscated should depend on the risk of re-

[1] As per *Anonymisation Code of Practice*, Information Commissioner's Office (ICO), 2012 (http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation)

[2] Controls covering inbound, processing and outbound flows of HSCIC data, including: Identity Data Standard, The HSCIC Index and Enterprise De-identification Solution, and continuous review of all of the HSCIC's security controls to ensure maximum robustness. System wide activities, such as the consultation on Accredited Safe Havens (ASHs) are also relevant and appropriate links will be made.

identification. This in turn should depend on the context in which it is used, other information to which the data user has access or may have access in the future[3] and other controls in place.

v. Where others access data held by the HSCIC, pseudonymisation should be applied in all appropriate[4] circumstances, alongside other techniques to minimise risk of re-identification. Additional external assurance should be applied to the pseudonymisation techniques the HSCIC currently employs.

vi. A further stage of the review should be undertaken to consider the pseudonymisation required in different situations and evaluate in detail the following three broad options for pseudonymisation of data collected by the HSCIC:

- pseudonymisation of data centrally (after receipt by the HSCIC)[5];
- pseudonymisation of data at source (before disclosure to the HSCIC);
- a mixture of pseudonymisation at source and pseudonymisation centrally.

vii. A project and steering group with broad representation from external subject matter experts should be set up to evaluate the three broad options further, in particular as to:

- other possible options;
- security, controls, governance and transparency;
- implementation (benefits, cost and time);
- whether there exist any functions which any of the options cannot deliver and to determine the importance of any such functions (to include the quality of linkage for pseudonymised data); and
- an understanding of how patient objection information could be managed under each of the options.

viii. There is a range of purposes covering non-direct care for which health data that flows into the HSCIC are used. This includes local commissioning, commissioning at scale and research. The data may flow into the HSCIC to support these processes, or as in the case of the Commissioning Dataset (CDS) the data that the HSCIC captures are a by-product of the commissioning process. The steering group should consider this range of different purposes to ensure that the solution is, or solutions are, appropriate for each purpose.

ix. Examples of implementation of different pseudonymisation models should be considered extensively as part of the next stage of the review. It is recommended that links are made to the project to revise the "Health informatics -- Pseudonymization" ISO standard. Consideration should also be given to work by the Clinical Practice Research Datalink (CPRD), QResearch, ResearchOne, the Welsh Primary Care Data Extract service and the Scottish Primary Care Information Resource (SPIRE)

[3] As per ICO supplementary guidance.
[4] The first preference should be to provide fully anonymised data in line with the NHS anonymisation standard: http://www.isb.nhs.uk/documents/isb-1523/amd-20-2010/1523202010spec.pdf If there is a clear need for pseudonymised data to benefit the Health and Social Care system and appropriate security controls and assurances are in place, as well as appropriate approvals given, then access to data could be given pseudonymised per customer per study. The security control requirements are being considered in more detail outside this review.
[5] Where the HSCIC currently employs pseudonymisation, it is performed centrally, typically after data quality and data linkage work.

programme. Consideration should be given to whether a pseudonymisation standard is required.

3. Context

3.1. The Health and Social Care Information Centre (HSCIC) is responsible for collecting, transporting, storing, analysing and disseminating England's health and social care data.

3.2. The Health and Social Care Act 2012 gives the HSCIC statutory powers to collect identifiable data without data subject consent, including requesting data controllers to supply such data, when directed to do so by Secretary of State, NHS England, National Institute for Clinical Excellence (NICE), Monitor or Care Quality Commission.

3.3. The HSCIC, as data controller for the collected identifiable data, has a responsibility to ensure compliance with both the common law duty of confidentiality and also the eight principles of the Data Protection Act (DPA) including protecting the subjects of that data from inappropriate identification.

3.4. In accordance with the DPA principles, the HSCIC has a duty to minimise its collection, processing, holding and dissemination of identifiable data to that essential for purposes it is to serve - the LEAST principle.[6]

3.5. The Caldicott 2 report[7] emphasises the importance of both sharing data and protecting privacy.Caldicott2 (p66) also recommended that "there should be an evaluation of the benefits, costs, risks and management issues of adopting a system or systems of pseudonymisation at source"

3.6. Having due regard to this background, the HSCIC commissioned a review to understand better the role that pseudonymisation should play in its use of health and care data.

3.7. The overall aim of the review was to:

> *ensure all aspects of the application of pseudonymisation were understood and that it was used appropriately within the HSCIC*

3.8. While the review was underway, resolutions that Primary Care patient data should be pseudonymised at source were passed at the British Medical Association's Local Medical Committees 2014 conference and its Annual Representative Meeting. These provide context for the review around GPs' views on patient data.

3.9. This document presents the recommendations arising from the first phase of this review and options for further consideration in the use of pseudonymisation. It proposes criteria against which to evaluate these options and initial considerations to inform these evaluations.

4. Scope

4.1. The scope of the review included:

[6] The LEAST principle as described in Ian Herbert's paper 'Fair Shares for All: Sharing and protecting electronic patient Sharing and protecting electronic patient healthcare data' (report for the BCS Primary Healthcare Specialist Group, March 2012): "the least data, copied the least number of times, held for the least time and used by the least number of people necessary for the purpose",

[7] https://www.gov.uk/government/publications/the-information-governance-review

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

4.2. The scope of the review excluded a number of items per se [8]:

- The use of pseudonymisation in point to point contexts independent of the HSCIC;
- Assessment of the merits of care.data or other HSCIC programmes;
- 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' vs 'opt out', except to gain an understanding of how patient objections could be managed under different pseudonymisation approaches;
- Any general ethical aspects of using identifiable or de-identified data.

5. Review Approach

5.1. The review recommendations were informed by interviews and workshops with knowledgeable individuals who each had a particular interest in and perspective on pseudonymisation. All participants shared their expertise very constructively. There was widespread agreement that collecting, linking and sharing individual level data promises to deliver significant improvements in the prevention and treatment of illness, in the understanding of health and care needs, and as to how well those needs are met by the services that are commissioned. All participants also agreed that the use of de-identified data should be the preferred choice where this is consistent with meeting legitimate needs. Participants were unanimous in acknowledging that there will be some cases where identifiable data is required.

5.2. In addition further consideration was given to the current HSCIC services and discussions took place with the leads of some of these services in order to inform possible models for pseudonymisation.

6. Key points from stakeholder engagement

6.1. Between the end of November 2013 and the end of January 2014, interviews were carried out with 31 subject matter experts. Following the interviews a workshop was held on January 22nd to examine collectively some of the key issues that emerged from the interviews. On January 23rd a webinar was held to look at some of the technical aspects of pseudonymisation. This document provides a brief summary of the key themes and points that were identified through this engagement.

6.2. The main issue that stakeholders discussed was the relative merits of different models of pseudonymisation, whether undertaken at source or centrally. The stakeholders represented a variety of views on this subject which are summarised below:

[8] However the creation of persistent databases of rich linked data covering an increasing portion of patients lifespan (and that therefore will become intrinsically more intrinsic even without any explicit identifiers) will greatly increase the need to minimise the risk of r-identification as far as possible.

6.3. Themes and Key Points:

6.3.1. Importance of trust

6.3.1.1. One of the strongest themes to emerge was the importance of trust.

6.3.1.2. Some stakeholders who favoured pseudonymisation at source suggested that it would help people trust that the HSCIC was taking all necessary measures to safeguard their data.

6.3.1.3. Some participants also expressed the idea that the HSCIC needs to be an exemplar in terms of information governance and data security and some interviewees stated that adopting pseudonymisation at source would contribute to achieving this.

6.3.1.4. Some participants made the specific point that auditable governance controls are needed within the HSCIC to prevent inappropriate re-identification. This could include records of who has accessed data[9], for what purpose, when and why data is shared in identifiable format with third parties.

6.3.1.5. Some participants believed that GPs can be viewed as proxies for their patients in the sense that if GPs trust that data is being used appropriately and with the necessary safeguards in place (such as pseudonymisation) then that trust will be shared by patients. They explained that such trust is important for patients sharing information fully with medical professionals, which underpins good healthcare.

6.3.1.6. Some participants made the point that because of the level of trust that exists between patients and their GPs, primary care records contain data that are more sensitive and so may require enhanced safeguards, as compared with, e.g. Hospital Episodes Statistics (HES) data.

6.3.2. It needs to be established whether there are any uses or benefits that can only be achieved with identifiable data

6.3.2.1. The participants were of the view that collecting, linking and analysing patient level data could provide very significant benefits to the health and care services and the population at large.

6.3.2.2. A significant number of participants advised that there should be greater clarity about how the data may be, and will be, used, and how it will not. They advocated the implementation of the LEAST principle - see 3.4 – so that identifiable data is not held in bulk for long periods – possibly indefinitely – for undefined future uses. It was felt that this principle supported the adoption of pseudonymisation wherever identifiable data are not needed.

6.3.2.3. Linked to the preceding point, some participants suggested that more could be done to make people aware of how much can be achieved using pseudonymised data. It was suggested that an open library be created demonstrating how such data can be used to meet a variety of operational requirements. It was noted that some of the work carried out by CPRD, QResearch, THIN and ResearchOne may provide a useful starting point.

[9] Recommendation 1 from Caldicott 2 review (page 34)

6.3.3. Debate needs to be informed and measured

6.3.3.1. Some participants stated that the dialogue around pseudonymisation and the broader topic of big data was unhelpfully polarised and needed to become more constructive, nuanced and realistic.

6.3.3.2. Some participants cautioned that whilst pseudonymisation may have a role to play, it won't be the case that "adopting pseudonymisation" in and of itself delivers the safe use of data.

6.3.3.3. One participant noted that by using pseudonymisation, a degree of protection is afforded that is in line with a reasonable assumption about what an informed person might choose.

6.3.3.4. Amongst those who were sceptical about pseudonymisation at source, some cited that the cost of implementing it would be very significant, to the point that these costs would be prohibitive although no evidence on costs was provided. This will be explored in the next stage of the review.

6.3.3.5. Some participants gave the view that pseudonymisation at source would provide little if any reduction in the risk of inappropriate re-identification because the data was already being carried through secure, encrypted channels and held in secure datacentres.

6.3.3.6. Related to the idea that discussion needs to be realistic and measured, it was noted by one participant that the risk of inappropriate re-identification can never be reduced to zero, so we should not start with an assumption that the mere presence of risk is unacceptable.

6.3.3.7. There was a significant level of disagreement between participants about whether the reliable linkage of data requires, or is even affected by, the data being identifiable or not.

6.3.3.8. A related but importantly separate question was also raised by some participants in respect of data quality and the risk that pseudonymised data would not afford the same opportunities for data quality issues to be identified and addressed. This can be explored in the next stage of the review.

6.3.4. Additional Points

6.3.4.1. The topics of different consent models and local versus national databases cropped up a number of times in the course of the review, but were not directly relevant to the topic of data pseudonymisation. However they do affect the need for, and deployment of, pseudonymisation at source.

7. Pseudonymisation required in different circumstances

7.1. Chapter 7 of the Information Commissioner's Office (ICO) Code of Practice on Anonymisation outlines that different levels of granularity of released data are appropriate dependent on the other safeguards to minimise the risk of re-

identification. It identifies that published data need to be fully anonymised. The HSCIC typically does this through aggregation and small number suppression. This is because anyone (including the individual themselves) could attempt to re-identify individuals within the dataset using a wide range of other information at their disposal.

7.2. The Code of Practice identifies that more granular information is frequently needed for richer analysis, for example in research. It outlines that data made available under limited access with robust controls in place can be more granular and it would be for this type of data release that pseudonymisation or anonymisation should be considered. Further guidance from the ICO outlines that the level of risk of re-identification depends upon the richness and type of data released to the recipient and the range of other data that the recipient has or could have access to. In reality, this would be difficult to assess, unless access were provided in a controlled environment that did not allow other datasets in and limited information going out, or only limited and carefully selected individual data are released.

7.3. At a minimum, all fields that could be used to identify a person or small group of people (such as NHS number, postcode and date of birth) should be encrypted, removed if not required or blurred.

7.4. However, there are other variables within datasets such as broader demographic information, dates when various contacts with the NHS occurred and types of treatment. These could not, in isolation, be used to re-identify an individual, but could together and in conjunction with enough additional information be used to re-identify someone. These variables also underpin much of the research that can be conducted using datasets, so the utility of the information also needs to be taken into account.

7.5. Further work needs to be undertaken to reach a view on the appropriate levels of granularity for various sets of circumstances. This should consider the level of benefit that can be derived from information of different levels of granularity. It should also look to quantify the extent to which a combination of broader demographic information has the potential for individual re-identification.

8. Potential models for pseudonymisation of data collected by HSCIC

8.1. A range of controls should be utilised to guard against inappropriate identification, as outlined in the Information Commissioner's Office Code of Practice for Anonymisation. These should cover the collection, storage, generation (including linking) and the dissemination of data about individuals by the HSCIC. These controls should cover policy, security, governance and transparency. Pseudonymisation should be considered alongside these other controls.

8.2. External customers of the HSCIC who have their access approved based on a legitimate purpose to access health data should be given access to data with the lowest possible risk of re-identification that meets this purpose. Pseudonymisation and anonymisation are two of a wider range of techniques that can be applied to data leaving the HSCIC to deliver this. The HSCIC currently uses pseudonymisation techniques; there should be additional external assurance undertaken as regards the processes and technologies that are employed.

8.3. The remainder of this section focuses on inbound data flows, as this area has a much wider range of possible approaches.

8.4. Three broad possible models for pseudonymising the data collected by the HSCIC have been identified for further consideration in the next stages of the review:

- Model 1: pseudonymise centrally, whereby all datasets held centrally are pseudonymised centrally;
- Model 2: pseudonymise at source, whereby all datasets held centrally are pseudonymised at source;
- Model 3: a mixed model, whereby some datasets held centrally are pseudonymised at source and others are pseudonymised centrally.

8.5. Model 1 - pseudonymise centrally for all datasets collected by the HSCIC

8.5.1. Under this model, the HSCIC would receive datasets in person-identifiable form where there exist a requirement and approvals, these datasets then being pseudonymised centrally. The personal confidential items would be used for data linkage and data quality checks before being separated from the other data items and replaced with pseudonyms in the datasets accessed by HSCIC analysts. This option would represent the least change from the current operating model of the HSCIC.

8.5.2. The number of HSCIC staff with access to the personal confidential items would be limited and regularly audited, and information about this would be made available transparently.

8.5.3. This model is closest to how the HSCIC currently employs pseudonymisation.

8.5.4. Diagram 1 overleaf shows an example of how this model could operate.

8.6. Model 2 - pseudonymise at source for all datasets collected by the HSCIC

8.6.1. Under this model, all datasets collected by the HSCIC would be pseudonymised at source. Consideration would be needed as to the precise scope of this. In addition to the activity record datasets covering secondary care and mental health, it could include National Tariff Services, the Medical Research Information Service (MRIS) component of the data linkage service and data held by Data Services for Commissioners Regional Offices. Although the scope of this review covers only data captured by the HSCIC, this model could in principle cover all NHS-to-NHS data flows for secondary uses.

8.6.2. This model would represent a substantial change to NHS systems and to central HSCIC systems and business processes, the extent of which would need to be understood. The importance of the functions that currently use the clear data would also need to be assessed along with an evaluation of whether they could be delivered using pseudonymised data.

8.6.3. If there were a plan for the HSCIC to link data sets across care settings, data submissions from providers across these care settings would need to be pseudonymised using the same key to enable that specific linkage. Different keys could be used for different linkages, but again they would need to be consistent across all care settings from where data would be linked. Therefore, a very secure key management function would be required.

8.6.4. Diagram 2 overleaf presents an example of how this model could operate.

8.7. Model 3 – mixed model

8.7.1. Under this model, some datasets (including, but possibly not limited to, GP data) would be pseudonymised at source whereas other HSCIC datasets would be pseudonymised centrally. If there were a plan to link data sets across care settings, source and central pseudonymisation would need to be performed using the same key and salt code to enable that linkage to take place. Keys could be refreshed periodically and different keys could be used for different linkages.

8.7.2. Some interviewees described their view of the unique nature of the GP dataset, arising from the individual relationship between the GP and the patient and patients' expectations of greater levels of privacy built up over a number of years, supporting a case for greater controls around the primary care dataset.

8.7.3. There are also fewer data sources for GP information than for information covering other sectors due to the fact that there are only four GP system suppliers.

8.7.4. The mixed model would help guard against casual or accidental person identification for the pseudonymised datasets. Unauthorised re-identification of pseudonymised data would be illegal. An important question is how such a model could be set up with the necessary auditable governance structures and boundaries to add further safeguards against the risk of data pseudonymised at source being re-identified. This would need to be considered for the HSCIC data repository and for any other customers of the different sets of data. Ensuring transparency as to who has access and the safeguards in place would be central.

8.7.5. Depending on which model is chosen at the next stage of the review, this mixed model could be considered as either an interim or final model, subject to the evaluation of options. Diagram 3 overleaf presents an example of how this model could operate.

Diagram 1 – an example of how model 1 (pseudonymise centrally) could operate. This diagram is not intended to be a comprehensive map of data flows or a technical specification, but as a starting point for further consideration:

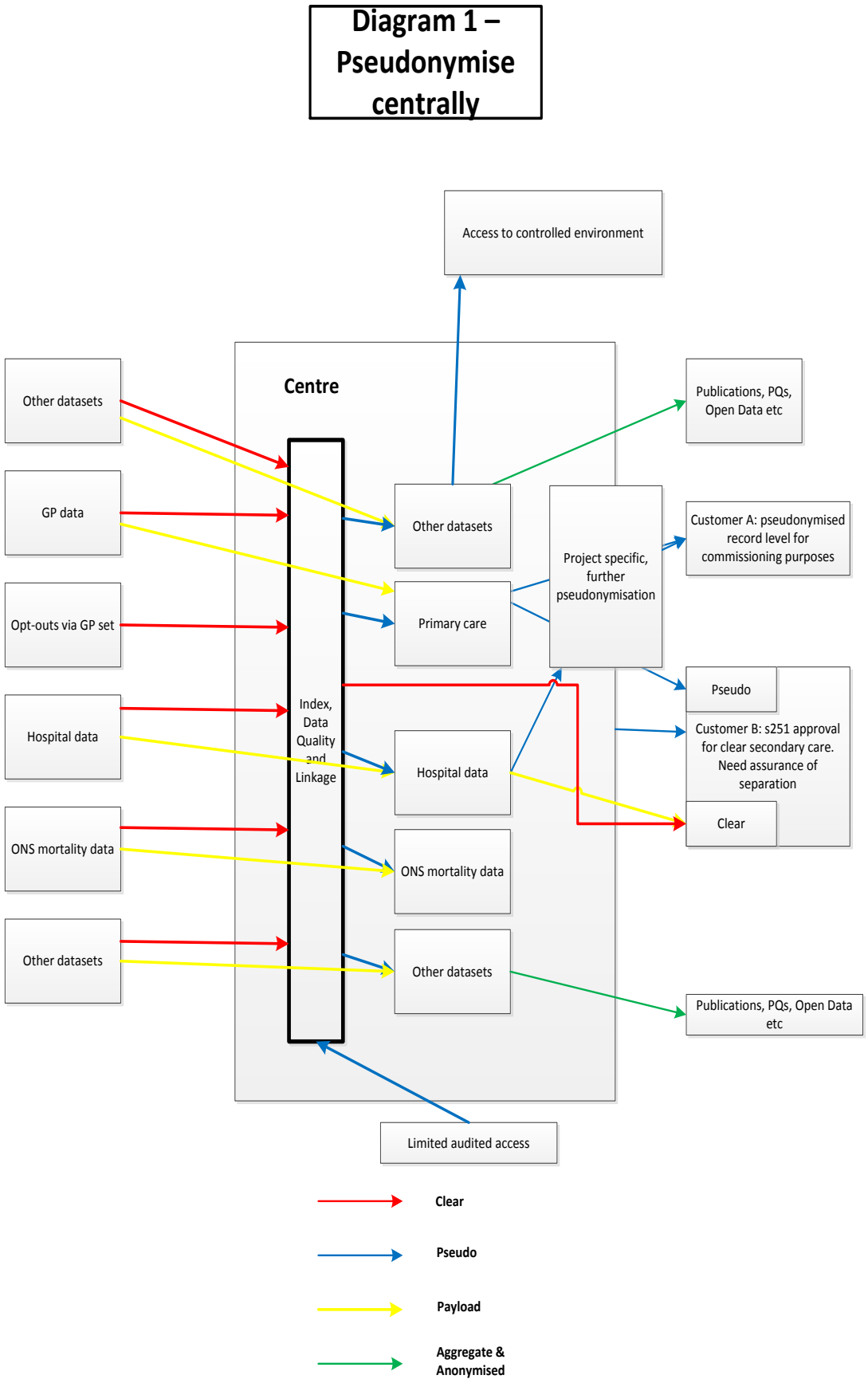


Diagram 2 – an example of how model 2 (pseudonymise at source) could operate. This diagram is not intended to be a comprehensive map of data flows or a technical specification, but as a starting point for further consideration:

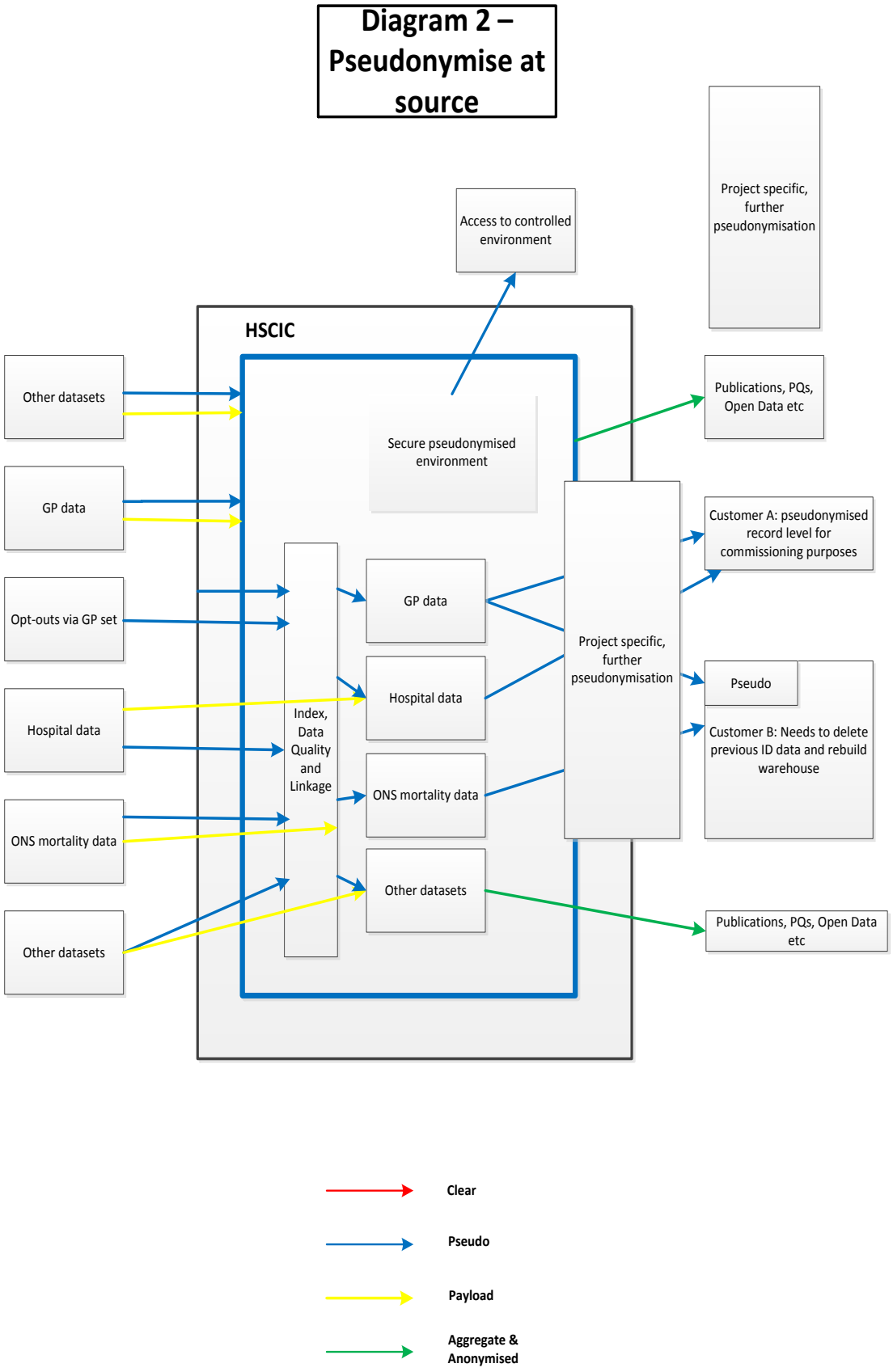
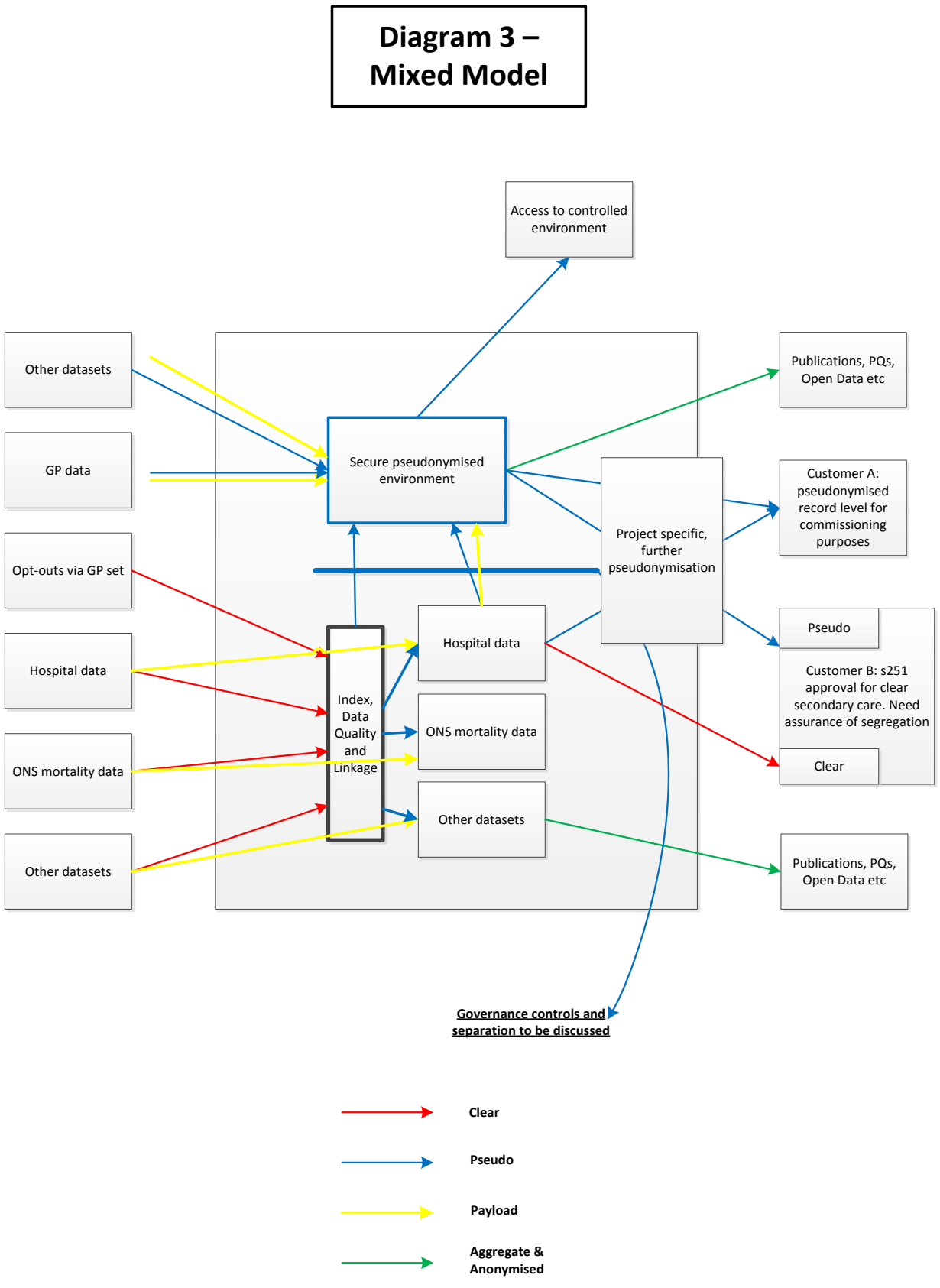


Diagram 3 – an example of how model 3 (mixed pseudonymisation model) may operate. This diagram is not intended to be a comprehensive map of data flows or a technical specification, but as a starting point for further consideration:



9. Criteria for evaluation of models

9.1. This review has identified some criteria against which the models can be evaluated and provides some initial considerations. The next stage of the review can evaluate the models against these criteria.

9.2. Security, controls, governance and transparency

9.2.1. Under all models, within the HSCIC, the number of individuals with access to personal confidential data needs to be minimised, auditable and transparent. The risk and impact of such individuals gaining unauthorised access should also be minimised. The extent to which each model can minimise these should be assessed by the Pseudonymisation Review.

9.2.2. The model that is chosen, in conjunction with the range of other controls in place to guard against inappropriate patient re-identification, needs to be broadly accepted as a suitable set of safeguards in order to gain the public's trust.

9.2.3. Pseudonymisation can be unpicked by creating a look-up between the unpseudonymised and pseudonymised values. This could occur through clear and pseudonymised fields coming into contact with each other or through a pseudonymisation key or salt becoming available and used to generate a look-up between clear and pseudonymised data. The safeguards around this that can be put in place for each model should be considered and under all models there should be a secure key management function.

9.2.4. The 'pseudonymise centrally' and 'mixed' models would require central deployment of the pseudonymisation algorithm and keys. This would create the potential for pseudonymised and clear data to come into contact with each other at the HSCIC or at any customer with approved access to both sets of data, thus enabling re-identification. To develop these models, the next stage of the review would need to consider any necessary development to the technical and governance models for pseudonymisation management that sufficiently mitigate the resulting risks. Examples where there is a separation of functions between organisations, such as CPRD linkage, should be considered. The developed models would need to be transparent and auditable.

9.2.5. The complexity of data supply varies markedly between settings. For GP data, there are four main system suppliers who would be involved for data submission. However for other settings, such as acute care, mental health and community, there are currently many more entities responsible for data submission. Applying the 'pseudonymise at source' model across all care settings would involve the same pseudonymisation algorithm, keys and salts potentially shared across thousands of organisations to enable consistent pseudonymisation for linkage. To develop this model, the next stage of the review would need to consider the security arrangements for the management of the keys, salts and pseudonymised data to avoid inappropriate re-identification whatever the source.

9.2.6. Under all models, consideration would need to be given to the effects of periodic changes of keys: there could be a positive impact on security, but the means by which continued performance of time-series analysis could be performed would require further consideration.

9.2.7. Under all models, if different datasets have different granularity of demographic details[10], linking them would open the potential for the more granular demographic details to be made available on the otherwise less-granular dataset. This should be further considered.

9.3. Implementation (cost and time)

9.3.1. This should be examined in more detail in the next stage of the review. Current assumptions are presented below.

9.3.2. The 'pseudonymise centrally' and 'mixed' models – see 8.4 - would require a specialist governance structure, partitions and technical set-up in order to prevent inappropriate re-identification of pseudonymised datasets. Much of this is already in place, but further developments may be needed for which the cost and time would need to be explored.

9.3.3. Securing a pseudonymised inbound flow of GP data is thought to be reasonably straightforward to implement, as two of the four major GP system suppliers already use the technology that can apply pseudonymisation at source. Discussions would be needed to understand the cost and time implications. Separate consideration could then be applied to other new datasets, outside of primary care, which might from time to time start to flow.

9.3.4. Implementing 'pseudonymisation at source' for datasets that the HSCIC already collects is likely to have greater cost and time considerations. In particular, there would be many more data providers to engage with, each of whom would have systems at varying levels of maturity

9.3.5. If datasets currently collected by the HSCIC were pseudonymised at source, derivations from data items such as postcode and date of birth would need to be undertaken at source and sent to the HSCIC. Data providers would need to apply standard lookups and rounding assumptions e.g. through use of k-anonymity [11]. These derivations provide important data for analysis. Examples of these are age or age band (to understand the effect age has on a specific treatment or the prevalence of diagnosis by age), location (that can be used to understand effects of deprivation, for example). Each submitting organisation or system supplier would need to be responsible for:

- Carrying out the derivations
- Sending the derived data items to the HSCIC
- Carrying out the data quality assessment of the raw data used to produce the derived data items
- Producing the data quality reports from those assessments

9.3.6. Many of the data processing systems for existing data flows are designed to receive identifiable data, so are set up to expect certain field lengths and perform certain validation routines. Many of them have grown incrementally; changes would be complex and testing extensive.

[10] for example if the GP dataset has only age range, but the hospital dataset has month and year of birth to enable analysis of neonatal conditions

[11] A criterion to ensure that at least k records in a data asset has the same quasi-identifier values. See Standard ISB 1523- http://www.isb.nhs.uk/library/standard/128

9.3.7. There exist external organisations, such as the Care Quality Commission, that currently receive clear data and build their own databases to discharge their statutory functions. They would need to make similar changes to their systems.

9.4. Any functions that some models cannot deliver?

9.4.1. This section outlines the HSCIC functions that are currently delivered using personal identifiable data, such as data linkage, and provides an assessment of the steps needed to evaluate whether they can be delivered using pseudonymised data.

9.4.2. The HSCIC's privacy impact assessment states that the HSCIC should:

> "obtain and process only the minimum necessary patient identifiable data from other organisations as required by the DPA act 1998".

9.4.3. This is similarly stated in the LEAST principle. This makes particularly important the question of whether there are any functions that could only be delivered with clear personal identifiers – this question should be answered on the basis of what could be done rather than on what has always been done.

9.4.4. A number of important functions performed by the HSCIC are currently reliant on using clear data. To evaluate the models, detailed consideration needs to be given to whether each of them can be undertaken using pseudonymised data. Where the use of pseudonymised data is possible, the cost and time of migration needs to be assessed; where it is not possible, the impact of that function not being delivered in the health and social care system needs to be assessed. Some functions are identified below for further consideration at the next stage of the review.

9.4.5. Identification of data validity problems such as invalid NHS numbers and use of inappropriate default dates of birth and postcode

9.4.5.1. Regardless of which model is adopted, poor data quality will impact on the ability to pseudonymise all person confidential data effectively, which could result in sub-optimal analysis and decision making. This is in addition to the impact it may already be having on direct patient care. Therefore, the levels of pseudonymisation failure, and the reasons for it, need to be understood to ensure that statistical bias, or error, is not introduced to the subsequent analysis of that data. Whoever carries out the pseudonymisation has a responsibility to:

- Assess the quality of the data used for pseudonymisation against mandated standards
- Make the results of those assessments available to all prospective users of the pseudonymised data to inform their view of its fitness for their purposes: If pseudonymisation were carried out at source, each submitting organisation would be responsible for:
  - Carrying out the data quality assessments on the data items used in the pseudonymisation process
  - Producing the data quality reports from those assessments

These assessments would need to be produced in a standard form and collated centrally for onward transfer to customers to enable them to assess whether the quality of their data is suitable for their purposes.

9.4.6. Delivery of effective data linkage

9.4.6.1. Detailed further work is required on data linkage, as the review confirmed the existence of different views around this.

9.4.6.2. Data linkages that are purely reliant on entire matching of values in fields between and within datasets would be unaffected. Data linkages that are reliant on some characteristics of values within the field may be affected. The impact would be less where the data quality is high for the personal identifiers (NHS number, postcode, date of birth) and consequently their pseudonymised versions.

9.4.6.3. Most data linkages performed in the HSCIC (for example HES to Mental Health Minimum Dataset – 'MHMDS') use a deterministic approach using various combinations of person-identifier fields (NHS number, date of birth, postcode, sex). They make use of various features inherent in the clear value of date of birth and postcode including the use of subsets of the field elements of date of birth or of treating certain values within the field differently for example postcodes for communal establishments.

9.4.6.4. By way of example, for the HES to MHMDS linkage used to produce the August 2013 publication[12], of the 4.2 million HES and MHMDs records, around 135,000 matches were informed by specific or partial values of postcode and date of birth and would attain a different match score. Further work is needed on assessing this impact for other datasets.

9.4.7. Data quality feedback loop between centre and providing organisations

9.4.7.1. The HSCIC and data suppliers currently discuss specific sets of records where data quality problems have been identified. Personal confidential data is currently used to highlight these records, so an alternative approach that is not reliant on patient confidential data would be needed.

9.4.8. External customers

9.4.8.1. A small number of other organisations receive personal confidential secondary care data to fulfil their statutory functions or with 'Section 251' approval[13]. Examples include Care Quality Commission and Public Health England. They build historical databases from datasets such as HES, so detailed discussions would be needed with them to understand the impact of the different pseudonymisation models on their ability to perform these functions. A previous exercise was undertaken to roll out an identifier in HES that was pseudonymised to a different key for each customer (the HES ID), so lessons could be learnt from this approach.

9.4.9. Other HSCIC services that are currently dependent on using Patient Confidential data

Work would be needed to ascertain whether elements of these services could be delivered with pseudonymised rather than clear data. Examples include:

9.4.9.1. The Medical Research Information Service (MRIS) component of Data Linkage and Extract Service. This service enables cohort tracking for

[12] http://www.hscic.gov.uk/pubs/hesmhmdslink1112add

[13] Approval in respect of arrangements made under s.251 of the National Health Service Act 2006 (2006 c. 41), wherein the Secretary of State may make regulations for the setting-aside of the common-law duty of confidentiality.

approved research studies using a range of details, including person name, and cleaning of patient lists.

9.4.9.2. Payment by Results Secondary Uses Service, which currently uses personal identifiers to construct spells to generate prices to inform commissioners how much to pay NHS providers.

9.4.9.3. Data Services for Commissioners Regional Offices (DSCROs) which provide a range of data linkage services between national and local data flows to enable CCGs and GP practices to perform commissioning functions.

9.4.9.4. The impact on these services of using pseudonymised data is likely to be very high, so the next stage of the review would need to quantify this and determine the overall importance of these services.

9.5. Patient indexing and management of opt outs

9.5.1. Via the GP extract, patients can register two types of objections in respect of their information:

9.5.1.1. 'Type 1 objections' encode an objection to the information leaving the GP practice;

9.5.1.2. 'Type 2 objections' encode an objection to any of their data from any of their datasets leaving the HSCIC in identifiable form.

9.5.2. Patients can also apply directly to the HSCIC to have their personal confidential data removed, under Section 10 of the Data Protection Act 1998, on the basis that processing it causes them substantial distress.

9.5.3. Detailed consideration is needed as to how Type 2 objections and Section 10 requests could be implemented under the three models. In particular, consideration should be given to whether identifiers need to flow in clear to enable implementation of these requests and whether these would need to be pseudonymised centrally.

10.Annex 1: Glossary

Concept	Definition
Aggregated Statistics	Statistical data about several individuals that has been combined to show general trends or values without identifying individuals within the data. Values determined to be of 'small numbers' are supressed to minimise risk of unauthorised personal identification, either through blurring or through omission altogether. (Caldicott review: information governance in the health and care system).
Anonymisation	The process of rendering data into a form which does not identify individuals and where identification is not likely to take place (ICO Anonymisation Code of Practice). Record level anonymised extracts for public access would require both removal of fields and k-anonymity suppression. Aggregate level tabulations for public access would require small numbers suppression in addition to the aggregation.
Anonymised Data	Data in a form that does not identify individuals and where identification through its combination with other data is not likely to take place (ICO Anonymisation Code of Practice). As such, anonymised data cannot be linked at a patient or service-user level. Anonymised data may be published both at record and aggregate level (ISB Standard 1523: Anonymisation Standard for Publishing Health and Social Care Data).
Caldicott Guardian	A senior person responsible for protecting the confidentiality of patient and service user information and enabling appropriate information sharing by providing advice to professionals and staff (Caldicott review: information governance in the health and care system).
Care.data	The Care.data programme will enable HSCIC to make the necessary step change to respond to the projected increased demand for data by increasing the breadth of data which is collected, linked and disseminated whilst protecting personal confidential data and reducing burden on the system. Care.data will enable investigation on what and how care is delivered, both in individual care settings and across care pathways and geographies.
Care Quality Commission	The Care Quality Commission (CQC) makes sure hospitals, care homes, dental and GP surgeries and all other care services in England provide people with safe, effective, compassionate and high quality care, and encourages these services to make improvements.

Care records	Care records are personal records that comprise documentary and other records concerning an individual (whether living or dead) who can be identified from them and relating: • to the individual's physical or mental health and healthcare • to spiritual counselling or assistance given or to be given to the individual • to counselling or assistance given or to be given to the individual, for the purposes of their personal welfare, by any voluntary organisation or by any individual who: – by reason of the individual's office or occupation has responsibilities for their personal welfare; or – by an order of a court has responsibilities for the individual's supervision. This record may be held electronically or in a paper file or a combination of both (**Caldicott review: information governance in the health and care system**).
Clear data	Data that are identifiable
Clinical Practice Research Datalink (CPRD)	The Clinical Practice Research Datalink (CPRD) is the new English NHS observational data and interventional research service, jointly funded by the NHS National Institute for Health Research (NIHR) and the Medicines and Healthcare products Regulatory Agency (MHRA). CPRD services are designed to maximise the way anonymised NHS clinical data can be linked to enable many types of observational research and deliver research outputs that are beneficial to improving and safeguarding public health.
Consent	Consent is the approval or agreement for something to happen after consideration. For consent to be legally valid, the individual must be informed, must have the capacity to make the decision in question and must give consent voluntarily. This applies to both explicit and implied consent. **Explicit consent** can be given in writing or verbally, or conveyed through another form of communication such as signing. A patient may have capacity to give consent, but may not be able to write or speak. Explicit consent is required when sharing information with staff who are not part of the team caring for the individual. It may also be required for a use other than that for which the information was originally collected, or when sharing is not related to an individual's direct health and social care. **Implied consent** is applicable only within the context of direct care of individuals. It refers to instances where the consent of the individual patient can be implied without having to make any positive action, such as giving their verbal agreement for a specific aspect of sharing information to proceed. **(Caldicott review: information governance in the health and care**

	system).
Data breach	Any failure to meet the requirements of the Data Protection Act, unlawful disclosure or misuse of personal confidential data and an inappropriate invasion of people's privacy (**Caldicott review: information governance in the health and care system**).
Data controller	A person (individual or organisation) who determines the purposes for which and the manner in which any personal confidential data are or will be processed. Data controllers must ensure that any processing of personal data for which they are responsible complies with the Act • Joint data controllers control how data is processed jointly, i.e., they must agree and make such decisions together. • Data controllers in common agree to pool data and are both responsible for how it is used but each may process the data independently for its own purposes. All of the data controllers in common are still responsible for ensuring it is adequately protected (**Caldicott review: information governance in the health and care system**).
Data linkage	A technique that involves bringing together and analysing data from a variety of sources, typically data that relates to the same individual (**ICO Anonymisation Code of Practice**).
Data processor	In relation to personal data, means any person (other than an employee of the data controller) who processes the data on behalf of the data controller. Data processors are not directly subject to the Data Protection Act. But the Information Commissioner recommends that organisations should choose data processors carefully and have in place effective means of monitoring, reviewing and auditing their processing and a written contract (detailing the information governance requirements) must be in place to ensure compliance with principle 7 of the Data Protection Act. (**Caldicott review: information governance in the health and care system**)
Data Protection Act	The Data Protection Act 1998 (DPA) is a United Kingdom Act of Parliament which defines UK law on the processing of data on identifiable living people. It is the main piece of legislation that governs the protection of personal data in the UK.
De-identified data	Data treated so that it lowers the risk of individuals being identified and will not breach confidentiality. There are 2 types of de-identified data: 1. De-identified data for publication - data that can be publicly disclosed as it has been anonymised and there is a low risk of individuals being identified. 2. De-identified data for limited disclosure or access - data that has been through a process of pseudonymisation; however there remains a risk of individuals being identified.

	De-identification may include pseudonymisation, application of derivations or removal of fields (**HSCIC Data Linkage and Extract Service website** & **Caldicott review: information governance in the health and care system**).
Derivations	Identifying items may hold information required by a customer for analysis. These may be replaced by derivations which blur, group or band the item so that the risk of identification decreases. For example, postcode can be grouped into super output area (SOA); postcode can be blurred to partial postcode (e.g., LS1 ***); date of birth can be banded into 5 year age group (e.g., 30 – 35 years old) (**Pseudonymisation Implementation Project – Guidance on Terminology**).
Direct care	A clinical, social or public health activity concerned with the prevention, investigation and treatment of illness and the alleviation of suffering of individuals. It includes supporting individuals' ability to function and to improve their participation in life and society. It includes the assurance of safe and high quality care and treatment through local audit, the management of untoward or adverse incidents, person satisfaction including measurement of outcomes undertaken by one or more registered and regulated health or social care professionals and their team with whom the individual has a legitimate relationship for their care (**Caldicott review: information governance in the health and care system**).
Enterprise De-Identification Solution (EDS)	EDS will ensure that secondary use data (i.e., not for direct Patient or Service User care) can be provided to interested parties whilst ensuring that public expectations and legal protections on the confidentiality are met. Enterprise wide in scope, the Programme's ambition is that all HSCIC secondary use data requiring de-identification for external or internal purposes will flow through the EDS.
Hospital Episode Statistics	HES is a data warehouse containing details of all admissions, outpatient appointments and A&E attendances at NHS hospitals in England. This data is collected during a patient's time at hospital and is submitted to allow hospitals to be paid for the care they deliver. HES data is designed to enable secondary use, that is used for non-clinical purposes, of this administrative data. It is a records-based system that covers all NHS trusts in England, including acute hospitals, primary care trusts and mental health trusts. HES information is stored as a large collection of separate records - one for each period of care - in a secure data warehouse. The HSCIC applies a strict statistical disclosure control in accordance with the HES protocol, to all published HES data. This suppresses small numbers to stop people identifying themselves and others, to ensure that patient confidentiality is maintained.

Identifier	A data item that used individually or used in combination with other items could reveal the identity of a person (**ISB Standard 1523: Anonymisation Standard for Publishing Health and Social Care Data)** **Direct Identifier:** Any data item that on its own could be used to uniquely identify and individual, including name, address, widely used unique person ID (e.g., NI Number, NHS Number, Local Hospital Number), telephone number, email address, etc. **Indirect Identifier:** A data item that when used in combination with other items could reveal the identity of a person (including postal code, gender, date of birth, event date or a derivative of one of these items).
k-anonymity	Suppression of data items to ensure that there are at least k records in a data asset that have the same indirect identifier values. For example, if the indirect identifiers are age and gender, then there must be at least k records with 45-year old females. It is necessary to remove any direct identifiers in order to satisfy k-anonymity (**ISB Standard 1523: Anonymisation Standard for Publishing Health and Social Care Data**).
Patient or Service User	Specifically refers to a user of NHS or Social Care Services, not a user of HSCIC services.
Personal Confidential Data (PCD)	Personal information about identified or identifiable individuals, which should be kept private or secret. For the purposes of this document 'Personal' includes the Data Protection Act (1998) definition of personal data, but it is adapted to include dead as well as living people; 'confidential' includes both information 'given in confidence' and 'that which is owed a duty of confidence' and is adapted to include 'sensitive' as defined in the Data Protection Act. This type of data may only be shared when there is a lawful basis to do so, e.g. with Patient or Service User consent or approval under section 251 of the National Health Service Act 2006. (**Caldicott review: information governance in the health and care system**).
Processing	Processing in relation to information or data, means obtaining, recording or holding the information or data or carrying out any operation or set of operations on the information or data, including: • organisation, adaptation or alteration of the information or data; • retrieval, consultation or use of the information or data; • disclosure of the information or data by transmission, dissemination or otherwise making available; or • alignment, combination, blocking, erasure or destruction of the information or data. (**Caldicott review: information governance in the health and care system**).

Pseudonymisation	The process of replacing identifiers in a record with alternate identifiers (pseudonyms), from which identities of individuals cannot be intrinsically inferred, for example replacing an NHS Number with another random number, or replacing an address with a location code. Pseudonyms themselves should not contain any information that could identify the individual to which they relate (e.g. should not be made up of characters from the date of birth etc.) (**Pseudonymisation Implementation Project – Guidance on Terminology**). See also : (**ICO Anonymisation Code of Practice**)
QResearch	QResearch is a working example of a multisource data linkage project done at scale (15 million patients) where the data were all pseudonymised at source using free open source software
Re-identification	The process of analysing data or combining it with other data with the result that individuals become identifiable. Also known as 'de-anonymisation' (**ICO Anonymisation Code of Practice**). & The process of discovering the identity of individuals from a data set by using additional relevant information (**ISB Standard 1523: Anonymisation Standard for Publishing Health and Social Care Data**)
ResearchOne	A working example of data linkage which is done on data which has been pseudonymised at source
Scottish Primary Care Information Resource (SPIRE)	The SPIRE project is collaboration between the Scottish Government and NHS National Services Scotland (NHS NSS) to develop a new service to simplify and standardise the process for extracting data from GP practice systems for a number of purposes e.g. audit, disease surveillance, benchmarking, planning, research and QOF payments.
Section 251	Section 251 of the National Service Act 2006 allows the Secretary of State to make provisions for the common law duty of confidentiality to be overridden to enable disclosure of confidential patient information for medical purposes, where it was not possible to use anonymised information and where seeking consent was not practical, having regard to the cost and technology available.
THIN	THIN is collaboration between INPS and Cegedim Strategic Data Medical Research UK (CSD MR UK). INPS has written unobtrusive data collection software for THIN, which is incorporated into Vision. CSD MR UK, whose staff were instrumental in developing the GPRD in the late 1980s, is an organisation providing access to research data. The staff at CSD MR UK has spent over 20 years facilitating the research use of UK GP Primary Care databases. CSD MR's clients include prestigious academic research groups such as the Universities of Nottingham and Pennsylvania, as well as major pharmaceutical companies.