Pseudonymisation Implementation Project (PIP)

Reference Paper 3

Guidance on De-identification

Final v1.0 - 20 November 2009

Guidance on De-identification			
Programme	NPFIT	Document Record ID Key	
Sub-Prog / Project	Pseudonymisation Implementation Project (PIP)	NPFIT-PIP-GUIDANCE-3NPFIT-FNT-TO-BPR-0025.01	
Prog. Director	J Thorp	Version	01
Owner	.	Status	Final
Author	Wally Gowing	Version Date	20 November 2009

Document Status:

This is a controlled document.

Whilst this document may be printed, the electronic version maintained in FileCM is the controlled copy. Any printed copies of the document are not controlled.

Related Documents:

These documents will provide additional information.

Ref	Doc Reference Number	Title	Version
1	NPFIT-FNT-TO-BPR-0022.01	PIP Implementation Guidance	FV1
2	NPFIT-FNT-TO-BPR-0023.01	Reference Paper 1 - Terminology[1]	FV1
3	NPFIT-FNT-TO-BPR-0024.01	Reference Paper 2 – Business Processes and New Safe Havens[2]	FV1
	NPFIT-FNT-TO-BPR-0025.01	Reference Paper 3 – De-identification[3]	FV1
4	TBA	Reference Paper 4 – Technical White Paper[4]	FV1
5	dh_4069254	NHS Code of Practice on Confidentiality[5]	
6	NA	PIP Planning Template and Guidance[6]	

[1] http://www.connectingforhealth.nhs.uk/systemsandservices/sus/delivery/pseudo

[2] http://www.connectingforhealth.nhs.uk/systemsandservices/sus/delivery/pseudo

[3] http://www.connectingforhealth.nhs.uk/systemsandservices/sus/delivery/pseudo

[4] http://www.connectingforhealth.nhs.uk/systemsandservices/sus/delivery/pseudo

[5] 5 www.dh.gov.uk/en/Managingyourorganisation/Informationpolicy/Patientconfidentialityandcaldicottguardians/DH_4100550

[6] http://www.connectingforhealth.nhs.uk/systemsandservices/sus/delivery/pseudo

Table of Contents

Background to Guidance on De-identification

1 Introduction

1.1 Purpose and scope

This paper is one of a set of documents produced as part of the Pseudonymisation Implementation Project (PIP) and provides a reference document for local organisations implementing NHS wide guidance on local NHS data usage and governance for secondary uses.

The purpose of this paper is to:

- Set out those factors that impact on developing the guidance concerning de-identification and pseudonymisation
- Provide informatics staff with the background and rationale for proposed de-identification regimes by acting as reference document
- Provide specific guidance for users on the basis of their organisation and local circumstances.

The scope of this paper is limited the new safe haven and business process aspects of the implementation to support secondary uses.

1.2 Related papers

This guidance builds on the earlier PIP documents

- PIP Implementation Planning Guidance
- PIP Maturity Model

This guidance is a supporting document to the Implementation Guidance on Local NHS Data Usage and Governance for Secondary Uses.

The other reference documents are shown in the related document reference section at the front of this document.

The Techniques White Paper (Ref 4) provides some additional and specific guidance in support of this, the De-identification Reference Paper.

1.3 Background and context for implementing Local Data Usage and Governance

The background and context for implementing local data usage and governance for secondary uses is covered in depth in Sections 1 and 2 of the Implementation Guidance on Local NHS Data Usage and Governance for Secondary Uses. This should be consulted to provide the overall context for De-identification and impact on NHS business processes.

Of particular significance is the requirement for a secure environment to exist in implementing de-identification.

1.4 De-identification guidance - Work in Progress

The guidance on de-identification has been developed from applying principles and lessons learnt in the development and use of pseudonymisation in SUS. This guidance covers the main issues involved in de-identification and the White Paper on techniques will provide supplementary information. However, issues will arise as implementations commence and as experience is gained. Therefore the guidance relating to de-identification should be regarded as work in progress and will be extended via the SUS websites as and when appropriate.

1.5 Context of De-identification for external use

NHS organisations may allow their data to go via extract to external bodies for specific use, such as specialist analysis (e.g. from a PCT to a university). If this is carried out under contract and the external body is not a recognised Data Processor under the Data Protection ACT (DPA), then it should be regarded as 'external use' within this paper.

For this situation, the NHS organisation is acting as pseudonymisation service provider, which means that the relevant ISO standard 25237:2008 Health Informatics Pseudonymisation[7] should be followed. The environment for this usage of the data should be regarded as potentially hostile and additional safeguards as per the ISO standard are required.

1.6 Patient label

Throughout this and the related papers, the term 'patient label' is used to describe the data item(s) that distinguishes one patient from another in a set of data; please note that this is purely for the purposes of clarity within the papers.

For identifiable data, the patient label may be the NHS Number (if present) but could be within an organisation, its Local Patient Identifier; another data item, or combination of data items, that can be used to uniquely identify one patient from another; in de-identified data sets, patient labels will vary but, for example, can be pseudonyms or table row numbers (the latter may not be unique as activity relating to the same patient may appear more than once in a table.)

1.7 Technical content

The contents of this paper are very technical in places, particularly Section 4 on Techniques. This is necessary in order to specify requirements and methods. The White Paper associated with this Reference Paper will provide further guidance and clarification.

[7] **http://www.iso.org/iso/catalogue_detail?csnumber=42807**

2 De-identification – aims and means

2.1 Aim of De-identification

There is an overarching Information Governance principle that users should only have access to those data that are necessary for the completion of the business activity which they are involved in. This is reflected in Caldicott Principles 1, 2 and 3, see Table 1 in the main guidance document. This principle applies to the use of patient level data for secondary or non-direct care purposes. The utilisation of de-identification tools enables users to make use of patient level data for a range of secondary purposes without having to access those data items, which may reveal the identity of the patient.

The aim of de-identification is to obscure person/patient identifier data items and combinations of them within patient records sufficiently that the risk of potential identification of the subject of a patient record is minimised to acceptable levels so as to be regarded as 'effective anonymisation'.

It should be noted that the risk of identification can never be totally eliminated. This is because a particular user of data may be familiar with individuals with specific or unique characteristics and therefore be able to identify them, or selecting such individuals combined with additional research could achieve identification through use of inference techniques.

The use of de-identification tools does not replace the general requirement to restrict access to data to that needed for a particular purpose. It should be complimentary. For example, if a report or piece of business analysis is undertaken as a monthly time series, there may be no need for users to access the full date of admission or discharge of a patient.

All organisations and individuals should be aware of their responsibilities to respect patient confidentiality and comply with the law, as embodied in Caldicott Principles 5 and 6 and the NHS Code of Practice on Confidentiality.

2.2 Person identifiers

Those data items within patient activity data that are considered to be sensitive are based on the rules laid down for the operation of the Secondary Uses Service SUS). There are 37 fields within the Commissioning Data Sets (CDS) submitted to and held within the SUS system that have been designated by the predecessors of National Information Governance Board (NIGB) Ethics and Confidentiality Committee (ECC) (the Patient Information Advisory Group (PIAG) and the related Security and Confidentiality Advisory Group (SCAG)) as sensitive in relation to identifying individual patients – see Figure 3. Those data items that may lead to identification of individuals can be grouped and shown (with examples but not limited to), as below:

- Strong identifiers – :
  - name (previously only held in SUS for overseas visitors & limited other cases, but not now submitted in CDS)
  - address (as above), postcode
- Weak or indirect identifiers:
  - supporting personal information – date of birth, ethnicity
  - operational reference numbers with meaning within the NHS - NHS Number, local patient identifier

The fields considered as ‘sensitive’ in relation to identifying patients in SUS are:

- Patient Name
- Patient Address
- Patient Date of Birth
- Patient Postcode
- Patient NHS Number
- Patient Ethnic Category
- Patient Local Patient Identifier
- Patient Hospital Spell Number
- Patient Pathway Identifier
- Patient Unique Booking Reference Number
- Patient Social Service Client Identifier.

In addition if the record is for a baby or a mother, the relevant fields contained in the record for the mother or baby respectively for name, address, date of birth, postcode and NHS Number are regarded as sensitive as well.

If concatenated forms of names or addresses are held in any fields, these should also be regarded as sensitive.

Additional potentially sensitive fields associated with SUS and CDS:

- Date of death – this is a data item not explicitly output in CDS, but can potentially be derived from method and date of discharge fields for hospital patients, and is potentially held within local systems. This is also regarded by ECC as sensitive, as it may lead to ready identification of individuals. The Data Protection Act does not apply to deceased persons, but the duty of confidence continues after a patient has died.
- When processing clinically sensitive data, the SUS Spell Identifier should not be used because of the danger of accidental linkage. The SUS Spell Identifier may refer to an episode that has been anonymised for legal reasons and linkage may be possible to earlier episodes containing identifiable data.
- The CDS Unique Identifier may be constructed from personal information which is can be readily interpreted in clear
- SUS Generated Record Identifier should not be used as a means to break pseudonymisation through linkage.

If any of the above data items in addition to any of the earlier 37 data items are included in any other local data and are transferred between organisations they must be considered sensitive and the same principles guiding their use and de-identification are assumed to apply.

2.3 Sensitive data in relation to different users

Whilst data items may be considered as sensitive as they may enable identity to be derived, the level of sensitivity is dependent on the user and the data sources and systems to which the user has access.

The NHS Number is technically a pseudonym as a one-to-one permanent mapping between a person and a number. It is considered a sensitive data item as a wide range of NHS employees (and latterly social services organisations) have access to systems that provide names and addresses against NHS Number; in effect it is a universal identifier within the NHS.

Other data items are considered within the secondary uses domain as ‘restricted identifiers’ as the data item is only generated and used within a specific organisations, so that facilities to derive identifiable data are restricted to the organisation and its systems.

An example is the Local Patient Identifier (LPI), which is a reference number assigned to a patient within a provider organisation to assist in the operation of the local patient administration system (PAS) and associated filing systems for case notes. The patient identity can be readily found from the LPI within the provider organisation creating and using the LPI; however this is not the case outside the originating provider organisation. Therefore, the LPI need not be pseudonymised within a related commissioner. The LPI may also be used to communicate from the commissioner to the provider without use of the NHS Number, except where the commissioner has access to the relevant PAS (which may happen with community hospitals).

2.4 Means of De-identification

De-identification of patient records can be achieved through all or a combination of:

- Not displaying sensitive data items
- Using derivations to replace the values of certain data items in systematic ways, such as using, :
    - electoral ward instead of postcode, displaying age instead of date of birth
    - banding of values, such as displaying age bands (e.g. 5-10) instead of date or year of birth
    - using post code sector (first 4 characters e.g. DE3 7) instead of the full post code (e.g. DE3 7FZ)
- Using pseudonyms on a one-off basis
- Using pseudonyms on a consistent basis
- Using in a non-sensitive context, data in original form that would otherwise be regarded as sensitive – this technique is known as quarantining of data. An example is the use of a provider’s LPI within a commissioner as illustrated in Section 2.3.

2.5 Pseudonymisation

Pseudonymisation is the method employed with data for secondary uses for de-identifying person identifier data items. When pseudonymisation techniques are consistently applied, the same pseudonym is provided for individual patients across different data sets and over time.

Removal of identifiers can also be used; however, this will prevent de-identified records from being linked.

It is possible to produce consistent pseudonyms using techniques, which do not allow the pseudonym to be reversed to permit the identity of the individual to be determined. The use of irreversible pseudonyms allows the linkage of records for the same individual at the same time as effectively anonymising these records.

2.6 Surrogates

A further term that is used in connection with de-identification and used in stating system requirements and in system design is that of the use of surrogates. A surrogate is a substitute for another entity, such as a data item. A pseudonym can be considered as a substitute data item (or surrogate), such as a replacement for the NHS Number in an activity record. Surrogates can also be used to enable links between data in the pseudonymised and identifiable states.

If surrogates are used as a basis for de-identification, these surrogates must be -

- Unique system-wide (e.g. replacing NHS Number), hence never reused
- System generated (i.e. not created by the user)
- Not manipulable by the user or application
- Without semantic or obvious meaning
- Not composed of several values from different domains.

The use of a surrogate field as a pseudonym for an identifier such as the NHS Number may require a table being maintained giving a one to one correspondence between the surrogate and the identifier.

In this case the surrogate may be used as a primary key for indexing or sorting records, in other linked tables, resulting in the term surrogate key.

2.7 Missing NHS Numbers

The creation of a consistent pseudonym to replace the NHS Number will enable linkage of records relating to the same individual. If the pseudonym is reversible, this will also allow the identity of the individual to be determined under authorised circumstances. If the NHS number is not included these processes are not possible.

There will be legitimate reasons why the NHS Number may not be present, e.g. in anonymised CDS records involving sensitive diagnoses or treatments or an overseas visitor not having an NHS number.

Where the NHS Number is expected to be present and is missing and records are required to be linked, ‘fuzzy matching’ methods can be used to enable the submitted record to be utilised. Linkage must not be undertaken if there is a risk that it will reveal personal details where these are protected by regulation or statute.

2.8 Application of pseudonymisation

Pseudonymisation will need to be undertaken within an individual organisation or a related group of organisations, such as through a shared service in order to enable data relating to individual patients received from NHS wide sources such as SUS to be linked with data transferred directly from care providers. The pseudonyms created will be specific to that organisation or that shared service.

If data is to be made available in pseudonymised form to other organisations, e.g. to support research, then pseudonyms different from those used internally must be used.

2.9 Storing data and keys

Pseudonyms must be held logically separate from identifiable data as part of a layered approach to maintaining the security of personal data.

Identifiable data must only be accessible by authorised users for legitimate and auditable reasons.

It is also required that the values of keys used in encryption and seeds used in hash function operation must be stored separately and securely with access restricted to relevant authorised users.

3 Displaying data

3.1 Displaying data

Users may view data through on-screen or paper reports, or in some cases through extracts from systems which enable subsequent local analysis. The rules set out in Figure 1 and Figure 2 apply to whichever form of display is used with patient data that is to be de-identified.

3.2 Displaying data for NHS Business Use

People making use of data for secondary purposes will be dependent on reports provided to them or data they can access to tailor reports. Therefore, the form of data that can be accessed provides the basis for control to achieve de-identification of patient records.

The display of data for routine business use is shown in Figure 1 and is based on CDS and MHMDS contents. Data used for spatial analysis must be handled differently and this is considered separately in Section 3.4.

Table key - Providers includes secondary, community, ISTP/C and ambulance services; Commissioners includes PCTs and Specialist Commissioning Groups; PHOs are subject to the same rules as Commissioners.

Figure 1 - Display of sensitive data items for routine business secondary use reporting

Data item	Originating Providers	Commissioners
Name	Do not display	
Address	Do not display	
Date of birth*	Replace by age in years	
Postcode**	Postcode sector and/or derivations	
NHS Number	Pseudonymised or do not display	
Ethnic category***	Identifiable if relevant to report, otherwise do not display	
Local patient identifier	Pseudonymised or do not display	Identifiable
Hospital Spell Number	Pseudonymised or do not display; SUS Spell ID may be a suitable proxy	Identifiable
Patient pathway identifier	Pseudonymised or do not display	Identifiable
SUS Spell ID	Identifiable – but do not display if there is a potential to reveal confidential data through linkage****	Identifiable – but do not display if there is a potential to reveal confidential data through linkage****
Unique Booking Reference Number	Pseudonymised or do not display	
Social Services Client Identifier	Pseudonymised or do not display	
Date of death	Truncate to month and year*****	

*** Neo-nates** – requests for information concerning patients of less than one year of age need to be treated as a special case. As analyses can rely on age at time of treatment being measured in days, effectively providing date of birth, safeguards must be taken in releasing such data. A possible route is to provide data on the basis of banding, such as less than 28 days and age by months thereafter.

**** Derivations** – care must be taken with derivations to avoid potential disclosure of data. This is on the basis that Output Area can be as small as 40 households although generally it is greater than 125 households. Unless there is a specific requirement to have the data at this low level then the default position should be to use derivations at Lower Super Output Area level.

***** Ethnic Category** – the nature of the meaning of ethnic category, its sensitivity and the limited number of values means that this data item has to be treated differently from other data items. It is pointless to pseudonymise Ethnic Category as the range of possible values is small, currently 22 [DN check this number]. This means that it is easy to determine the value of any pseudonym from a small set of records for a given area. Ethnic Category is classified as a sensitive data and it is appropriate to display it (in clear) only where it is relevant to the purpose of the report.

It is appropriate to include Ethnic Category in generic tools built to allow end users to ‘cut and slice’ data in a variety of ways as it is valid for users to need to query the data against that dimension. If Ethnic Category is to be available in this manner and used as a dimension in generating output, then systems should provide a message relating to that dimension giving a warning along the lines that ‘this report might display small numbers / sensitive data, remember your duty of confidentiality’.

****** SUS Spell Identifier** – care must be taken on linking episodes if any episodes involve ‘sensitive’ data covered by the Human Embryology Act and STD Directives, as records should be anonymised in such cases. Previous episodes in the spell may contain identifiable data running counter to the legal requirements. Therefore any Spell involving an episode with sensitive data of this type must be anonymised.

******* Date of Death** – the default should be to not display the actual date of death, however, this may be relevant to some reports and be required to be output.

Patient Label - It follows from Fig 1 that examples of use of patient labels for routine business secondary use are

- Between organisations
    - The LPI for communication between a commissioner and the originating provider
- Within an organisation
    - The LPI for business use outside of the provider that supplied the activity.
    - The pseudonym of the NHS Number
    - The row number within the report if a pseudonym of the NHS Number is not displayed

Other aspects of communication between organisations are covered in the Business Processes and New Safe Haven Reference paper. (Ref 2)

The differences in displaying data for secondary purposes between providers and commissioners mean that there is a need to provide different data displays for users in PCT provider arms than those in PCT commissioner arms.

3.3 Provision and Display of data for External Use

The provision to and use of patient level data external to the secure environment, as defined in Section 3.3 of the PIP Implementation Guidance needs to be considered in the context of a potentially hostile environment. This means that data for external use provided through NHS organisations should be treated differently from in-house NHS business use. The requirements impacting on the contents of any extract supplied are set out in Figure 2 and are based upon the rules governing release of data from the Hospital Episode Statistics service, known as the HES Rules.

Figure 2 Display and handling of sensitive data items for external use

Data item	External (via extract)	
	One-off	Repeated
Name	Do not supply any name data items	
Address	Do not supply any address data items	
Date of birth	Replace by age or age band in years	
Postcode	Postcode sector plus derivations	
NHS Number	Pseudonymised / anonymised	Pseudonymised with consistent values; different values for different purposes for same user
Ethnic category	Do not provide unless relevant to purpose of analysis	
Local patient identifier	Pseudonymised or do not display	
Hospital Spell Number	Pseudonymised or do not display	
SUS PbR Spell ID	Pseudonymised	
Patient pathway identifier	Pseudonymised	
Unique Booking Reference Number	Pseudonymised	
Social Services Client Identifier	Pseudonymised or do not display	
Date of death	Truncate to month and year	

3.4 Spatial analyses

Spatial analysis based on patients' places of residence can only be undertaken in a robust way through the use of full postcodes or geocode / grid references) to derive higher aggregations of geography.

It is implicit that geographic aggregations however derived, can be at least as sensitive as a full postcode or geocode. For this reason care should be taken with any aggregation on a small area basis, such as less than Output Areas as described earlier.

Spatial analysis should be undertaken in the following ways:

- With identifiable data without any data being revealed to the user – e.g. a plot of points for a given set of criteria with no ‘row level’ data available to the user
- If patient characteristics are to be displayed, then data should be displayed as in Figure 1, except that a different set of pseudonyms from those used elsewhere within the local organisation must be used to prevent cross tabulation or inference attacks taking place.
- Provision of an extract of data for use in a local system. In this case, provision of data items should be as in the External columns in Figure 2 and the pseudonyms used must be different from those used in other business use within the organisation. This is again to prevent cross tabulation or inference attacks taking place.

Patient Label - examples of patient labels for spatial analysis are

- The pseudonym of the NHS Number, computed so that is drawn from a different set from that for other business use
- Row number within the report if no pseudonym of the NHS Number is generated or is a pseudonym is not displayed.

4 Techniques

4.1 Techniques used in creating pseudonyms and associated terms

There are two characteristics that distinguish different types of pseudonyms:

- Reversibility – pseudonyms may be reversible or irreversible
- Replicability – the ability to apply a consistent and repeatable pseudonym to the same data item on different occasions and when it appears in different records.

In considering alternative ways of creating pseudonyms, two other factors are important:

- The format of the pseudonym and whether this is appropriate for its intended use - this arises because those approaches which generate pseudonym using cryptographic functions tend to generate long and user un-friendly strings which are suitable for machine processing, but unsuited for use in other contexts
- The methods that should be adopted when using a surrogates field as a pseudonym

The following paragraphs consider alternative ways of creating different types of pseudonyms:

- Irreversible Pseudonyms - the primary mechanism for creating irreversible pseudonyms is through the application of a cryptographic hash function. These are one-way functions that take a string of any length as input and produce a fixed-length hash value (or digest)[8]. Hash functions are available in both SQL Server (from SQL Server 2005 on) and Oracle (obfuscation toolkit). Considerations around the use of hash functions are described at further length below.
- Reversible Pseudonyms - 1 - the creation of a reversible pseudonym generally involves the maintenance of a secure lookup table which holds the source data and which is linked to less sensitive data elements by a surrogate key which is randomised in some way to ensure that there is no relationship between the value of the key and the clear text. The lookup table can either be created to include all potential values, when the range of potential values is bounded (as for example is the case for date of birth) or updated when new values are found in the data. Access to the lookup table must be limited to authorised users. The source text on the secure table may be encrypted to provide a further layer of security if required.
- Reversible Pseudonyms – 2 - in principle, an alternative approach to the creation of a reversible pseudonym is to apply an encryption; the pseudonym being the code text, while access to the clear text is by decryption. However, the encryption functions available in commercial systems such as SQL Server and Oracle[9] may be an inappropriate as a mechanism to create reversible pseudonyms as they are implemented in a way that ensures that a given plain text will generate differing encryptions from case to case.
- Formatting issues can be handled to some degree through data transformation – for example, SUS reduces the length of a pseudonym on presentation by expressing it to Base 36; alternatively other transformations can be used, such as Base 30 or Base 32.

[8] One common use of hash functions is to "destroy" any structure that may exist in the input, while preserving most of its entropy. Validity of using hash functions for entropy extraction is not based on their cryptographic properties but rather on our belief that a good hash function destroys most of the dependencies that may exist in the bits of its input

http://research.microsoft.com/en-us/people/mironov/hash_survey.pdf

[9] Note there are NHS Enterprise Wide Agreements (EWA) concerning Microsoft and Oracle products.

Another approach is to use a further set of randomly generated values to operate against the surrogate key. These techniques will be explored further in Reference Paper 4.

- The mechanisms available for creating suitable surrogate fields and keys include:
    - sequential (random element introduced by sequence in which data item, e.g. NHS Numbers arise)
    - random sampling without replacement
    - cryptographic hash function – such as SQL Server use of MD5 or SHA1 and Oracle's DBMS_CRYPTO function
    - specific system functions for surrogate creation
    - – such as functions within SQL Server (SQL Server identity column) and Oracle (SEQUENCE)
    - adding or subtracting a consistent random number generated from the above.
- Any or all the above may be combined to provide an effective solution in accordance with the rules set out below.

When considering which approach to adopt, the following factors should be taken into account:

- There are a number of technical issues around the use of current hash functions (e.g. see discussions on relevant Wikipedia web pages) in supporting secure systems has identified an number of issues with those currently in use. However, the facilities available within facilities available to NHS organisations through existing system can be regarded as sufficiently robust.
- A more direct concern arises when a hash is used to pseudonymise a bounded set of numbers where the set is relatively small.
    - For example, there are only just over 40,000 days between 1900 and 2010 and if it is known that these have been pseudonymised by the simple application of a hash function then the creation of a look-up table which will break the pseudonymisation is a simple matter.
    - An approach, which has been successfully trialled by one NHS Trust, involves building a table of 75,000 entries, creating a key by sampling without replacement in the range 0 to 9E18, and assigning successive dates from 1880 to each new entry as it occurs. The key was then hashed to provide a pseudonym of consistent form with other pseudonyms applied to the set.
- More generally, hashed values should always be 'seeded' with a local constant value. The effect of this is to localise the pseudonymisation derived from the hash to hashes which use the same seed and algorithm and provide another layer of security.
- It is vital that data is consistently formatted before pseudonymisation to maintain the ability of pseudonyms to link data. This is particularly true of postcodes where data may variously been provided as:
    - A single blank separating the outbound and inbound postcode e.g. CH2_1 XZ
    - Outbound postcode left adjusted and inbound postcode right adjusted e.g. M1___1 XZ
    - With no blank e.g. L12S13K
    - With some other number of intervening blanks

    The simplest approach in this case is to remove any blanks for consistency, as in G23WT for the BBC's postal address of G2 3WT.

4.2 Using a Master Patient Index

In order to reduce the amount of processing and time taken to pseudonymised data, the organisation's master patient index (MPI) can be used. For a PCT, the MPI would be based on demographic records for its responsible and resident population from PDS and for a provider, the MPI would be sourced from its PAS system.

Such MPIs provide the means of creating pseudonyms for the identifiable data items for each person/patient using techniques outlined earlier and using 'look-up tables' to assign the pseudonymised values as records for individual patients are encountered.

The use of the MPI also provides the opportunity to create internal system/organisation pseudonym patient label for each patient. This can be achieved through allocating the patient a random number through sampling without replacement or alternative means of creating a suitably formatted pseudonym or surrogate. For greater security, a two-step pseudonymisation process would be undertaken. The data item can be used as an internal system identifier or it can take the place of the NHS Number as the key pseudonym for a patient.

As records for patients with a NHS Number already present in the MPI are encountered, the previously created pseudonym is used. If the NHS Number is not on the MPI, then a pseudonym needs creating; the NHS Number is added to the MPI and the NHS Number and pseudonym are added to the look up table.

If records are received without an NHS Number (whether missing data or because deliberately anonymised), the patient can be assigned a temporary pseudonym. Subsequently 'fuzzy matching' can be used to try to link the records for patients with temporary pseudonyms to those with known NHS Numbers in the cases of missing data. This method ensures that, for PCTs in particular, complete data sets can be used encompassing all activity undertaken on their behalf.

4.3 Techniques for accessing identifiable data:

Access to identifiable data, can either be in response to a specific user initiated request for a pseudonym to be reversed or set automatically, for example by reference to a user's role -

- De-pseudonymisation in response to a specific request should be the normal mode of operation where the user needs to know the identity of only a small subset of data – for example where a subset of patients at risk of readmission and in need of intervention is extracted from a much larger dataset. As well as minimising the risk of unnecessarily allowing patient labels to be viewed, this approach supports a much more targeted approach to audit.
- Automatic access to clear data may be relevant in respect to staff that are using "secondary data" to support quasi-operational processes and routinely need to identify all the patients returned from the systems, such as legitimate users of output from use of the PARR algorithm.

Specific examples of the two approaches are set out below:

- Different and separate access routes to pseudonymisation and identifiable data so pseudonymisation users do not see identifiers or vice versa
- Unlock for non-display of identifiers and no pseudonyms – all users all access the same screens which show pseudonyms (or suppressed identifiable fields); authorised users request access to identifiable data, and then are required to re-authenticate to unlock the view of non-identified data to reveal NHS Numbers and Dates of Birth and other identifiable data. This could be achieved through several look ups on a record identifier,

but a map of pseudonym to identifiable value is usually required to provide acceptable response times to the end user.

4.4 RULES

Rules for applying pseudonymisation techniques are:

1. Whichever method of pseudonymisation is used, each field has a different base for its pseudonym – e.g. with encryption, say key 1 for NHS Number, key 2 for date of birth; so that it must not be possible to deduce values of one field from another if the pseudonym is compromised.
2. Pseudonyms to be used in place of NHS Numbers and other fields that are to be used by NHS staff must be of reasonable length and formatted on output to ensure readability. Consideration also needs to be given to the impact on existing systems both in terms of the maintenance of internal values and the formatting of reports. For example, in order to replace NHS Numbers in existing report formats, then the output pseudonym should generally be of the same field length, (i.e. 10 characters but should not be only digits to avoid confusion with genuine NHS Numbers).
3. Pseudonyms generated from a hash must be seeded.
4. Pseudonymisation should be undertaken prior to user access and not ‘on the fly’, that is undertaking the generation of pseudonyms whilst processing and displaying data. This is because of the chances of error leading to inadvertent display of identifiable data. This method may be acceptable in producing extract files, but only if checks are made the output prior to dispatch to the user.
5. Pseudonyms for external use must be generated to give different pseudonym values (e.g. via use of a different hash seed) in order that internal pseudonyms are not compromised.
6. In the absence of explicit approvals to the contrary, data provided to external organisations should apply a distinct pseudonymisation to each data set generated for a specific purpose so that data cannot be linked across them. A consistent pseudonym may be required for a specific purpose, separate data sets are provided over a period of time and the recipient needs to link the data sets to create longitudinal records.
7. Display only the pseudonymised data items that are required, e.g. do not display pseudonymised date of birth if it is not relevant to a report
8. De-pseudonymisation requests for, and access to data in the clear, must be fully logged and approval by the appropriate authority documented.
9. Pseudonymised data must be treated in the same way as identifiable data in terms of security and access, as risks of re-identification do exist.
10. Pseudonymisation does not obviate the need to maintain the highest standards of security and confidentiality in local working and in system design and implementation.

4.5 Extracts of identifiable data

If extracts of identifiable data are generated, the user must be reminded of the obligations on them. The obligations include the subsequent management of extracted data with an indication of what responsibilities a registered user must comply with.

4.6 Log and audit access to identifiable data

It is necessary to log access to identifiable data. This is in order to provide the basic information to support the Care Record Guarantee to inform patients as to who has accessed/seen their data and to support forensic analysis in the event of untoward incidents.

The logging process should happen automatically in transaction processing systems and would be expected for provider based clinical systems. However, for database systems primarily operated for non-direct care purposes, such as corporate warehouses, this is not feasible as the records for legitimate accesses could be as numerous as the database itself and would also contain identifiable data, exacerbating the risk of inappropriate disclosure of patient identifiable data.

The aim of logging becomes to create the ability to know who has accessed identifiable data and to be able to replicate the query undertaken.

The key items to be logged therefore are:

- Who has accessed which data bases containing identifiable data
- Date and time of access
- Query or access process undertaken, including the parameters of the query.

The log of accesses should itself form a structured database to enable queries and audit.

The log of accesses should be audited via sampling of users or subject matter on a regular basis. The aim of the audit is to check for unusual patterns of access.

5 Implementing De-identification in systems and organisations

5.1 Principles

There are many ways in which the facilities and techniques outlined in earlier sections can be implemented in systems used for non-direct care purposes. The way in which individual organisations undertake implementation will depend on the particular configuration of processes, technologies and systems utilised. However, there are some principles that need be respected in providing data for secondary use purposes. These are:

- The only circumstances in which pseudonyms need not be generated and used is if data is not to be extracted from a system within the organisation or supplied to external organisations; in which case data must be displayed without showing any identifiable data.
- Identifiable data items (such as NHS Number, Date of Birth, postcodes and LPI) should be stored separately (e.g. in separate tables) from the pseudonymised and non-identifiable data (the payload). Where possible, storage should also be physically distinct. (E.g. on a separate server).
- Pseudonyms for external usage must be different from internal pseudonyms in order to minimise the risk of compromising pseudonyms.

5.2 Shared services

Shared service arrangements for providing informatics and commissioning support services to multiple organisations are used extensively across the NHS. The common feature of these services is that data processing functionality and resources are shared by the services' customer organisations. The requirement to implement de-identification can be met by:

- Using pseudonyms specific to each organisation through use of different keys or seeds for each constituent organisation; or
- Using pseudonyms specific to the shared service, that is through use of the same key or seed for all constituent organisations; or
- Using both approaches for different uses of the data; for example a common set of pseudonyms might be used to support invoice validation and contract management across all organisations, but organisation specific pseudonyms created to support data analysis undertaken by public health departments.

The decision on which approach to take is a local one and should be determined by the purpose of the shared services and how it operates.

5.3 Provider stand-alone and legacy systems

Context

There are two sets of circumstances where de-identification of data for secondary uses in Provider organisations may not be immediately feasible. These are where the addition of pseudonymisation facilities or the modification of reporting functionality to existing systems using patient identifiable data would be complex and disproportionately uneconomic.

Such systems are

- stand-alone systems, usually supporting specific clinical areas
- Legacy systems, which may include Patient Administration Systems (PAS).

Potential solutions

The main potential solution is through the application of de-identification via export of reports to another system, effectively via middleware. In the case of bringing data together from multiple sources, for example for reporting on patient pathways such as Referral to Treatment, this may be best achieved with a database solution.

PIP Requirements

Meeting the need to de-identify patient level data for secondary uses in the context set out above may be complex and time consuming. It is still incumbent upon the organisation to determine how this is achieved. If a solution is not implemented, the organisation will in effect be operating outside the legal framework for using identifiable data.

In these circumstances, it is necessary for the organisation to mitigate the risk posed by such systems through

- Compiling a register of such systems
- Implementing appropriate restrictions on access to identifiable data made available through these systems through physical and electronic means
- Training of relevant staff on secondary use IG and good practice
- Developing exit strategies to resolve the IG issues for each system.

5.4 Technical guidance

Further technical guidance on implementing pseudonymisation will be provided through the 'White Paper'. (Ref 4). This will cover amongst other subjects:

- Use of hash function
- Sample code
- Creation of look up tables
- Use of schema/user separation
- Use of joins to create identifiable user views.
- Data transformations.

Annex 1

Figure 3 SUS data items for pseudonymisation

(Key M for mother when record is for a baby and B for baby when record is for mother)

Item No.	Logical Name	Attribute
1	NHS NUMBER	NHS_NO
2	NHS NUMBER (MOTHER)	NHS_NO_M
3	NHS NUMBER (BABY)	NHS_NO_B
4	NHS NUMBER (OLD)	OLD_NWCS_NHS_NO_OLD
5	PERSON BIRTH DATE	DOB
6	PERSON BIRTH DATE (MOTHER)	DOB_M
7	PERSON BIRTH DATE (BABY)	DOB_B
8	POSTCODE OF USUAL ADDRESS	POSTCODE
9	POSTCODE OF USUAL ADDRESS	POSTCODE_2
10	POSTCODE OF USUAL ADDRESS (MOTHER)	POSTCODE_M
11	PATIENT PATHWAY IDENTIFIER	PATIENT_PATHWAY_ID
12	LOCAL PATIENT IDENTIFIER	LPI, LPI M, LPI B
13	MHMDS LOCAL PATIENT IDENTIFIER	MHMDS LOCAL PATIENT IDENTIFIER
14	Patient UBRN	UBRN
15	UNIQUE BOOKING REFERENCE NUMBER (CONVERTED)	UBRN_CONVERTED
16	Address Line 1	ADDRESS_LINE1
17	Address Line 2	ADDRESS_LINE2
18	Address Line 3	ADDRESS_LINE3
19	Address Line 4	ADDRESS_LINE4
20	Address Line 5	ADDRESS_LINE5
21	Address Line 1 (Mother)	ADDRESS_LINE1_M
22	Address Line 2 (Mother)	ADDRESS_LINE2_M
23	Address Line 3 (Mother)	ADDRESS_LINE3_M
24	Address Line 4 (Mother)	ADDRESS_LINE4_M
25	Address Line 5 (Mother)	ADDRESS_LINE5_M
26	Hospital Provider Spell Number	HOSPITAL_PROVIDER_SPELL_NO
27	Patient Name 1	PATIENT_NAME1
28	Patient Name 2	PATIENT_NAME2
29	Patient Name 3	PATIENT_NAME3
30	Patient Name 4	PATIENT_NAME4
31	Patient Name 5	PATIENT_NAME5
32	Patient Preferred Name	PATIENT_PREFERRED_NAME
33	Patient Surname	PATIENT_SURNAME
34	Person Identifier	PERSON_IDENTIFIER
35	SS Client Identifier	SS_CLIENT_ID
36	Usual Address	USUAL_ADDRESS
37	Usual Address (Mother)	USUAL_ADDRESS_M