

# NHS Number and the systems used to manage them

## An overview for research users

---

Andy Boyd, Richard Thomas, John Macleod  
Population Health Sciences, Bristol Medical School,  
University of Bristol

April 2018



## **Acknowledgements & Background**

This project was funded as an Innovation award as part of the collaborative research programme entitled 'Cohorts and Longitudinal Studies Enhancement Resources' (CLOSER); ESRC grant reference: ES/K000357/1. CLOSER is a consortium including eight of the UK's major cohort and longitudinal studies. The CLOSER network brings these teams together to:

- stimulate interdisciplinary research across the major longitudinal studies
- provide shared resources for research
- assist with training and development for researchers in the use of longitudinal data at all career stages
- share information and expertise in longitudinal methodology.

This project – CLOSER work package 8 – aims to highlight means of linking longitudinal research studies to routine health records. The project draws on methodologies developed in the Project to Enhance ALSPAC through Record Linkage (PEARL) – a Wellcome Trust award (WT grant reference: WT086118/Z/08/Z) – that has developed linkages within the Avon Longitudinal Study of Parents and Children (ALSPAC).

Much of the information presented within this report has been distilled from NHS Digital (and predecessor organizations) guidance documents. This includes our adaptations of some technical diagrams and graphics. The source documents have been acknowledged where appropriate.

### **Citation of the Report:**

Boyd A, Thomas R, Cornish R, Macleod J. (2018). NHS Number and the systems used to manage them: an overview for research users. Bristol, UK: University of Bristol.

## Contents

Summary .....	4
<i>Scope</i> .....	5
NHS Numbers and the systems used to manage them .....	6
<i>History</i> .....	6
<i>Value within a Longitudinal Population Studies context</i> .....	7
NHS Number .....	9
Scottish Community Health Index (CHI) .....	13
Northern Ireland Health and Care (H&C) Number .....	14
Systems for managing the NHS Number (in England, Wales and IoM) .....	14
<i>Historical Systems</i> .....	14
<i>The NHS ‘Spine’</i> .....	15
<i>Patient Demographic Service (PDS) and the National Back Office (NBO)</i> .....	15
Governance issues specifically related to NHS Numbers.....	17
Quality and coverage issues of relevance to LPS community .....	19
The use of NHS Number and PDS demographic data for record linkage purposes.....	21
Appendix A: MRIS Record Linkage Protocol .....	25

## Summary

**NHS Numbers are the only unique, national, patient identifier within the UK's health and social care system. They have both clinical and research value, and are of particular value to the longitudinal population study community in establishing linkages to participants' records. Historically, creation and allocation of NHS Numbers was not coordinated across providers or systems; introducing duplication and error within NHS record systems and meaning record linkage algorithms had sub-optimal information to accurately establish linkages. The creation of the modern NHS Number and the NHS Spine – a system for managing patient identifiers through the care system – has led to improvements in NHS Number management and quality. Despite these improvements, from a longitudinal perspective the system remains complex, varies across the UK, and is subject to increasingly stringent data owner controls and legislative requirements.**

The NHS Number is a unique identifier used to consistently identify patients across the health and social care system. There are three separate NHS Number systems: one for England, Wales and the Isle of Man; one for Scotland; and, one for Northern Ireland. NHS Numbers (in England, Wales and the IoM) are currently managed within the 'Personal Demographics Service'; a database of NHS administrative patient information. A strong patient identifier is vital for clinical safety (e.g. ensuring the correct information is joined together) and for administration (e.g. record management and the distribution of funds within the NHS).

NHS Numbers and the systems that manage these are also of value to the research community, given that linkage to routine records is increasingly seen as an important data collection methodology within longitudinal research<sup>1,2</sup>. This resource report summarises the NHS Numbers used across the UK, outlines how these are managed with a focus on the Personal Demographic Service (PDS) and how these resources relate to the needs of longitudinal population studies.

This technical aspect of the report is supplemented with a short discussion on selected literature related to the quality of NHS Number, the NHS patient registers and record linkage. The aim of this is to provide an overview (rather than a comprehensive review) of linkage to NHS data and the technical issues associated with the linkage and biases that may emerge.

---

<sup>1</sup> Pell, J., 2014. Maximising the Value of UK Population Cohorts: MRC Strategic Review of the Largest UK Population Cohort Studies.

<sup>2</sup> Wellcome's Longitudinal Population Studies Working Group, 2017. Longitudinal Population Studies Strategy.

## **Aims**

This report aims to summarise information related to the NHS Number, the systems used to manage NHS Number's and to note design or historical features which are of relevance to NHS Number's application within the cohort and longitudinal study community. Within this, we will:

1. Provide an overview of NHS Number, describing its history, coverage and content;
2. Describe the particular format of the different versions of NHS Number;
3. Summarise the process by which NHS Numbers are allocated;
4. Describe the systems developed to manage NHS Number;
5. Describes types of quality and coverage issue impacting on NHS Number and associated patient demographic data;
6. Summarise some of the literature assessing the quality and completeness of the NHS Number and associated demographic/registry information.

The report is not based on a systematic review of all literature relating to NHS Number, the systems used to manage it, or the sum of research studies utilising NHS Number in their research. Rather, it is based on the knowledge and experience of the authors gained from linking participants in the ALSPAC birth cohort study to their NHS Numbers held in the national repository for these data.

## **Scope**

The report focuses on the NHS Number systems in England, Wales and the Isle of Man. This system is distinct from the Scottish NHS Number system and to the system used in Northern Ireland.

## NHS Numbers and the systems used to manage them

### *History*

The origins of the NHS Number system lie in the Second World War. Where, in anticipation of the war, a nascent national health service was established through the Emergency Medical Services (1938). Shortly after war was declared, a population census was conducted (taken on 29th September 1939), a national population register was established and national identity cards were issued (as part of the census enumeration process)<sup>3</sup>. Each identity card recorded a person's identity and allocated an identity number. The local population registers taken during the 1939 registration census were compiled into a central register held at the Central National Register Office (CNRO) at Southport (where the 1939 archive comprises of approximately seven thousand transcript books containing forty million registrations)<sup>4</sup>. The National Register was adapted to meet the needs of the National Health Service (1948) and continued to run under 'wartime conditions' (where it was used in the management of rationing) until it was discontinued in February 1952. Following this, the CNRO became the National Health Service Central Register (NHSCR), where the National Register became a patient register, with the associated Identity Numbers becoming the original NHS Numbers. The primary purposes of the NHSCR was to: 1) help General Practitioners (via Family Practitioner Committees (formerly Executive Councils) of the NHS) to register patients through maintaining a central register of all NHS patients in England and Wales; 2) control the issue of new NHS numbers and manage the risk of inflated of practice lists; and, 3) feed back vital event notifications to General Practices (e.g. notifying GPs of patient deaths)<sup>5</sup>. Until 1990 the NHSCR was managed as a paper-based system, based on each NHS patient having an entry in a transcript book (indexed by NHS Number). This entry was updated to reflect patients' current status (including cancer and death registrations and practice change/migration).

The initial coverage of NHS Numbers and the National Register was high (given it was tied to the allocation of rations)<sup>4</sup>. However, historically, the use of the NHS Number was inconsistent, both in terms of quality and quantity, across NHS organisations and different NHS record keeping systems<sup>6</sup>. Improving Information Technology capability coupled with a desire to improve clinical care<sup>7</sup> and operating efficiencies within the NHS has driven initiatives to develop a fully interoperable electronic health care system; underpinned by the mandatory and standardised use of NHS Number<sup>8</sup>. Historical milestones in relation to NHS

---

<sup>3</sup> The 1939 National Registration census is of great interest to those interested in genealogy and the academic community. The census records have been digitised and are available via genealogical databases or government offices. See: <http://www.1911census.org.uk/1939.htm>

<sup>4</sup> Agar JE. Identity cards in Britain: past experience and policy implications. *History & Policy*. 2005(33).

<sup>5</sup> The history of the NHS Central Register is recorded within the 'General Register Office: National Health Service Central Register: Correspondence and Papers' archive held in the National Archive (<http://discovery.nationalarchives.gov.uk/details/r/C13347>).

<sup>6</sup> Information Standards: NHS Number Standard for General Practice (England) v1.2. March 2009. Great Britain: Information Standards Board for Health and Social Care.

<sup>7</sup> Risk to patient safety of not using the NHS Number as the national identifier for all patients. (2009). Great Britain: NHS Connecting for Care. Available from: <http://webarchive.nationalarchives.gov.uk/20130502105747/http://www.nrls.npsa.nhs.uk/resources/?entryid45=61913>

<sup>8</sup> Five Year Forward View. (2014). Great Britain: NHS England.

Number<sup>9</sup> include: 1) in 1991 NHSCR introducing an in-house developed computer system called the Central Health Register Enquiry System (CHRIS)<sup>10</sup> to replace the previous paper-based system; 2) In 1995 the introduction of the new NHS Number system; 3) From 2009, NHS standards mandated the precise structure of this new NHS Number, its systematic use and that it be incorporated into all new software being deployed within the NHS system; and, also in 2009, changes implemented to allocate NHS Number's as soon after birth as possible via the Information Authority's NHS Numbers for Babies (NN4B) Programme.

Over time the NHSCR has become an increasingly important research resource. Studies were able to 'trace' (i.e. establish a record linkage) their participants within the register and 'flag' them to receive vital status updates. The Medical Research Information Service (MRIS) was established to conduct these tracing and flagging services in the role of 'honest broker'<sup>11</sup>. By 1998 MRIS had flagged ~500 research studies in the NHSCR<sup>12</sup>. This role (within England, Wales and IoM) has transferred to NHS Digital, within Scotland this role is conducted by the National Records of Scotland (NRS), and within Northern Ireland by the Central Services Agency for the Health and Social Services<sup>13</sup>.

### ***Value within a Longitudinal Population Studies context***

Outside of their clinical and administrative value within the NHS, longitudinal studies are likely to use NHS Number to inform record linkage and data extraction processes. Within this use case, the systems used to manage NHS Number, the additional demographic and administrative data they contain and the manner in which these interface with other health record systems are likely to be as important as the NHS Number themselves<sup>14</sup>. Record linkage in this context refers to the process of establishing a link between a study participants record within a study administrative record and that same given individual's record(s) within a health or social care dataset (or across multiple datasets). Record linkage techniques typically rely on comparing individuals' personal identifiers (e.g. their name, address, sex, date of birth). Each of which are subject to natural change over time or reporting/recording error. Augmenting 'real world' identifiers such as these with identification numbers helps address change occurring from social processes, some forms of error and also helps the efficiency of data management<sup>15</sup>.

---

<sup>9</sup> For a wider perspective on the creation and evolution of the UK National Health Service see: Greengross P, Grant K, Collini E. The history and development of the UK National Health Service 1948 1999. Health System Resource Centre, United Kingdom. 1999.

<sup>10</sup> NHS Information Centre takes control of NHS Central Register (Press Release). April 2008. Great Britain: NHS Information Centre. Available from: <http://webarchive.nationalarchives.gov.uk/20090410030330/http://www.ic.nhs.uk/news-and-events/press-office/press-releases/april-2008/nhs-information-centre-takes-control-of-nhs-central-register>

<sup>11</sup> Strategy 2010/13 and Annual Business Plan 2010/11 (March 2010). Great Britain: The NHS Information Centre for health and social care. Available from: [http://content.digital.nhs.uk/media/1386/Business-Plan-201011-and-Strategy-201013/pdf/NHS\\_IC\\_draft\\_Bus\\_strat\\_plan\\_2010-11\\_v7a.pdf](http://content.digital.nhs.uk/media/1386/Business-Plan-201011-and-Strategy-201013/pdf/NHS_IC_draft_Bus_strat_plan_2010-11_v7a.pdf)

<sup>12</sup> Ibid. 10.

<sup>13</sup> NHSCR data: quality assurance of administrative data used in population statistics. (December 2016). Great Britain: Office for National Statistics. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/nhscrdataqualityassuranceofadministratedatausedinpopulationstatisticsdec2016>

<sup>14</sup> Womersley J. The public health uses of the Scottish Community Health Index (CHI). *Journal of Public Health*. 1996 Dec 1;18(4):465-72.

<sup>15</sup> Gill LE. OX-LINK: the Oxford medical record linkage system. In Proc. Int'l Record Linkage Workshop and Exposition 1997 (pp. 15-33).

Once a link is established between study participants and their NHS Number, this can be used to 'flag' participants within centrally managed patient registers. Such flagging can be a means to undertaking retrospective data collection on existing cohorts and prospective long-term data collection designed to capture critical outcomes, improve study administration and enable the linkage of additional health and social care records.

The demographic and health service status information contained within the patient index databases used to manage NHS Numbers has additional value for longitudinal studies. These data can be used to:

- Inform participant tracing strategies – for example, the ALSPAC birth cohort study use linkage to PDS to gather contact information on participants who have become lost to follow-up<sup>16</sup>, the 6 Day Sample of the Scottish Mental Survey used Scottish NHSCR tracing to mail fair processing materials<sup>17</sup>;
- To collect key outcome information (e.g. fact of death) to avoid insensitive patient contact<sup>18</sup> and as a means of collecting important outcome data. For example, the Hertfordshire Aging Study used linked cause of death to identify that small size at birth and during infancy was associated with increased risk of cardiovascular mortality in men and women<sup>19</sup>;
- To assist study recruitment and sample selection. For example, the Born in Bradford cohort used live feeds to the PDS's predecessor system to validate NHS Numbers during recruitment<sup>20</sup>;
- 'Flag' a cohort of participants to the NHS to receive ongoing vital event notifications and to permit linkage to other data sources (e.g. Hospital Episode Statistics databases. For example, the ALSPAC birth cohort study, an initial linkage to NHS Number via the NHSCR system was used to inform linkage to participants' records within the Central Practice Research Datalink database<sup>21</sup>, to identify and extract records from the Hospital Episode Statistics database<sup>22</sup>, and to determine participants' registered GP practice and therefore enable extraction of participant primary care records held within General Practice systems<sup>23</sup>.

---

<sup>16</sup> Bray I, Noble S, Boyd A, Brown L, Hayes P, Malcolm J, Robinson R, Williams R, Burston K, Macleod J, Molloy L. A randomised controlled trial comparing opt-in and opt-out home visits for tracing lost participants in a prospective birth cohort study. *BMC medical research methodology*. 2015 Dec;15(1):52.

<sup>17</sup> C.E. Brett, I.J. Deary. Realising health data linkage from a researcher's perspective: Following up the 6-Day Sample of the Scottish Mental Survey 1947. *Longitudinal and Life Course Studies*, 5(2014), pp. 283-298

<sup>18</sup> Singh S. How not to contact the dead. *Br J Gen Pract*. 2013 Jun 1;63(611):309-.

<sup>19</sup> Syddall HE, Simmonds SJ, Martin HJ, Watson C, Dennison EM, Cooper C, Sayer AA, Hertfordshire Cohort Study Group. Cohort profile: The Hertfordshire Ageing Study (HAS). *International journal of epidemiology*. 2009 Jan 8;39(1):36-43.

<sup>20</sup> Raynor P. Born in Bradford, a cohort study of babies born in Bradford, and their parents: protocol for the recruitment phase. *BMC public health*. 2008 Dec;8(1):327.

<sup>21</sup> Cornish RP, Henderson J, Boyd AW, Granell R, Van Staa T, Macleod J. Validating childhood asthma in an epidemiological study using linked electronic patient records. *BMJ open*. 2014 Apr 1;4(4):e005345.

<sup>22</sup> Mars B, Cornish R, Heron J, Boyd A, Crane C, Hawton K, Lewis G, Tilling K, Macleod J, Gunnell D. Using data linkage to investigate inconsistent reporting of self-harm and questionnaire non-response. *Archives of suicide research*. 2016 Apr 2;20(2):113-41.

<sup>23</sup> Cornish RP, John A, Boyd A, Tilling K, Macleod J. Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. *BMJ open*. 2016 Dec 1;6(12):e013167.



## NHS Number

The NHS Number in England, Wales and IoM has existed using two distinct formats.

### *Original NHS number format*

The original NHS number format (based on WWII National Identity Card Numbers) used an alpha-numeric code; an initial four-character code followed by numbers.

Those alive in 1939 and included in the National Registration census had systematically applied numbers which comprised a four alpha-character code indicating the enumeration district and a subsequent set of digits where the first digits indicated the 'schedule' being completed by the numerator (which indicate an individual property or institution) and subsequent digits assigned incrementally based on how many people were in the property at the time of the census (e.g. for area ABCD, schedule line 24, 3rd person in the household the card number would have been "ABCD 24/3")<sup>24</sup>.

Subsequent entries to the register (e.g. new births) would receive a number using the same format, where the alpha-component corresponded to the birth registration district and the time period. This system continued until 1996.

### *Current NHS Number format*

The new format NHS number, introduced in 1996, is a unique 10-digit number assigned to every individual registered with the NHS in England, Wales and IoM<sup>25</sup>. The NHS Number is constructed to a standard; which specifies how the NHS Number is created, used, transmitted and displayed<sup>26</sup>. It is 10 numeric digits in length, and is displayed and printed using a '3 3 4' format (e.g. 123 456 7890). The first nine digits are the identifier and the tenth is a check digit used to confirm the number's validity<sup>27</sup>. As a means to improve quality and consistency of NHS Number use across the NHS it is now mandatory for NHS software systems to validate inputted numbers and to reference the NHS Number master reference file<sup>28</sup>. This validation check uses the Modulus-11 check process that can form a useful quality control measure within LPS data management systems. The validation check can be expressed in any programming language, but is illustrated here as Stata syntax<sup>29</sup> (see Panel 1).

The new format NHS Number was introduced with a communication campaign aiming to improve NHS Staff and public understanding and support for the systematic use of the NHS Number on all NHS records (see Figure 1).

---

<sup>24</sup> See the website for the 1939 National Registration census for a more detailed description: <http://www.nationalarchives.gov.uk/help-with-your-research/research-guides/1939-register/>

<sup>25</sup> Connecting for Health - Explaining the NHS Number. [Online].; 2010. Available from: <http://webarchive.nationalarchives.gov.uk/20100305114152/http://www.connectingforhealth.nhs.uk/systemsandservices/nsts/faqs/explaining>.

<sup>26</sup> NHS Number Standard Specification v1.0. 31<sup>st</sup> May 2012. Great Britain: Information Standards Board for Health and Social Care. Available from: [http://webarchive.nationalarchives.gov.uk/20150107144958/http://www.isb.nhs.uk/documents/isb-0149/amd-136-2010/index\\_html](http://webarchive.nationalarchives.gov.uk/20150107144958/http://www.isb.nhs.uk/documents/isb-0149/amd-136-2010/index_html)

<sup>27</sup> Ibid.

<sup>28</sup> Ibid.

<sup>29</sup> Stata v15. StataCorp. 2011. Stata Statistical Software: Release 15. College Station, TX: StataCorp LP.

Figure 1. NHS Number public awareness poster (2013)<sup>30</sup>

**NHS**

# You are unique - and so is your NHS number

**Find it.**

The NHS number makes it easier to find your health records and coordinate the care you receive. Using your NHS number to identify you correctly is an important step towards improving the quality of your treatment and care.

**Your unique 10-digit NHS number will look similar to this example. 943 476 5870**

**Use it.**

You can find your NHS number by checking:

- . your medical card (if you have one);
- . your prescription;
- . a letter or other document from the NHS; or
- . with your GP practice.

Your baby's NHS number can be found in their red book (if they have one).

**Share it.**

If you know your NHS number, you can help healthcare staff find your records more easily by showing your number to them at appointments or quoting it when you contact them.

**hscic** Health & Social Care Information Centre

If you have any questions you can email [nhsnumber@hscic.gov.uk](mailto:nhsnumber@hscic.gov.uk)

Ref: 4787

<sup>30</sup> Public and NHS staff information materials can be found at: <http://webarchive.nationalarchives.gov.uk/20160921150902/http://systems.digital.nhs.uk/nhsnumber/staff/comms>

**Panel 1. Stata representation of the NHS Number Modulus 11 check-sum.**

```

* STEP 1: Multiply each of the first nine digits by weighting factor
* part a: convert nhs_id to string and split each digit into it's own variable
tostring nhs_id, gen(nhs_id_st)
gen t1 = substr(nhs_id_st,1,1)
gen t2 = substr(nhs_id_st,2,1)
gen t3 = substr(nhs_id_st,3,1)
gen t4 = substr(nhs_id_st,4,1)
gen t5 = substr(nhs_id_st,5,1)
gen t6 = substr(nhs_id_st,6,1)
gen t7 = substr(nhs_id_st,7,1)
gen t8 = substr(nhs_id_st,8,1)
gen t9 = substr(nhs_id_st,9,1)
gen t10 = substr(nhs_id_st,10,1)

* part b: convert single digit variables to numerics
foreach v of var t1-t10 {
destring `v', replace
}

* part c: multple first 9 digits by weighting factor
replace t1 = t1*10
replace t2 = t2*9
replace t3 = t3*8
replace t4 = t4*7
replace t5 = t5*6
replace t6 = t6*5
replace t7 = t7*4
replace t8 = t8*3
replace t9 = t9*2

*****
* STEP 2: Add the results of each multiplication together
gen total = t1+t2+t3+t4+t5+t6+t7+t8+t9

*****
* STEP 3: Divide the total by 11 and establish the remainder
gen remainder = mod(total, 11)

*****
* STEP 4: Subtract the remainder from 11 to give the check digit
gen check_digit = 11-remainder

*****
* STEP 5: Check the remainder matches the check digit. If it does not, the NHS NUMBER is invalid
* If the result is 11 then a check digit of 0 is used.
* If the result is 10 then the NHS NUMBER is invalid
gen id_valid = 1 if check_digit ==t10
replace id_valid = 0 if check_digit !=t10
replace id_valid = 1 if check_digit ==11 & t10 ==0
replace id_valid = 0 if check_digit ==10
replace id_valid =. if nhs_id ==.

*****
* Finalising file & Labelling
label define valid_lb 0"no" 1"yes"
label values id_valid valid_lb
* drop temp vars & save outfile

```

### ***Original system for allocating NHS numbers***

The initial allocation of NHS numbers and the population of the National Register was conducted during the National Registration census by enumerators systematically visiting households. Subsequent ID allocation for babies was managed by the 'Registrars of Birth' at the point of Civil Registration<sup>31</sup>. This system continued until 2002.

### ***Current system for allocating NHS numbers***

Some individual will not have an NHS Number (those born outside of England, Wales or the IoM, or, those born in England, Wales or the IoM before October 2002 and who had not previously registered with the NHS). Since October 2002, newborn babies are allocated an NHS number soon after birth; with ID allocation being managed by NHS staff in Maternity Units as they complete the Statutory Birth Notification<sup>32</sup>. For other patients, they will be allocated an NHS number as they join the NHS, although temporarily joining an NHS GP practice list or visiting for less than three months does not result in the issue of an NHS number. This process of allocating NHS Numbers is described in detail elsewhere<sup>33</sup> and is summarised in Panel 2.

The master copy of an individual's NHS number is held within the PDS database (described below). Note that individual's PDS record may be invalidated and a new PDS record created. This will result in the allocation of a new NHS Number.

The NHS have provided guidance illustrating 'complex cases' situations where NHS Numbers may not have been allocated (e.g. overseas visitors, patient only ever registered in Scotland or Northern Ireland, private patients, long-term mental health patients)<sup>34</sup>. These are distinct from scenarios where NHS Number exists but cannot be traced due to data quality issues. Where NHS Numbers cannot be identified, temporary ID codes can be used instead. There are a range of codes describing particular 'complex case' scenarios (e.g. ZZ99 3VZ is used for patients with no NHS Number and are currently of no fixed abode).

## **Panel 2. NHS Number allocation process**

After registering at a General Practice, the patient acceptance details are returned to the NHSCR (Central Register). The return is checked to see if patient has already registered with another NHS GP or has NHS number, if so:

- NHSCR advises the currently responsible organisation to deduct the patient from their database.
- The currently responsible organisation advises the patient's previous GP to deduct the patient and forward their medical records to their organisation.

---

<sup>31</sup> It is a legal requirement in the U.K. to officially register every live-born and still-born (until September 1992, defined as deliveries 28+ weeks gestation, from October 1992 this threshold was reduced to 24+ weeks) birth within 40 days from delivery.

<sup>32</sup> Data Standards: NHS Numbers for Babies. 29th October 2002. Great Britain: NHS Information Authority.

<sup>33</sup> NHS Number Standard Specification. (2012). Great Britain: NHS Information Standards Board. Available from: <http://webarchive.nationalarchives.gov.uk/+http://www.isb.nhs.uk/library/standard/191>

<sup>34</sup> NHS Number Your Unique Patient Identifier: Guidance on categories of complex case. July 2012. Great Britain: Connecting for Care. Available from: <http://webarchive.nationalarchives.gov.uk/20160921163126/http://systems.digital.nhs.uk/nhsnumber/saff/guidance/complex.pdf>

- The previous GP's organisation forwards the patient's medical records to the new responsible organisation.

If not:

- NHSCR issues an NHS number to the patient.
- NHSCR creates a registration record on its own system.

In either case:

NHSCR confirms acceptance and patient details, including NHS number, with the newly responsible organisation.

- The newly-responsible organisation confirms acceptance and patient details, including NHS number, with the patient's new NHS GP.
- The organisation may send the patient a medical card (FP4), including their NHS Number.

## **Scottish Community Health Index (CHI)**

NHS Scotland maintain a register of all patients within The Community Health Index (CHI). The CHI Number is the unique patient identifier within this index. The CHI database has evolved from a regional database<sup>35</sup> to having national coverage and it now contains a record for all NHS Scotland patients who are resident in Scotland, are non-Scottish patients or temporary visitors. The use of CHI Number in NHS Scotland systems is mandatory<sup>36</sup> and the majority of babies are allocated with their CHI Number at birth (babies born at home or in a remote community may have a lag before CHI Number allocation given there is no immediate CHI-allocation or registration facility available in such circumstances<sup>37</sup>). The CHI system comprises regional database, linked by a search index, and interfaces with the (Scottish) National Health Service Central Registrar (NHSCR)<sup>38</sup> which contains birth registration data extending back to the 1939 National Registration census.

The CHI Number is formatted as a 10 digit numeric code. Like the NHS Number, the final digit is a check digit that be used to validate the code using the Modulus 11 process. The code contains embedded personal identifiers, the first six digits correspond to Date of Birth (in DDMMYY format) and the ninth digit represents patient sex (the 9<sup>th</sup> digit will be an odd number for males and an even number for females).

Embedding personal identifiers within the CHI Number has potential benefits in ensuring quality control (validating incoming linked health records against a study participant database using embedded date of birth is a useful quality check)<sup>39</sup> yet also complicates the ethico-governance use of the identifier (given that includes additional 'real world' identifiers that could be used to disclose a patients identity).

---

<sup>35</sup> Womersley J. The public health uses of the Scottish Community Health Index (CHI). *Journal of Public Health*. 1996 Dec 1;18(4):465-72.

<sup>36</sup> The use of the CHI (Community Health Index) to support integrated care across the NHS in Scotland. (2013). Great Britain: NHS Scotland.

<sup>37</sup> Ibid.

<sup>38</sup> <http://www.shsc.scot/media/16292/2014-12-16-what-is-the-community-health-index.pdf>

<sup>39</sup> Walesby K, Harrison JK, Russ TC. What big data could achieve in Scotland. *J r Coll Physicians edinb*. 2017 Jun;47:114-9.

## Northern Ireland Health and Care (H&C) Number

Within Northern Ireland, practice registration data and patient demographic information is maintained within the Medical Card Register; a list of patients registered with a family doctor<sup>40</sup>. The unique patient identifier is known as the 'Health and Care Number (H&C Number)'. The H&C number uses the same 10-digit numerical value as the NHS Number used in England, Wales and IoM. It is presented using the same '3 3 4' format, with the tenth digit being a modulus 11 check-digit. The H&C Numbers are distinct from NHS Numbers as they are drawn from a reserved range (from 320 000 001 to 399 999 999 plus check digit)<sup>41</sup>. The H&C Number is being used as the master patient identifier within the new Northern Ireland 'Electronic Care Record' system<sup>42</sup>.

## Systems for managing the NHS Number (in England, Wales and IoM)

### *Historical Systems*

Until 1991 the NHSCR was managed as a paper based system. NHS Numbers, along with patient demographic and status information was organised in transcript books (indexed by NHSID) with a unique entry for each person. Each transcript book holds registrations for one registration area for a specific time period. In 1991, NHSCR introduced an in-house developed computer system known as CHRIS (Central Health Register Enquiry System)<sup>43</sup>. The CHRIS system (which went live on the 1<sup>st</sup> January 1991) was populated with the digitised records of all the live patients recorded in the register. It was subsequently updated with the records of new born babies and patients immigrating into England, Wales and IoM who then registered with GP. The CHRIS system was used by the MRIS team to provide services to researchers. These services included<sup>44</sup>:

- Record Linkage – a service to 'flag' lists of research study participants to their NHS register record. The linkage methodology details are provided in Appendix 1;
- Current Status – a status check of whether an individual is alive (reporting geographical location of practice registration) or dead (reporting death registration details);
- List-cleaning – providing the current status of 'flagged' patients (reporting fact and date of death, current forename, surname, date of birth, geographical location of practice registration);

---

<sup>40</sup> Methodology Paper – Mid-year population estimates for Northern Ireland. (2017). Great Britain: Northern Ireland Statistics and Research Agency.

<sup>41</sup> See NHS Data Dictionary, Available at:

[https://www.datadictionary.nhs.uk/data\\_dictionary/attributes/h/health\\_and\\_care\\_number\\_de.asp](https://www.datadictionary.nhs.uk/data_dictionary/attributes/h/health_and_care_number_de.asp)

<sup>42</sup> Purvis J. NIECR-a quiet revolution. The Ulster medical journal. 2015 Jan;84(1):1.

<sup>43</sup> NHS Information Centre takes control of NHS Central Register (Press Release). April 2008. Great Britain: NHS Information Centre. Available from:

<http://webarchive.nationalarchives.gov.uk/20090410030330/http://www.ic.nhs.uk/news-and-events/press-office/press-releases/april-2008/nhs-information-centre-takes-control-of-nhs-central-register>

<sup>44</sup> See archived records of MRIS website for detail (archive snapshot taken May 2008):

<http://webarchive.nationalarchives.gov.uk/20080527165020/http://www.ic.nhs.uk/our-services/improving-patient-care/medical-research-information-service>

- Long-term follow up – providing long-term morbidity and mortality data for a ‘flagged’ population (deaths and cancer registration details, exits and entries to the NHS, geographical location of practice registration).

The CHRIS system was replaced by the PDS as part of the NHS Spine program. Following a review of NHS data sharing activities in the wake of the care.data program<sup>45</sup> there was substantial change (in England) in the way in which researchers could link study cohorts to NHS Number and patient register details. These changes are outside the scope of this resource report.

### ***The NHS ‘Spine’***

The NHS Spine is the name given by the NHS to a collection of local and national (English, Welsh, IoM) databases and systems containing both patients clinical and demographic information. The Spine has been developed and is maintained by NHS Digital. The Spine comprises databases of information and ‘messaging’ functions capable of transferring and transmitting information between systems to agreed standards. Given that the NHS consists of over 27,000 individual organisations (e.g., primary and secondary care sites, pharmacies, opticians, dentists), a standardised system for transferring information is seen as essential to: allow efficient communication and flow of information; a means to improve patient safety through improved quality of information (e.g. reducing errors in identifying patients and accurately and efficiently collating relevant patient information); cost benefits through administrative efficiencies. The Spine is based around the notion of each patient being allocated a unique identifier (the NHS Number) and this number being used consistently across all record keeping and supported by rigorous data quality checks.

The Spine connects a wide range of systems, including the:

- Personal Demographics Service (PDS) – a database storing NHS Number and patient demographic information;
- Summary Care Record (SCR) – a centralised summary of patient’s clinical information;
- Secondary Uses Services (SUS) – a system centralising data from patient records to provide anonymised/pseudonymised business reports and statistics for (largely NHS) research, planning and public health delivery;
- GP2GP - the system transferring patient records between GP practices;
- Electronic prescription service – the system used to send electronic prescription records from General Practice’s to pharmacies.

### ***Patient Demographic Service (PDS) and the National Back Office (NBO)***

The PDS has replaced the NHSCR and is now the national (English, Welsh, IoM) electronic database of patient demographic details (i.e. the master patient register). It contains patient identifiers including the NHS Number, name, address, date of birth and associated non-clinical medical information such as their General Practice registration details. The PDS and the NBO share responsibility for managing the NHS Number and the associated patient

---

<sup>45</sup> Presser L, Hruskova M, Rowbottom H, Kancir J. Care. data and access to UK health records: patient privacy and public trust. Technology Science. 2015 Aug;2015081103.

demographic data held within the PDS database. The PDS role relates to allocating NHS Numbers, managing the record to reflect 'complex case' circumstances (described below), managing appropriate access to the record, flagging the record to reflect patient data sharing decisions, and flagging the record for research purposes. It is within the NBO remit to manage the quality of NHS Numbers through identifying problem records (e.g. duplicate records) and reconciling problems with incoming data feeds.

The aim of the PDS is to support the functionality of the Spine systems through enabling a patient to be readily identified by NHS staff and associated, quickly and accurately, with their correct medical details. It is designed to support clinical functions and the linkage requirements of the research community. Managed access for research purposes can also be provided by NHS Digital through the Data Access Request Service (DARS)<sup>46</sup>.

### **PDS contents**

Unlike the wider Spine, the PDS only holds patient demographic data. The database contains a wide range of administrative variables (see Table 1). Some of these are likely to change over time to reflect changing national programmes (e.g. variables flagging governance related values are likely to change to reflect new legislation or new public permission options).

**Table 1: Demographic fields within the Patient Demographic Service database.**

Source: adapted from NHS Digital website (21<sup>st</sup> September 2016, available from:

<http://webarchive.nationalarchives.gov.uk/20160921142924/http://systems.digital.nhs.uk/demographic/s/pds/contents>)

<b>PDS variable</b>	<b>Variable description</b>
NHS Number	The unique patient identifier.
Patient name	Including any previous names, aliases and preferred name, e.g. Chris rather than Christopher.
Date of birth	The patient's date of birth.
Place of birth	The patient's place of birth.
Additional birth information	The delivery time and birth order for multiple births.
Date of death	The patient's date of death. <sup>a</sup>
Death notification status	Indicates a formal death certificate has been issued for the patient and the death has been registered.
Gender	Administrative gender.
Address	Includes main, temporary and correspondence addresses.
Alternative contacts	The patient's legal guardian, proxy, family/close contact.
Telecommunication contact details	Contact details such as telephone number, fax number and email address.
Preferred contact times	Patient's preferred contact times.
Preferred contact method	The patient's preferred contact method, e.g. telephone contact by proxy, no telephone contact, sign language required in face-to-face contact or minicom.
Preferred written communication format	Specialised patient contact requirements, e.g. large print, Braille, audio tape.
Preferred language	Information on patient's preferred language of communication.
Interpreter required	Indicates that the patient requires an interpreter.

<sup>46</sup> The DARS access process is outside of the scope of this report. Application details are available via the DARS website (available March 2018): <https://digital.nhs.uk/data-access-request-service/process>



NHS Care Record consent to share status	Indicates that the patient has agreed to share their health record.
Nominated dispensing contractor	The patient's nominated dispensing contractor that could include a community pharmacy, dispensing appliance contractor and a dispensing doctor.
Reason for removal	Indicates that the patient is no longer registered with a GP, e.g. if a patient has moved from England or Wales.
Previous NHS contact indicator	Indicates that the patient confirms they have had previous NHS treatment. This allows the PDS National Back Office to check for a duplicate of a record.
Patient call-back consent status	Indicates that the patient is willing to be called back from a Choose & Book call-centre.
Shared secret	An encrypted password used to validate a patient's identity when contacted from a Choose & Book call-centre.
Sensitive record indicator	Indicates that either the record is not accessible to PDS users, or that the content of the record is being reviewed to ensure the data is correct.
Primary care	The GP Practice with whom the patient is registered.
Date of registration	The date the patient registered with the GP Practice.
Serial change number	The mechanism for synchronising local and national records.
NHAIS information	The PDS holds certain information to allow it to interact with the NHAIS system that administers primary care. This information is not routinely accessible by the NHS, but NHAIS demographic information is regularly synchronised with PDS demographic information.

<sup>a</sup>Contemporaneous records suggest the PDS also contains information on who reported the death, distinguishing between formal (where a registrar has issued a certificate), informal (where the NHS have recorded a patient death following a Medical Certificate of the Cause of Death being issued) and local (where a death has occurred but has not been formally notified) death status.

### **‘Complex cases’**

All patients should have a PDS record, although situations relating to ‘complex cases’ may mean it is difficult to trace some individual’s records<sup>47</sup>. For example, individuals serving within the Armed Forces, or those in Prison are likely to have an NHS Number and a PDS record – but are also likely to have different demographic details (either those provided by the study or those held by the NHS).

## **Governance issues specifically related to NHS Numbers**

NHS Numbers (in all formats<sup>48</sup>) are considered to be personal identifiers under the terms of the current Data Protection Act 1998 and the imminent Data Protection Act 2018 (EU General Data Protection Regulations). This introduces controls on how NHS Numbers should be used. It is however arguable – given that almost all UK citizens are routinely

<sup>47</sup> NHS Number Your Unique Patient Identifier: Guidance on categories of complex case. July 2012. Great Britain: Connecting for Care. Available from: <http://webarchive.nationalarchives.gov.uk/20160921163126/http://systems.digital.nhs.uk/nhsnumber/scaff/guidance/complex.pdf>

<sup>48</sup> i.e. this includes the English, Welsh and IoM NHS Number as well as the Scottish CHI, despite the fact that the NHS Number is an effectively random number whereas the CHI is a composite identify containing other identifiers (date of birth and sex). This is because both patient identifiers are information that relate to an individual.

allocated an NHS Number regardless of health status - that an NHS Number alone does not constitute 'health data' and therefore does not need to be treated as sensitive personal information (this argument clearly would not hold where the context of the NHS Number conferred health status, e.g., a list of patients receiving chemotherapy would be health data even if it only contained NHS Numbers). This has benefits given that the use of sensitive personal information requires additional controls than the use of personal information. Regardless, as NHS Numbers are personal information (and sometimes sensitive personal information) this introduces governance restrictions on how NHS Numbers are used. NHS Scotland have investigated this issue with the Information Commissioner's Office (ICO) – the national regulator for Data Protection Act in the UK – who consider that NHS Numbers should only be used for the purposes for which they were devised (note that this can, as is the case in Scotland, explicitly include for research purposes)<sup>49</sup>. The justification for this is that using any form of personal identifier for multiple purposes may introduce data quality issues, raise information security concerns and degrade the initial purpose of the identifier (as seen with the Social Security Identifier number in the USA<sup>50</sup>). Research data managers should be aware that the current ICO guidance is that non-health service organisations (i.e. research studies) may use (where they have an appropriate ethico-legal basis) NHS Numbers but not as the primary identifier within the research studies databases (i.e. the NHS Number is to be used as a tool to facilitate interactions with the health services only).

Access to PDS content (or the equivalent in Scotland and Northern Ireland) is covered by a complex range of ethico-legal considerations that are standard to any research application to access identifiable health information. It is outside the scope of this report to describe these requirements in any detail. However, there are also specific governance considerations that relate to certain PDS data that impact on research end users<sup>51</sup>. These are separate to the national 'opt-out' data sharing restrictions being introduced across England during 2018<sup>52</sup>. These include records relating to:

*Vulnerable patients:* Vulnerable patients (e.g. those suffering domestic violence, in the public eye, in witness protection programs) can request that their PDS entry is marked as 'sensitive' (a process known as 's-flagging' within the NHS). This process restricts access to contact information, but does not remove the information from the record itself. This may impact on research participant tracing initiatives.

*Adoptions:* The UK adoption process is managed by the General Register Office (GRO). The GRO will notify the PDS of an adoption. A new record (including a new NHS Number) will be created within the PDS reflecting the post-adoption details of the child. No link will be established between the pre-adoption and post-adoption records (or ID numbers). This impacts on longitudinal studies attempting to link identifiers collected during recruitment (e.g. during pregnancy) with participant records where the study is unaware of the adoption process. This may bias or complicate the interpretation of study findings in certain investigations (e.g. those relating to outcomes in vulnerable children).

*Gender reassignments:* It is a legal right (Gender Recognition Act 2004) that individuals can ensure that their official records accurately reflect their status. Once a patient informs their GP about their gender reassignment, information about this is transmitted to the PDS. The

---

<sup>49</sup> The use of the CHI (Community Health Index) to support integrated care across the NHS in Scotland. (2013). Great Britain: Scottish Government.

<sup>50</sup> Puckett C. The story of the social security number. Soc. Sec. Bull. 2009;69:55.

<sup>51</sup> Managing inappropriate access to patients' demographic information using IT and local systems and services. Great Britain: Health and Social Care Information Centre.

<sup>52</sup> A system being introduced in 2018 to allow patients to determine how their identifiable health care records are used. See here (available March 2018): <https://digital.nhs.uk/national-data-opt-out>

PDS will then create a new PDS entry with a new NHS Number. No link will be established between the old, and the new record. This could impact both on longitudinal studies tracing activities and strategies for participant follow-up through record linkage. This may bias or complicate the interpretation of study findings in certain investigations (e.g. those relating to outcomes in individuals following gender reassignment).

## Quality and coverage issues of relevance to LPS community

Accurate research depends on good quality data. Fortunately, this research need is increasingly aligned with the needs of the health services; where there is a clinically driven demand to ensure NHS Numbers are accurate (in order to enable the accurate collation of records belonging to the same individual and the efficient administration of the system) and a financially driven demand to ensure patient registers are accurate (given that NHS payments are frequently allocated on a capitation basis)<sup>53</sup>. For this reason (within England, Wales and the IoM) the NBO are mandated to manage the quality of NHS Numbers (e.g. checking for patients allocated multiple NHS Numbers) and NHS standards and policy require that General Practitioners and system designers conduct checks sufficient to minimise list inflation<sup>54</sup>. While even in comparatively recent history there have been high levels of error within patient registration status (e.g. the Office for National Statistics estimated that in 2010 the GP registered population exceeded the national population by ~2.8m, or 5.2% inflation<sup>55</sup> and that in Northern Ireland only an estimated three out of five patients leaving Northern Ireland to go abroad deregister from their family GP<sup>56</sup>), more recent evaluations have suggested good levels of quality across the system<sup>57</sup>.

### Quality and coverage

It is not possible to provide a comprehensive set of quality and coverage issues relating to NHS Number systems given the complexity of the systems, their variation over time and across the UK Home Nations and the variety of factors that could impact on quality and coverage (e.g. recording error, fraud, data handling error, linkage error). Rather than detail all possible issues, we highlight the following known issues that are pertinent to longitudinal studies.

1) *The master patient index (e.g. the PDS) is unlikely to include all citizens, and those absent are likely to be systematically different from those included.* This has the potential to introduce misreporting and may complicate recruitment to birth cohorts or assessments comparing cohort samples with wider population statistics. For example:

- Not every individual will have an NHS Number, and not all will be included in the patient index. For example: 1) those receiving entirely private health care treatment

---

<sup>53</sup> Tackling list inflation for primary medical services. (2013). Great Britain: NHS England. Available from: <https://www.england.nhs.uk/wp-content/uploads/2013/10/tack-infla.pdf>

<sup>54</sup> Tackling list inflation for primary medical services. (2013). Great Britain: NHS England. Available from: <https://www.england.nhs.uk/wp-content/uploads/2013/10/tack-infla.pdf>

<sup>55</sup> Tackling list inflation for primary medical services. (2013). Great Britain: NHS England. Available from: <https://www.england.nhs.uk/wp-content/uploads/2013/10/tack-infla.pdf>

<sup>56</sup> Methodology Paper – Mid-year population estimates for Northern Ireland. (2017). Great Britain: Northern Ireland Statistics and Research Agency.

<sup>57</sup> NHSCR data: quality assurance of administrative data used in population statistics. (December 2016). Great Britain: Office for National Statistics. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/nhscrdataqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016>

may not have an NHS Number (although many will as their treatment will involve some degree of NHS involvement); 2) coverage in the 1939 National Register census was impacted by people not completing the census in order to avoid conscription; those migrating in and out of the system may have partial coverage (and this may include those moving between the UK's Home Nations).

- Historically, NHS Numbers were allocated as part of the Birth Registration process, which occurred days or weeks after birth. This could lead to missed registrations from those emigrating or where children died during the neonatal period. Dickinson et al. (1997)<sup>58</sup> found that in Cumbria during the 1950s, 31% of children who died in the neonatal period either were not entered on NHSCR, or if they were entered, there was no record of their death. However, by the 1970s the register was 99% complete.
- There are variations in policy for registering stillborn deliveries across the UK Home Nations<sup>59</sup>. Within England a stillborn delivery must be registered within 42 days, in Scotland registration must occur within 21 days and in Northern Ireland registering stillborn deliveries is optional (and can take place within a year from delivery). These policy variations may have introduced differences to NHS Number allocation while allocation was linked with birth registration.

2) *Recording quality and coverage has been impacted by changes in which migration is recorded and managed.* This could impact on research considering migration (as patients moving house within or across geographical areas may not change their GP and would therefore not generate a migration notification), research using spatial analysis techniques or on studies tracking & tracing strategies. It is worth considering that these inconsistency are likely to have temporal (i.e. where linked to change in policy or system) and socio-demographic patterns (e.g. students attending University while retaining a GP registration based on their home address). Issues include:

- Patients moving house will have delayed registration at a new General Practice. It is not a legal requirement in the UK to update your health registration status when you move home. The time period between moving and creating a new registration (known as 'lagging') is likely to be patterned by health and social characteristics, with parents of young children, the elderly and those with chronic health conditions most likely to have a short 'lag' period and male patients aged 18-44 most likely to have a longer 'lag' period<sup>60</sup>.
- Practice relating to permanent emigration has changed over time. Historically, permanent embarkation (emigration) was recorded as a result of patients handing in their 'medical card' at the point of emigration. The move away from using 'medical cards' has meant this practice no longer occurs and may as a result lead to some under-reporting of emigration<sup>61</sup>.

---

<sup>58</sup> Dickinson HO, Parker L, Harris D, Botting B, Lawson A. Audit of ascertainment of deaths to children born in Cumbria, UK, 1950-89 through the NHS central register. *Journal of Epidemiology & Community Health*. 1997 Aug 1;51(4):438-42.

<sup>59</sup> See: <https://www.gov.uk/register-stillbirth>

<sup>60</sup> Tackling list inflation for primary medical services. (2013). Great Britain: NHS England. Available from: <https://www.england.nhs.uk/wp-content/uploads/2013/10/tack-infla.pdf>

<sup>61</sup> NHSCR data: quality assurance of administrative data used in population statistics. (December 2016). Great Britain: Office for National Statistics. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/nhscrdataqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016>

- There is at times ambiguity between place of registration and place of residence. For example, at times it has been policy that patients must register with a GP near their home, at other times patients are free to choose their GP, or where the geographical information reported in the patient register is based on patient residence or the patient's registered practice.

3) *System and data management error will impact on data quality.* These factors have the potential to impact on all aspects of epidemiological research, from reducing linkage quality, the onward use of NHS Number to identify and extract other records and the misreporting or misclassification of patient status. Coupled with the issues relating to coverage of NHS Number described above, it is worth those designing linkage algorithms to consider the merits of linkage passes that do not involve the NHS Number (even when theoretically present in both datasets).

- *The system will be impacted by data management error.* For example, the quality of information being returned to 'flagged' cohort studies could be impacted by transcription errors introduced through the paper based system, or duplicated NHS Numbers or misallocated NHS Numbers introduced through disjointed administration processes.
- *Policy and system changed impacts on the range of data collected and the way they are shared.* This can manifest itself in a reduced granularity of information provided (e.g. the NHSCR postings notifications to flagged longitudinal studies use to distinguish between patient movements due to 'Diplomatic Embarkation', 'Prisoner' and 'Psychiatric Patient', these are no longer recorded as separate posting classifications, but are instead aggregated to the broader category of 'Cancellation'<sup>62</sup>).

## The use of NHS Number and PDS demographic data for record linkage purposes.

Studies using principally NHS data or linking NHS data to other sources to answer health research questions are increasingly common and are well documented; with the Understanding Patient Data website providing examples of good practice<sup>63</sup>. There are fewer examples of methodological studies focusing on issues associated with linkage to NHS records (e.g. quality assessments). This overview provides a summary of publications of interest to researchers establishing linkages to NHS records via NHS Number and the PDS.

### **Linking error**

A study by Harron (2017)<sup>64</sup>, aimed to utilise identifier error variation to suggest new ways to improve data linkage. In this example HES records were linked via NHS number with data from the PDS. Findings showed an 0.11% identifier error (disagreements between HES-PDS records or missing values) for sex and date of birth and 53% of records for postcode. Identifier error rates were also found to differ significantly by age, ethnicity and sex. Other

---

<sup>62</sup> NHSCR data: quality assurance of administrative data used in population statistics. (December 2016). Great Britain: Office for National Statistics. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/nhscrdataqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016>

<sup>63</sup> See: <http://understandingpatientdata.org.uk/case-studies>

<sup>64</sup> Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. Utilising identifier error variation in linkage of large administrative data sources. *BMC medical research methodology*. 2017 Dec;17(1):23.

important groups can be disproportionately affected by linkage error, for example when linking mother and baby data, babies who die in the first couple of days of life may not be linked because their name or NHS Number may not have been allocated before death<sup>65,66</sup>.

Another problem occurs when patient identifiers are either incorrect or missing, as this will lead to underestimates in readmission rates<sup>67</sup>. This problem results in hospitals with poor quality identifiers appearing to have superior performance metrics. The GUIDance for Information about Linking Data sets (GUID) highlights areas of the linkage pathway when error can be introduced and suggest information that can be used to assess/account for linkage error<sup>68</sup>.

As well as considering the linkage pathway, another way to minimise error in data linkage is to improve both data quality and the algorithm used for linkage<sup>69</sup>. In a study by Hagger-Johnson (2015), the algorithm used to link together care records belonging to the same patient using NHS hospitals in England HES data, was thought to have a missed match rate of at least 4%<sup>70</sup>.

### ***Multi data source approach***

To minimise issues with linkage error, studies such as the one by Britton et al. (2012)<sup>71</sup>, aimed to test and compare the validity of linked data sources using a multi sourced data approach. In this example, self-reported stroke events are compared against information in hospital medical records vs HES in the Whitehall II longitudinal study. This study found that a third of self-reported stroke events were not recorded in HES, however, the majority of missing events could be validated through hospital notes and GP records. Similarly, Dickinson et al. (2001)<sup>72</sup> assessed the completeness and accuracy of cancer registrations within the NHSCR. The findings of these studies highlight the importance of a multi-source data approach when working with routinely collected data records.

---

<sup>65</sup> Hummler HD, Poets C. Mortality of extremely low birthweight infants-large differences between quality assurance data and the national birth/death registry. *Zeitschrift fur Geburtshilfe und Neonatologie*. 2011 Feb;215(1):10-7.

<sup>66</sup> Anthony S, Bruin KM, Graafmans WC, Dorrepaal CA, Borkent-Polet M, van Hemel OJ, Jansen FH, Ouden A. The reliability of perinatal and neonatal mortality rates: differential under-reporting in linked professional registers vs. Dutch civil registers. *Paediatric and perinatal epidemiology*. 2001 Jul 1;15(3):306-14.

<sup>67</sup> Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, Dibben C, Goldstein H. GUID: GUIDance for Information about Linking Data sets. *Journal of Public Health*. 2017 Mar 28:1-8.

<sup>68</sup> *Ibid*.

<sup>69</sup> Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. *Journal of innovation in health informatics*. 2017;24(2):234-46.

<sup>70</sup> Hagger-Johnson G, Harron K, Fleming T, Gilbert R, Goldstein H, Landy R, Parslow RC. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ open*. 2015 Aug 1;5(8):e008118.

<sup>71</sup> Britton A, Milne B, Butler T, Sanchez-Galvez A, Shipley M, Rudd A, Wolfe CD, Bhalla A, Brunner EJ. Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): Extracting information from hospital medical records versus the Hospital Episode Statistics database. *BMC medical research methodology*. 2012 Dec;12(1):83.

<sup>72</sup> Dickinson HO, Salotti JA, Birch PJ, Reid MM, Malcolm A, Parker L. How complete and accurate are cancer registrations notified by the National Health Service Central Register for England and Wales?. *Journal of Epidemiology & Community Health*. 2001 Jun 1;55(6):414-22.

### *Linkage techniques*

The Millennium Cohort Study (MCS) have augmented research questions and validated responses using linkage variables in the form of birth and hospital records<sup>73</sup>. In a study focused on linking the MCS to hospital records, they described linkage to the NHS using a range of personal identifiers, rather than including NHS Number which was not available to the study team at that time.

This study illustrates the variation in linkage methodologies between England, Wales, Scotland and Northern Ireland. Deterministic matching in England is more common. This uses different combinations of identifiers (e.g. names, date of birth, addresses and hospital information) in a hierarchical approach, to account for missing and incomplete data<sup>74</sup>. Hockley et al. (2008) describe how deterministic algorithms designed to minimise false matches have the disadvantages of a high missed match rate<sup>75</sup>. In contrast, linkage to Scottish records uses probabilistic matching approaches, here weights are assigned, and a threshold set, all weights which score higher than threshold were deemed to be true links. The probabilistic method used in Scotland yielded a higher matching rate than the deterministic methods used in other countries. Hockley attributed this to be, in part, due to more complete data and longer experience in linking routinely collected data and that a lack of completeness of data in other UK countries meant there would not have been significant improvement if the probabilistic method had been used.

A number of other studies conclude that probabilistic data linkage is known to produce more accurate linkage and less biased results than deterministic linkage, particularly in settings where data quality is poor<sup>76, 77, 78</sup>. This is demonstrated in a paper by Hagger-Johnson et al. (2017)<sup>79</sup>, that reported a study aiming to compare an existing NHS Hospital Episode Statistics (HES) identifier (HESID) deterministic algorithm with one that included an additional probabilistic step and use this to quantify missed and false matches. Missed matches are more common for ethnic minorities, those living in areas of high socio-economic deprivation, foreign patients and those with 'no fixed abode'. This is significant because it will create issues with bias when researchers are analysing data. The study concluded that probabilistic linkage using HES data reduced missed matches and bias in estimated readmission rates.

---

<sup>73</sup> Hockley C, Quigley MA, Hughes G, Calderwood L, Joshi H, Davidson LL. Linking Millennium Cohort data to birth registration and hospital episode records. *Paediatric and perinatal epidemiology*. 2008 Jan 1;22(1):99-109.

<sup>74</sup> *Ibid.*

<sup>75</sup> Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. *Revista de saúde pública*. 2009 Oct;43(5):875-82.

<sup>76</sup> Baldwin E, Johnson K, Berthoud H, Dublin S. Linking mothers and infants within electronic health records: a comparison of deterministic and probabilistic algorithms. *Pharmacoepidemiology and drug safety*. 2015 Jan 1;24(1):45-51.

<sup>77</sup> Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies. *PLoS One*. 2015 Aug 24;10(8):e0136179.

<sup>78</sup> Gill LE. OX-LINK: the Oxford medical record linkage system. In *Proc. Int'l Record Linkage Workshop and Exposition 1997* (pp. 15-33).

<sup>79</sup> Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linking to enhance deterministic algorithms and reduce linkage errors in hospital administrative data. *J Innov Health Inform*. 2017; 24(2).

In a similar study by Harron et al. (2016)<sup>80</sup> comparing deterministic and probabilistic linkage approaches for linking maternal delivery records with the birth record of their babies within pseudonymised HES records. This study explores the merits of deterministic linkage which require exact or approximate agreement on set of identifiers (e.g. sex, postcode, dob). Exact deterministic results in very few false-matches, yet can result in low match rates. This study firstly used exact deterministic linkage to bring together maternal and baby records, and then supplemented this with probabilistic linkage of remaining unlinked records. Harron concluded that probabilistic linkage of the maternal and newborn record is an efficient way to enrich maternity data.

These studies illustrate that linking individuals to NHS records using NHS Number and other personal identifiers present either within the register database, or clinical information within the records themselves poses significant challenges. This is likely to be the case whether linking within a single data source (e.g. linking participants in a longitudinal study to their HES via information in the patient register), or where using a multi-source approach to cross-reference information from a range of sources (e.g. linking related records – such as cancer registrations - from a range of local and national sources). In order to achieve the most accurate and complete linkage, a number of approaches and techniques can be adopted. Firstly, using data from a number of sources can add depth and enable assessments of validity. Secondly, contextual consideration of the matching techniques/algorithms used is important and variations in approaches may yield quality and coverage benefits. Regardless of approach, the underlying quality and coverage of robust unique identifiers (i.e. NHS Number) is important in maximising linkage outcomes. Any interpretation of linkage validity and issues relating to potential bias will need to take into account the manner in which the records are populated and the nature of the record management systems.

---

<sup>80</sup> Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking data for mothers and babies in de-identified electronic health data. PLoS One. 2016 Oct 20;11(10):e0164667.



## Appendix A: MRIS Record Linkage Protocol

The following summary information was taken from the MRIS website<sup>81</sup>.

### The matching process

We will use the information you give us about your study subjects to find their current status or carry out long-term follow-up. The cost of the service will depend on the quality of the demographic data that we receive from you.

Factors which affect matching rates include  
what information is available  
how accurate and up-to-date the information is

There are four types of matching procedures  
Band A Auto-match (low cost and resource)  
Band B Operator match  
Band C Manual with NHS number match  
Band D Manual without NHS number (high cost and resource)  
We start with auto-matching and continue until we get a match or until the customer has specified to stop (according to budget).

### Band A - Auto-match

Where we can match records automatically by computer without an operator.

#### To do this we require:

full surname  
full forename  
date of birth

#### What counts as an auto-match?

A unique match following strict matching criteria  
Matching criteria is tight so that automatic mismatches are minimal

#### What does not count as an auto-match?

More than one case with two or three parts of date of birth matching  
One complete match if another case exists with two out of three parts matching

### Band B - Operator match

Where an automatic match was not found or where more than one possible match was clarified by an operator. Additional address information is helpful at this stage.

### Band C - Manual with NHS number match

Where Band A and B weren't successful or for those cases who died before 1991 and an NHS number is supplied. A manual search of non-computerised records is done.

### Band D - Manual without NHS number

Cases where a manual search is required, and no NHS number was provided.

**Any further information that is available might be useful for matching. Please contact us for more details.**

<sup>81</sup> MRIS website snapshot taken on 27<sup>th</sup> May 2008. Available from:  
<http://webarchive.nationalarchives.gov.uk/20080527180240/http://www.ic.nhs.uk/our-services/improving-patient-care/medical-research-information-service/what-we-can-do-for-you/the-matching-process>

**CLOSER**

UCL Institute of Education  
20 Bedford Way  
London  
WC1H 0AL  
United Kingdom

Tel: +44 (0)20 7331 5102

Email: [closer@ucl.ac.uk](mailto:closer@ucl.ac.uk)

Web: [www.closer.ac.uk](http://www.closer.ac.uk)

Twitter: [@CLOSER\\_UK](https://twitter.com/CLOSER_UK)

YouTube: [CLOSER](https://www.youtube.com/CLOSER)