Office for
National Statistics

Patient Register: quality assurance of administrative data used in population statistics, Dec 2016

The quality assurance undertaken on administrative data for patient register data used within Population Statistics Division (PSD) publications.

Contact:
Pete Large
pop.info@ons.gsi.gov.uk

Release date:
20 December 2016

Next release:
To be announced

Table of contents

1 . Executive summary

This report shows how the Quality Assurance of Administrative Data (QAAD) Toolkit has been applied to NHS Patient Register data for England and Wales.

The NHS Patient Register provides a broad source within England and Wales that covers a large proportion of the population, with a good degree of accuracy for statistical purposes. The source provides demographic type data; no medical data are provided from this source.

The NHS Patient Register is used directly, and indirectly, in the production of a range of our high-profile statistics relating to population and migration estimation.

The Patient Register has a number of issues when used for statistical purposes. The source has a number of both under- and over-coverage issues (when compared with the target statistical population) and time lags in the data. The source has limited audit and there is potential for distortive effects because of its role in General Practitioner (GP) finance. The effect of these issues will vary by geography, age and sex, the main variables in demographic statistics.

Measures are in place to take account of the statistical quality issues through statistical production methods and the use of complementary data sources.

As a broad coverage source the NHS Patient Register provides an important source of data that is essential to the production of population statistics, with a high degree of accuracy. The Register is supplemented by complementary data and methods in order to produce quality statistical outputs.

2 . Introduction

The NHS Patient Register is a record of all persons registered with a General Practitioner (GP) in England and Wales. The NHS Patient Register was used to maintain an accurate list of all persons registered with a GP, allowing the timely transfer of medical records and correct payments to doctors. It contains a list of everyone who is, or ever has been, registered with a GP in England and Wales since the NHS was founded in July 1948.

The NHS Patient Register extract provided to us currently contains approximately 58 million records, extracted from 87 separate National Health Applications and Infrastructure Services (NHAIS) sites; 82 of which are in England and 5 of which are in Wales. Every NHAIS site holds a register of the patients registered with GPs within their area of responsibility.

Uses in Population Statistics Division (PSD)

Internal migration

Change in address for an individual between annual extracts is used (with other sources) to help estimate internal migration, that is, migration within England and Wales[1]. Internal migration is a component of our mid-year population estimates for England and Wales, a National Statistics release.

Local authority distribution of international migrants

The data are not used to calculate the level of international migration: they are used as 1 of the sources to distribute estimates of international migration into the UK to geographic areas. The Patient Register holds a “flag” for persons whose previous place of residence was outside of the UK. These data are used (together with other sources) to help allocate separate estimates of international migration into the UK to local authority areas. International migration is a component of our mid-year population estimates for England and Wales, a National Statistics release.

Small area population estimates (SAPE)

The change in the number of people on the Patient Register at Super Output Area level is used (together with other data) to calculate our SAPE using a ratio change method; SAPE for Super Output Areas are a National Statistics release.

The number of people on the Patient Register for Output Areas is used to apportion Super Output Area estimates to Output area.

Assessment of quality assurance level

The Quality Assurance of Administrative Data (QAAD) Toolkit sets out 4 levels for the quality assurance that will be required of a dataset: A0 – no assurance; A1 – basic assurance; A2 – enhanced assurance; and A3 – comprehensive assurance. The UK Statistics Authority state that the A0 level is not compliant with the Code of Practice for Official Statistics. The assessment of the assurance level is in turn based on a combination of assessments of data quality risk and public interest.

The toolkit sets out the level of assurances required as follows:

A0 – No assurance	Not compliant with the Code of Practice for Official Statistics
A1 – Basic assurance	Statistical producer has reviewed and published a summary of the administrative data quality assurance (QA) arrangements
A2 – Enhanced assurance	Statistical producer has evaluated the administrative data QA arrangements and published a fuller description of the assurance
A3 – Comprehensive assurance	Statistical producer has investigated the administrative data QA arrangements, identified the results of independent audit, and published detailed documentation about the assurance and audit

The following sections set out the level of quality assurance that is required for each of the uses of Patient Register data in PSD, explaining why the level is required. The issues described in this assessment are covered in more detail later in this document.

Quality assurance level for the NHS Patient Register: A3 – Comprehensive assurance

Where PSD uses the same data source for a number of different purposes it is normally more efficient to undertake some of the quality assurance only once. Each separate use may have a different level of assurance required. As a consequence, the highest level of quality assurance, from the different uses, will be applied to common quality assurance.

Summary of quality assurance level for NHS Patient Register

Use	Risk of quality concern	Public interest	Assurance level
Overall assessment			A3 – comprehensive
Internal migration	High	Medium	A3 – comprehensive
Local authority distribution of international migrants	Low	High	A2 – enhanced
Small area population estimates	Medium	Low	A2 – enhanced

Internal migration

Quality assurance level for internal migration: A3 – Comprehensive assurance

Level of risk of data quality concerns: High

The level of risk of data quality concerns is high because:

- the data are collected by multiple collection bodies; data are collected from approximately 8,000 GP practices
- the data are collated via intermediate bodies, the 87 National Health Applications and Infrastructure Services (NHAIS)
- audit and cleaning practices will differ by collection bodies, and therefore on a geographical basis
- there is limited independent audit – the last national audit, the National Duplicate Registration Initiative (NDRI) was undertaken in 2009 by the then Audit Commission
- there are known quality and statistical bias issues for the data source (see ‘Quality issues’ section)
- the NHS Patient Register is the main source for internal migration estimates
- the impact of these risks is minimised by having a high-quality Service Level Agreement in place

Public interest profile of the statistics: Medium

The level of public interest in the statistics produced is medium because internal migration estimates are published National Statistics. They are an important component of mid-year population estimates, which:

- are very high-profile National Statistics
- are politically sensitive as they are used in resource allocation
- have wide media interest
- are required under European legislation
- are used as the basis of population projections, which are used to allocate local authority funding and as a basis for housing plans
- have a wide variety of other onward uses

Additional information

The use of Patient Register data for internal migration relies on having an up-to-date residential address recorded. For administrative purposes this is less vital.

Due to the nature of the risk, for example, lags in registration, especially amongst young males, and the impact on subsequent statistics (population estimates and projections), an A3 level of assurance has been decided.

Local authority distribution of international migrants

Quality assurance level for local authority distribution of international migrants: A2 – Enhanced assurance

Level of risk of data quality concerns: Low

The level of risk of data quality concerns is low because:

- the data are collected by multiple collection bodies; data are collected from approximately 8,000 GP practices
- the data are collated via intermediate bodies, the 87 National Health Applications and Infrastructure Services (NHAIS)
- audit and cleaning practices will differ by collection bodies, and therefore on a geographical basis
- there is limited independent audit – the last national audit, the NDRI was undertaken in 2009 by the then Audit Commission
- there are known quality and statistical bias issues for the data source (see ‘Quality issues’ section)
- the NHS Patient Register is an important source for distributing international migration estimates geographically to local authority level – but does not affect the overall national level
- the impact on the statistic is also mitigated by the use of other data
- the impact of these risks is minimised by having a high-quality Service Level Agreement in place

Public interest profile of the statistics: High

The level of public interest in the statistics produced is high because local authority distribution of international migrants form part of a very high-profile published National Statistic; however, this source contributes only to the geographical estimation of flows to local areas.

Local authority distribution of international migrants form part of an important component of mid-year population estimates, which:

- are very high-profile National Statistics
- are politically sensitive as they are used in resource allocation
- have wide media interest
- are required under European legislation
- are used as the basis of population projections, which are used to allocate local authority funding and as a basis for housing plans
- have a wide variety of other onward uses

Quality assurance level for small area population estimates (SAPE): A2 – Enhanced assurance

Level of risk of data quality concerns: Medium

The level of risk of data quality concerns is medium because:

- the data are collected by multiple collection bodies; data are collected from approximately 8,000 GP practices

- the data are collated via intermediate bodies, the 87 National Health Applications and Infrastructure Services (NHAIS)

- audit and cleaning practices will differ by collection bodies, and therefore on a geographical basis

- there is limited independent audit – the last national audit, the NDRI was undertaken in 2009 by the then Audit Commission

- there are known quality and statistical bias issues for the data source (see ‘Quality issues’ section)

- the NHS Patient Register is the main administrative data source for SAPE, however, the data are only used to distribute to local authority totals

- the impact of these risks is minimised by having a high-quality Service Level Agreement in place

Public interest profile of the statistics: Low

SAPE are a published National Statistic of primarily local interest.

Selection of dataset

The Patient Register was first used as a source in the calculation of population statistics in the mid-1999 population estimates, where it replaced the electoral roll in the calculation of internal migration. Our research concluded that the Patient Register provided a better estimate of internal migration. Further research has resulted in the use of other data to supplement Patient Register data (see ‘Use of student data’ below).

Practice areas associated with data quality

The Quality Assurance of Administrative Data (QAAD) Toolkit describes 4 practice areas for quality assurance. The assurance undertaken for each practice area is described in the following sections.

Notes for: Introduction

1. Migration within Scotland and within Northern Ireland are calculated separately by, respectively, National Records of Scotland and the Northern Ireland Statistics and Research Agency. A separate adjustment for “cross border” flows between the countries is made.

3 . Operational context and administrative data collection

Detailed description of the administrative system and operational context

Organisations collecting the data

The data are collected by approximately 8,000 NHS General Practices in England and Wales [1]. These data were collated by 87 separate National Health Applications and Infrastructure Services (NHAIS) sites.

Original purpose for data collection

The NHS Patient Register is used to maintain an accurate list of all persons registered with a General Practitioner (GP), allowing the timely transfer of medical records and correct payments to doctors. By registering with a GP practice an individual secures access to GP services at that practice; although they can also be treated at another practice in an emergency for up to 14 days, thereafter they will have to register as a temporary patient.

Legal basis for collection

There is no legal requirement for patients to register with an NHS GP. GP practices are required to allow patients to register, except under specific circumstances, and patients are not required to prove identity or entitlement in order to register or to receive treatment.

Under the terms of their primary medical services contracts, GP practices cannot refuse an application to join its list of NHS patients on the grounds of race, gender, social class, age, religion, sexual orientation, appearance, disability or medical condition.

Other than that, they can only turn down an application if:

- the commissioner has agreed that they can close their list to new patients
- the patient lives outside the practice boundary
- they have other reasonable grounds

In practice, this means that the GP practice's discretion to refuse a patient is limited.

When applying to become a patient there is no regulatory requirement to prove identity, address, immigration status or the provision of an NHS number in order to register. However, there are practical reasons why a practice might need to be assured that people are who they say they are, or to check where they live, so it can help the process if a patient can provide relevant documents. There is however no contractual requirement to request this nor is establishing an individual's identity the role of General Practice.

Any practice that requests documentation regarding a patient's identity or immigration status MUST apply the same process for all patients requesting registration.

As there is no requirement under the regulations to produce identity or residence information, the patient MUST be registered on application unless the practice has reasonable grounds to decline. Registration and appointments should not be withheld because a patient does not have the necessary proof of residence or personal identification. Inability by a patient to provide identification or proof of address would not be considered reasonable grounds to refuse to register a patient. For more information please see Patient Registration - Standard Operating Principles for Primary Medical Care (General Practice), 27 November 2015, NHS England .

In January 2016, NHS England published the Policy Book for Primary Medical Services. This document says commissioners (NHS England and Primary Care Support Services (PCSS)) have “an obligation to prepare and keep up to date a list of patients accepted by the contractor or assigned to its list of patients and who have not been removed from that list.” (Chapter 9, paragraph 1.3).

Description of how the data are collected

When a child is born the midwife in attendance provides the mother with the child’s NHS number. This number is linked to during the birth registration and this updates the information on the Patient Demographics Service (PDS) system.

When an individual registers with a GP practice, the GP system will update both the PDS and NHAIS systems using the NHS number. If the NHS number is not easily determined then a temporary number is allocated and a trace to determine the true NHS number is carried out by the National Back Office. If the NHS number is unable to be traced a new one may be allocated, the provision of an NHS number does not reflect the entitlement to free NHS treatment.

The Patient Register extract for us is taken from the NHAIS system and used in the calculation of population statistics.

Registration with a GP

In most practices registration with a GP will be done by the GP practice’s administrative staff – for example the receptionist.

Standard registration with a GP is via form GMS1.

Registration with a GP revolves around the NHS number. “All NHS patients that present to a GP will either:

- already have an NHS number
- will be allocated an NHS number through the GP registration process

Everyone born in England, Wales or the Isle of Man, and everyone who has ever registered with a GP practice in England, Wales or the Isle of Man will have an NHS number.” (NHS Number Guidance for GP Practices V1.1, June 2011, HSCIC)

A number of situations can arise following completion of the form:

- when a patient registers with an NHS number guidance to GP practices is to use another demographic, such as date of birth, to confirm the number is correct

- when a patient registers without an NHS number the GP practice will either seek to trace the number themselves, or request the number via the GP Links Service

- patients born in Scotland do not get a PDS (Patient Demographics System) NHS number so patients new to the NHS in England and Wales, born in Scotland, will also be allocated a new NHS number by the PCSS

- when a patient transfers from Northern Ireland the NHS number will be obtained via the PCSS

- when a patient is new to the NHS (normally a first migration into the UK), and over 6 months old, a new NHS number will be supplied by the PCSS

- when a patient is new to the NHS, but they are not "ordinarily resident" (guidance is staying in the UK for under 3 months) a temporary NHS number will be allocated

- at the first registration after a birth, the NHS number may be available from maternity lists, discharge papers, or documentation (child health record) held by the parent, and can be obtained or confirmed via the PDS trace system

- where a number cannot be traced a temporary number can be issued

The guidance does not cover registration of babies under 6 months old born outside the UK.

The registration details are entered onto the local GP system.

Assignment to a practice

Where a patient is resident within a PCSS's area, and has been unable to register with a GP, the PCSS can at the patients request assign the patient to a practice, see Annex A for the form that accompanies this process. There are a number of safeguards within this process before the assignment is made.

These registration details are compiled onto the local patient registration database that covers the GP practice along with others within a given geographical area (the National Health Applications and Infrastructure Services (NHAIS) database.

Patients between lists

Patients who are between lists at practices either at the patient's, or separately doctor's request are included on the data held at NHAIS and supplied to us using 2 temporary "practice" codes.

Differences across areas in the collection and recording of the data

There is some common guidance on patient registration (see Patient Registration - Standard Operating Principles for Primary Medical Care (General Practice) and NHS Number Guidance for GP Practices V1.1 for example). However, within this guidance there is scope for difference in implementation. Methods vary from GP practice to GP practice, and even between individuals interacting with systems. Computer systems also vary with some common systems, but some local systems are unique to individual GP practices.

Some practices have direct access to central systems, notably the Personal Demographics Service (PDS) through their organisation's local system, or by using a secure web-based portal; other practices will work through their PCSS.

As a result of these differences there will be variation in the quality of data recorded, with the verification and checking of data entered.

In addition to national audit exercises, periodically GP practices or local NHS bodies have undertaken exercises to “clean” their lists by removing records where there has been no patient contact for an extended period. This will have a geographically differential impact on data produced from GP lists.

The NHS is administered separately in England and Wales, with different organisations. Standards, guidance and practice may differ between the 2 countries. The NHAIS system is used to compile the data from both countries and produce the NHS Patient Register data that we receive for England and Wales as a combined dataset.

We do not hold record level details for Scotland or Northern Ireland.

Data item issues

General: Data are often entered manually, based on handwritten forms, so are prone to transcription errors (for example, name misspelling).

Flag 4: Flag 4s are 1 of a number of flags that indicate a status of a patient’s previous address; in the case of Flag 4 the presence of the flag indicates that the previous address of the patient was outside the UK. When using these indicators for statistical use, there are 2 features of these flags that may have an impact:

- Flag 4s are removed on any move within the UK – a patient may arrive from abroad, live in 1 address and quickly move to another address and it is then no longer possible to tell that they have recently arrived from outside the UK

- Flag 4s are added where the address of previous residence is outside the UK; this may be as the result of a short residence outside the UK that does not fit with statistical definitions of migration – a patient may be a UK resident, move abroad for a relatively short period and then return to the UK

Date of birth: Where a date of birth is unknown a default date of birth is sometimes used. Common examples are 1 January 1900 and 1 January 1901. 1 January may also be used by patients from cultures where date of birth is not routinely recorded, and will indicate an approximate age.

Issues in design and definition of targets

Where a Patient Register contains records of people who should no longer be on the list this is known as list inflation. There are a number of reasons for this and GP practices can look to address this issue through list cleaning.

GP practices are paid based on the number of patients recorded on their Patient Register.

GP practices can face a high turnover of patients, which can also give rise to list inflation if subsequent registrations are not processed in a timely manner.

There are records that link to patients who do not de-register; there are 2 principal reasons why an individual may not de-register. Firstly, some people may be slow to re-register at other GP practices, despite having moved away from the area, as they are relatively healthy and do not see re-registration as a priority (this affects list inflation at the local level). Secondly, people who live abroad (who are required to de-register from their GP practice) fail to do so either because they do not see it as a priority when moving abroad or because they wish to have continued access to the NHS; these people are referred to as “ghost patients” (this affects list inflation at both the local and national level).

There are various reasons as to why list cleaning is not undertaken such as GP practices not having enough resource to carry out the work and there being a cost, beyond staff time, to the GP practice in carrying out the work.

Potential sources of bias and error in the administrative system

Time lags in updating addresses – patients: In addition to any lags by GP practices (see ‘Issues in design and definition of targets’ above), patients can be slow to inform their GP, or re-register with a new GP, following a change of address. We have published research on this topic in An analysis of patient register data in the Longitudinal Study – what does it tell us about the quality of the data? Whilst this can happen at any age, the typical delay does vary by age and sex, by family status and by health status. These delays may actually mean that, if an individual has a number of changes of address then only some of these addresses end up being recorded on the Patient Register. Typically the following groups tend to be quicker to register a change of address:

- mothers with young children
- those with ongoing health conditions
- the very elderly

Additionally, the following tend to be slower to register a change of address:

- young healthy adults, especially males – including students
- highly mobile individuals
- healthier persons, especially males
- males – in general

Parental home: Mobile young adults may choose or default to remaining registered at their parental address; this is similar to the lags issue above.

Shared custody: there is potential for a range of issues where there is shared custody of children at different addresses. These can potentially include:

- split residence not reflected in the record
- duplicate registrations
- use of different names (particularly surnames)

Care homes (and similar institutions): Where a patient moves into a care home this may not be recorded as a change in address, particularly if the intended stay is short. These stays may become longer so that there has in effect been a change in residence, although the GP register has not been updated. By contrast a move may be intended as long-term (and the move recorded on the Patient Register) but the patient may die soon after the move. There are 2 potential consequences for statistical use of data:

- a spell of residence away from the home address can be missed
- a death can be recorded with a place of residence given as different from that recorded on the Patient Register

Duplicate records: There are a number of occasions where duplicate records can occur. These can be classified as duplicates with the same NHS number, and duplicates with a different NHS number. Duplicate records with the same number could be:

- records for the same person, with the same number, held on 2 different lists – the Audit Commission said that the majority of these were temporary where a patient was in the process of transferring from one practice to another.
- records for 2 different people with the same number – the Audit Commission said that these were rare and the cause was not known; potential reasons could include error by patient in writing in an NHS number (and where a check is not undertaken or is passed by chance – for example, same date of birth), clerical (typing or legibility) error on NHS number input, mismatching a patient to another with similar details on registration and allocating the wrong number or fraud. (Anecdotally there have been reports of people "sharing" NHS numbers in certain communities.)

Duplicates with different numbers could be because:

- on registration with a GP a patient is incorrectly identified as a new entrant to the NHS (returning migrant, exit from armed forces, return from private practice)
- on registration with a GP no match is traced to previous records
- on registration with a GP a patient gives insufficient details to allow a match (as above, patients do not legally have to provide proof of identity).

Snapshot: The data that we currently use from the Patient Register are based on a (typically annual) snapshot. This gives the position at that point in time, as it occurs on the register. This approach means there is no "history" between snapshots of changes. For example, where there are 2 or more changes of address in a year only the first and last of these addresses are captured (implying a single move). Moves before exiting from the system (including emigration and death) can also be missed, as can moves shortly after entry to the system (including immigration and birth).

Coverage: The Patient Register is a broad coverage source, but some groups are not included (under-coverage of total population) and there are also some over-coverage issues. Nationally evidence shows that over-coverage tends to be the larger issue, with the Patient Register having 4.3% more people registered than the 2011 Census estimate of population. On a similar note, NHS England's Policy Book for Primary Medical Services says "a practice list can hold 3 to 8% of inaccuracy due to patient turnover alone" (Chapter 9, Part B, paragraph 9.5).

The following groups are not included in the Patient Register, and may be the cause of statistical under-coverage:

- patients solely registered with private GPs
- babies that have yet to be registered at a GP practice
- migrants into the UK who have yet to register
- armed forces (though some remain on GP lists)
- some armed forces dependants
- prisoners (other than those with a sentence under 6 months)
- some prisoners with a short sentence who have received medical assistance in prison
- patients who have been removed through ”no contact” measures
- patients with a temporary NHS number, where no ”permanent” number exists or where their permanent number is not on a GP register

The following groups, for some statistical purposes, may be thought of as over-coverage:

- patients who are no longer resident in the UK (emigrants)
- patients who are staying in the UK for only a short period
- duplicate records

Safeguards used to minimise the risks to data quality

List cleaning

List cleaning is the removal, deletion or suspension of patient records where a record is, or is thought to be, incorrect or where the record is for a patient who should not be recorded on a GP list (for example, armed forces, prisoners, emigrants, duplicates).

List cleaning exercises can be controversial, and can be unpopular with GP practices. Concern is often voiced as protecting patients from removal from lists. The British Medical Association’s (BMA’s) General Practitioners Committee (GPC) has voiced concerns about implementing list maintenance, particularly that described in the guidance referred to below.

National list cleaning

The Audit Commission last ran a “list cleaning” exercise, the National Duplicate Records Initiative (NDRI), in 2009 to 2010. This exercise identified a number of records that were, or were likely, to be incorrectly included on GP lists. These records were then passed to NHAIS areas to action. Differences in approach by NHAIS areas meant that there was geographical variation in the degree of list cleaning undertaken as a result of the initiative.

This exercise identified patient records where:

- the patient is deceased
- records were duplicates
- there was high occupancy implied at an address (addresses with registered occupancy of more than 10 patients were determined to be ”implausible” and flagged for investigation)
- the patient was a failed asylum seeker
- where the age of the patient was implausibly high with no contact.

Previous exercises have also undertaken:

- temporary number matching
- ”gone away” (no contact) matching

However, the Audit Commission noted that: “Reported deductions [resulting from the matches found by the initiative] at individual NHAIS sites ranged from over 13,000 to none. Although variances are to be expected because of the different population demographics at each site and differences in the number of matches released, the range of outcomes indicates that some sites have not followed up the NDRI matches effectively.”

Local list cleaning

As well as the national initiative undertaken by the Audit Commission, local list cleaning exercises have also been undertaken (see National Duplicate Records Initiative 2009/10 pages 22 to 23 for an example).

Guidance and policy on list cleaning

In January 2016, NHS England published the Policy Book for Primary Medical Services which identifies the issue of list cleaning and describes the action that should be undertaken to deal with it.

The Policy Book says (Chapter 9, part B):

“9.4 Some degree of list inflation is inevitable, but manageable if kept within reasonable bounds. Current trends of inflation are excessive and in some regions continue to rise. The Commissioner and PCSS provider is expected to engage in regular proactive list maintenance with general practices.”

This policy sets out operating principles which outline that list maintenance should be designed as a continuous rolling programme. The document goes on to detail elements of a rolling programme. The NHAIS system has a series of checks to reduce list inflation and they include:

- checking patients remain resident where households in multiple occupancy are implied
- check patients registered for 4 years or more at a university address are still resident
- check patients aged over 100 are still resident
- check patients with a previous address abroad remain resident 12 months after registration, and that their address is correct
- update addresses or remove patients where addresses are demolished
- start de-registration processes for patients resident at the same address for 5 years with no contact
- compare registration numbers to our “population figures” at ward or super output areas (SOA) level to prioritise work and target list maintenance

At the time of writing this list maintenance is not being widely implemented.

Data preparation

NHAIS undertake some data preparation to help screen temporary and incomplete records before data supply.

Identification of temporary records

Temporary records are created for NHS patients receiving treatment in an area that is not their usual residence. They can be identified because they are issued with a temporary NHS number. These patients can appear more than once on the NHS Patient Register and so records with a temporary NHS number are removed, by NHAIS, from the register before any further processing is done.

Identification of records with incomplete data

As part of the validation process, records are identified by NHAIS which fail a basic range or validation check. Records that fail basic checks are written to a separate file and are referred to as incomplete records.

Other policy and guidance

There is guidance and policy for GPs to follow in relation to the data they use and in particular the use of the NHS number.

Confirmation of information during interactions

The NHS Number Guidance for GPs says that it is good practice to confirm patient registration details during interactions with the health service. For example, a practice receptionist asking a patient to confirm their address and/or date of birth when booking or attending an appointment. Exact practice will vary.

Confirmation of registration details

Section 12.3 of the policy handbook reinforces the need for finding the correct NHS number on a patient registration. Measures required include:

- leave a registration pending and trace the number instead of allocating a new number
- contact the patient to obtain further details
- National Back Office to conduct monthly checks for duplicate registrations

Changes over time

Some of the key changes that may have had an impact on Patient Register data used for our statistics and research include:

- forthcoming: closure of NHAIS and the assessment of data from the Patient Demographic System (PDS) as a suitable replacement
- July 2016 HSCIC (Health and Social Care Information Centre) becomes NHS Digital
- February 2016 (Central Health Register Inquiry System) turned off (see sections 43 and 44 of Statistics and Registration Service Act 2007), data now sourced from the Patient Demographic System (PDS)
- January 2016 introduction of Policy Book for Primary Medical Service
- March 2013 Primary Care Trusts (PCTs) abolished and duties transferred to NHS England, NHS Wales and Clinical Commissioning Groups (a form of PCSS)
- 2009 National Duplicate Numbers Initiative
- April 2008 NHS Central Register transfers from ONS to NHS Information Centre
- 2005 NHAIS (National Health Application and Infrastructure Services) Patient Number Tracing starts
- 2004 National Duplicate Numbers Initiative
- 2000 to 2004 PCTs (Primary Care Trusts) created
- 1999 National Duplicate Numbers Initiative
- 1999 Primary Care Groups established
- 1997 GP Fundholding abolished
- 1991 NHSCR (NHS Central Register) goes electronic using the Central Health Register Inquiry System (CHRIS)
- July 1948 Foundation of the Office for National Statistics (ONS)
- 1939 Start of compilation of National Registration records and the start of what will become the NHS Central Register

Implications for accuracy and quality of the data

This section summarises the key implications of the operational context on the statistical data quality of “input” data from the Patient Register; that is the data before they are used to create statistics. A number of measures are used to deal with these issues in the production of “output” statistics, these are set out later in this document.

Coverage

The majority of the implications for the statistical quality of the input data relate to coverage. Users of the data need to take into account that the data from the Patient Register may not match the desired statistical population.

The overall impact of the operational context issues is that there is over-coverage, but under-coverage issues can also have an impact.

Under-coverage

Those in this category include:

- prisoners (except most short-term prisoners), armed forces (most) and some dependants, and private practice only patients not covered
- migrants and recent returnees to the UK are not covered until or if they register with a GP (unless they have an extant legacy registration)

This can lead to statistical bias due to under coverage of certain types of people, which will vary geographically.

Over-coverage

The largest source of over-coverage is thought to be those “gone-away”, mostly consisting of patients who now live abroad. Duplicate registrations and delays in processing changes can also lead to over-coverage.

Differences in approaches to list cleaning and also the type of people who end up over-covered, can lead to large geographical differences in the degree of over-coverage. It will also have a differential impact on the characteristics of people over-covered (for example by sex and age group).

Registration lags

As noted above, patients can be slow to update their address following a move. This means that the register will have the wrong location recorded for a number of patients. This can lead to localised, under-coverage and over-coverage, varying by geographical location. Because of the characteristics of those prone to delay re-registering (typically young healthy adults, especially male) this localised coverage difference will vary by age and sex. As much movement of young adults is associated with higher education, there can be substantial effects in student areas and areas of graduate working. However, university policies differ, and the effect in itself will vary – for example some universities are pro-active in ensuring new students re-register to a local GP, but may not be as concerned to encourage re-registration after leaving (graduating) the establishment. As a result, quality of data on students can be variable, and the quality of data of recent graduates can be poor.

Data input errors

The system relies heavily on handwritten completion of forms by patients and manual data-entry by practice staff. Although there are some safeguards in place (as above) data errors are inevitable. As a result, names and addresses will be prone to misspelling. Less often, dates of birth will be prone to some error (for example, swapping of month with day from dd,mm,yy to mm,dd,yy).

Babies

A collection of the above issues will apply to very young babies. Safeguards in place (the confirmation of details on interaction with practice) will rapidly eliminate the majority of these errors, but details for younger babies can be particularly error prone, including:

- errors in incomplete recorded address
- change of address soon after birth – or use of an alternative address (grandparents', father's or mother's address, and so on)
- incorrect recording of sex
- spelling or typo errors in name
- use of name variations – Joe as opposed to Joseph
- use of father's as opposed to mother's surname
- use of name that differs from other records due to parental dispute
- use of "baby boy", for example, where a name has not been decided

The impact of these issues on the quality of our statistics is described in the Producers' QA section below.

Notes for: Operational context and administrative data collection

1. 7,604 practices in England (July 2016) and 455 in Wales (September 2015).

4 . Communication with data supply partners

Interaction with supply partners

Communication with data supply partners is managed by the Data Sharing and Supplier Management (DSSM) team within our Data as a Service Division.

Liaison with the supplier starts in May for the August delivery. Communication is initiated by email, and will include phone calls, and conference calls if required. Over the lead up to the data delivery date the following issues are discussed, in approximate time order:

- whether there are any changes to the data or supply
- identify and agree any changes to the Memorandum of Understanding or Service Level Agreement
- settle costs
- ensure that metadata is up to date
- confirm that the Data Supply Template is up to date
- confirm the logistics of the data exchange

Currently there is considerable interaction with data suppliers as discussions are under way to obtain Patient Register type data from the new IT infrastructure that has been developed by NHS Digital. There are ongoing meetings and workshops and the minutes and actions recorded. As these discussions cover similar issues, much of the background and context is relevant to the existing Patient Register data supply.

Written supply agreement

Supply of Patient Register data is documented in a Service Level Agreement (SLA) with the Health and Social Care Information Centre (HSCIC), now NHS Digital. This SLA in turn forms part of a larger Memorandum of Understanding (MOU) with the Department of Health (DH) and HSCIC.

The SLA is consistent with our SLA template.

In addition, an annual data extract request is agreed and signed.

These documents are managed by DSSM.

Roles and responsibilities

The parties to the agreement are the Office for National Statistics (ONS) and NHS Digital (which was then referred to as NHS IC).

The SLA sets out management arrangements which include:

- SLA Managers, who are identified in the agreement
- Annex Managers, who are identified in the Annex
- the main formal liaison will be through an ONS and NHS Digital Steering Group on National Back Office (NBO) and Medical Research Information Services (MRIS), comprising the SLA Managers and Annex Managers from the NHS Digital and ONS.

Date of the agreement

The SLA was agreed and came into effect in July 2010.

The agreement is valid indefinitely; however, the terms of this agreement, and in particular the annexes, are subject to annual review. Amendments can be made only with the agreement, in writing, of both parties.

The MOU and Patient Record annex was last reviewed in November 2014, at which time annual reviews were suspended to allow investigation of alternative supply from the new IT infrastructure that NHS Digital has developed.

Legal basis for data supply

The data are supplied using the gateway set out in Sections 43 (England) and 44 (Wales) of the Statistics and Registration Services Act 2007.

NHS Digital are authorised to share the data under Section 251 of the National Health Services Act 2006.

Data supply and transfer process

The agreement sets out the technical detail of extracting the data to a password-protected file. This file will be saved to a DVD which will be collected in person by our staff. The password for the file will be separately emailed to relevant staff and will not be made available to the person collecting the DVD.

A copy of the source file will be kept by NHAIS on a dedicated area of the network with access to named staff only.

Security and confidentiality protection

An extract of the data, without name and address, is held by us with standard active directory restriction. Only named users have access to the server and file where the records are held, both before and after the transfer. The records themselves are held as plain text with no encryption.

Although not included in the MOU this data extract is also held in a secure production environment, where access is restricted to named, security cleared personnel.

The full data are held within a specially developed secure research facility, administered by our Data as a Service Division. Access to this facility is limited to named individuals with a demonstrable business need to access the data who have signed data confidentiality agreements and received appropriate training. Further details about this facility are published in Beyond 2011: Safeguarding Data for Research: Our Policy July 2013; this sets out that data that could be used to reveal an individual’s identity are only available to researchers as hashed data. (Hashing is a 1-way encryption process and as such the records are anonymous).

Schedule for data provision

The data are supplied, annually, in August. Data relates to those registered on the NHAIS system on the weekend closest to 31 July of that year.

Content specification

The SLA sets out the data that will be supplied. These are the data consistent with the description in Sections 43 and 44 of the Statistics and Registration Service Act 2007. The agreement sets out that the files will be in CSV format, 1 for each NHAIS area. It sets out the individual variables (fields) and detailed format and content information. This agreement is supplemented by a Data Specification Template that provides more detail on the file format, individual variables and their format. The data comprises name, address, date of birth, NHS number, sex and information relating to their history of registration. The agreement sets out those variables that are excluded from the reduced extract referred to above (fields 13 to 21).

The fields to be supplied, as set out in the agreement are:

Field Number	Field Name
Field 1	NHS Number
Field 2	Gender
Field 3	Date of Birth
Field 4	Postcode
Field 5	Date of Acceptance
Field 6	Site Supplying Data
Field 7	Site of Responsible GP
Field 8	Last Movement Type
Field 9	Previous Cipher/Q Code
Field 10	Patient's Q Code
Field 11	Date Q Code allocated
Field 12	Date patient record last modified
Field 13	Patient Surname
Field 14	1st Forename
Field 15	Other Forenames
Field 16	Address Line 1
Field 17	Address Line 2
Field 18	Address Line 3
Field 19	Address Line 4
Field 20	Address Line 5
Field 21	FP69 Flag

Q code is a reference to the NHAIS area for the patient. An FP69 flag is an indicator of patient inactivity.

Data management arrangements

The SLA says that NHS Digital and ONS will share disaster recovery plans.

The SLA says that data shared will be handled with “appropriate degree of quality” taking into account “legal requirements … ONS and NHS IC guidelines, IM security guidelines and other relevant standards.” Any breach of confidentiality must be reported to Senior Information Risk Officers immediately so that they can agree on any necessary actions.

Data usage

Sections 43 and 44 of the Statistics and Registration Service Act 2007 allow the sharing of this information only for the “production of population statistics”.

Data retention and disposal

The MOU does not set out any data retention or disposal agreements. Like all data, retention of Patient Register data is subject to the Data Protection Act 1998 (Schedule 1, Part 1, Principle 5): “Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes.“

Supplementary information

The SLA additionally sets out arrangements for:

- incident management
- escalation
- performance review
- reporting
- resolution and arbitration

The agreement also sets out quality criteria. The data, when we test it for validity, should have at least 97% of records in each file passing the initial validation phase. The data are validated by checking that no fields are set to NULL values or are incorrectly formatted. In the event that this condition is not met, further investigation will be necessary and NHAIS may be contacted for supporting information.

Change management process

Change management is via discussion between ONS and NHS Digital and subject to agreement in writing.

As part of the regular process of supplier engagement, our DSSM team engage in discussion with NHS Digital. As part of these discussions we ask whether there have been any changes that will affect the data and checks whether the metadata and Data Supply Template remain the same, or need updating.

We are assessing an alternative data source to the Patient Register from the new IT infrastructure developed by NHS Digital. The findings of this assessment will be published in due course.

Engagement with users

Our Population Statistics Division (PSD) continually engages with users to understand how well outputs meet their requirements. PSD’s user engagement activities include formal consultations on proposed changes to outputs, regular communication on plans through a quarterly newsletter, and external events open to all users. In addition, where evaluating changes to methods or sources has required specialist knowledge of local areas, PSD has organised Local Insight Reference Panels to elicit the views of relevant local authorities. From these activities, any issues relating to the sources, and their fitness for the proposed use, will naturally come out. Issues restricted to 1 output will generally be addressed by the team responsible for that output while the Stakeholder Engagement team in PSD takes an overview of any issues with more general implications, and ensures that this is considered in development of outputs across the division. It should be noted that users are more likely to comment on the overall methodology and the effect that it has on the final statistics than on a contributory data source.

Any issues around the quality of the statistics are described in the Quality and Methodology Information report accompanying each output. Issues around specific administrative data sources used in producing the statistics are considered in Quality Assurance of Administrative Data reports such as this.

When changes are proposed to methods (including changes in data sources being used in producing statistics) our Population Methodology and Statistical Infrastructure Division will assess the resultant methods prior to implementation to assure that they are of sufficient statistical quality to meet user needs and are an improvement on the previous method. An independent evaluation by academic experts may also be undertaken, should methodological changes be extensive. The methods are also subject to scrutiny by the UK Statistics Authority as part of the National Statistics accreditation programme under Principle 4 of the Code of Practice for Official Statistics (sound methods and assured quality).

The Responsible Statistician is named for each release and contact details for them are provided, so should someone have concerns over the statistics they are able to communicate them with us. Methodology documents are published to enable users to provide scrutiny.

5 . Quality assurance principles, standards and checks applied by data suppliers

Data suppliers' principles, standards (quality indicators) and quality checks

General approach to data collection and quality

A general principle in the collection of demographic data across the NHS systems is the collection of data should not interfere with the effective operation of the key function of providing healthcare. It is recognised that data are often entered in high-pressure operational environments. Therefore though correct entry and checking of data are encouraged, particularly the use of NHS number, it is accepted that this is not always possible. Therefore, front line checking of data when entering onto the system is kept to a minimum, for example the delay of admission of a patient into hospital because of an information mismatch is undesirable. As a consequence of this underlying approach a certain amount of error and uncertainty within the data needs to be accepted and, where possible, allowed for in the creation of statistics from the administrative data.

Use of NHS number

Much of the quality of Patient Register data revolves around use of the NHS number.

Principles

Each patient should have only 1 NHS number that is unique to the patient.

GP practices are encouraged to:

Find it: Find the NHS number for a person as soon as possible Use it: Use the NHS number to link a person to their record Share it: Share the NHS number with colleagues so they can use it

Practices are encouraged to help make patients aware of the number and suggest a range of ways of doing this.

Quality Checks

The guidance says there are 2 ways to check the NHS number.

Validate: GP practice systems and most NHS IT systems automatically check the format of the NHS number when it is entered. This format check uses an algorithm applied to the first 9 digits to confirm that the tenth digit, the “check digit”, is correct. Validation of NHS numbers substantially reduces the risk that the NHS number may have been recorded against the wrong patient, or that NHS numbers may have been incorrectly entered or incorrectly recorded.

Verify: When a patient makes an appointment a trace is made against the Personal Demographics Service (PDS), the national electronic database of NHS patient demographic details, using the NHS number and another reliable demographic item, such as date of birth, enables accurate identification of a patient record. The NHS number provided to the General Practice when a new patient registers has been verified.

The guidance goes on to suggest sources and usage of the NHS number in a number of different types of interaction with patients.

To avoid the generation of multiple NHS numbers for patients the Annex A of the guidance gives a series of questions to ask of patients to ensure that the patient is new to the NHS. Where a patient is not new to the NHS a new number should not be given. However, a temporary number can be given if a trace cannot be made.

Operation of quality checks

The operation of quality checks will vary from GP practice to GP practice, and potentially depend on the person undertaking the work as well. Quality can therefore vary geographically.

Quality checks before data supply

NHS Digital undertake quality checks before data supply. These include:

Identification of temporary records

Temporary records are created for NHS patients receiving treatment in an area that is not their usual residence. They can be identified because they are issued with a temporary NHS number. These patients can appear more than once on the NHS Patient Register and so records with a temporary NHS number are removed from the register before any further processing is done.

Identification of records with incomplete data

As part of the validation process, records are identified which fail a basic range or validation check. Records that fail basic checks are written to a separate file and are referred to as incomplete records.

Quality reports

NHS Digital publishes data detailing the number of patients registered with GPs, which is from the same source as the data supplied to us. A quality report is published alongside this publication. This report provides some basic quality information grouped around the EU dimensions of quality. The Welsh Government publishes similar data with a report including Key Quality Information.

In addition, the NDRI audit reports have highlighted a number of quality issues.

Audit arrangements

There is no current consistent national audit of GP registrations.

The Policy Book for Primary Medical Services was published in January 2016. This sets out a number of practices that may form the basis of regional oversight; however, these practices are not yet widely implemented. See ‘Safeguards used to minimise the risks to data quality’ and ‘National list cleaning’ sections.

The last national audit, known as the NDRI, was undertaken by the (then) Audit Commission in 2009 to 2010. This provided a thorough audit of GP registration, including a number of recommendations (pp 4 to 5), and the identification of a number of duplicate records. However, it is unclear to what degree the Commission’s recommendations were taken up, and the Commission says in its report that the “range of outcomes indicates that some sites have not followed up the NDRI matches effectively”.

Implications on official statistics

The Patient Register is an effective tool to assist in the provision of healthcare, but for the statistical uses described in this report it has a number of limitations. These limitations are well known to production teams in our Population Statistics Division (PSD) and effective measures have been developed to help deal with these limitations, so that the statistical power of using a high coverage source with generally high accuracy can be realised. Potential future users of the source will need to consider these issues and how to deal with them before additional official statistics based on the Patient Register are produced.

The following sections describe the impacts of each of the main input data quality issues on the official statistics that use Patient Register data. Each issue is taken in turn, with each relevant use addressed within the issue.

Coverage

Internal migration

The internal migration methodology is based mainly on the comparison of 2 snapshots of the Patient Register. By comparing address locations for the record of the same NHS number from year to year, and using only those cases where a change of address within England and Wales is recorded are included. List inflation due to “gone away” cases is therefore not an issue.

No adjustment is made to internal migration estimates for prisoners or armed forces. However, for onwards use within population estimates and projections, separate adjustments are made for these populations.

No adjustment is made for private-only medical patients, but this is thought to have only a small impact on statistical quality.

Local authority distribution of international migrants

International migration is estimated based on data from the International Passenger Survey. Patient Register data are used alongside data from other administrative sources to allocate estimates of international migration to local authorities. The estimates cover civilian migration and therefore coverage relating to armed forces is not required (separate adjustments are made for home and foreign armed forces for onwards use in other statistical outputs). However, the under-coverage associated with private-only patients, and migrants choosing not to register with a GP (together with other issues below), will have an impact on the accuracy of the geographical allocation of international migrants to local areas. This impact is acknowledged within the documentation accompanying the statistics, for example, the Quality and Methodology Information document (QMI) on population estimates says:

“At this level [local authority], individual migration estimates are subject to greater levels of uncertainty. However, the impact of uncertainty associated with net migration flows is small as a percentage of the local authority mid-year estimate.”

Small area population estimates (SAPE)

Separate adjustments are made for armed forces and prisoners, hence these forms of under-coverage are not an issue. The ratio change method for SAPE is designed with the limitations of the source in mind. The method minimises the impact of coverage issues from other sources. As the method constrains to local authority (LA), it is only differential change as a result of under- or over-coverage within an LA that has an impact, especially list cleaning or similar exercises that have a differential impact within an LA. Nevertheless, such issues will have an impact and affect the quality of the statistics.

For output area (OA) based estimates there is potentially a greater (proportional) impact as the absolute level (as opposed to change level) is used, so under- and over-coverage have a direct impact on the estimates. Population for OAs are calculated by taking the number of people on the Patient Register for OAs and using this to determine the share of the population for the super output area of which it is a part. For more information see methodology note on production of small area population estimates.

These impacts will form a part of the overall limitations on accuracy of the statistics. The Quality and Methodology Information document (QMI) on SAPE gives information on the overall accuracy of the statistics produced.

Snapshot

The Patient Register data we use are a snapshot based on a particular date. Intermediate change between snapshots will be missed, for example, where someone moves more than once in a year.

Internal migration estimates

The use of a snapshot approach for internal migration estimates means that where a patient makes more than 1 move in a year (between snapshots) only 1 move would be recorded. To adjust for this, data from the NHS Central Register (NHSCR) are used to uplift the estimates of internal moves. The current NHSCR system data have other issues that impact statistics (recording only moves between NHAIS areas) so a combined method using both sources is required.

Where an entry onto the register (births or immigration) is followed by an internal move, before the snapshot is taken, either the first location and the move is missed, or the move is missed and any subsequent move will reduce the data for the wrong area. An adjustment is made for births to estimate for this effect, although it is approximate as it uses moves of babies aged 1 as a proxy. For other entries (immigration) no adjustment is made.

Similarly, where a move follows a snapshot and precedes an exit from the system (death or emigration) the move is missed and the data may be removed from the wrong area on the exit. No adjustments are applied for this issue.

We are currently researching how the use of data from emerging new data sources within the health system might be better able to address these issues in future.

Local authority distribution of international migrants

The use of the Flag 4s is affected by using snapshots; this is discussed further in the Data item issues section below.

There can be issues where international and internal migration are both involved and these are covered in the Internal migration estimates section above.

Small area population estimates

The methods used are snapshot-based, so tracking intermediate change is not required.

Delays in registration

Alongside coverage issues, delays in registration is 1 of the 2 main issues with using Patient Register data for statistical purposes. The issue is described in the Potential sources of bias and error in the administrative system section above. Several measures are used to help address this issue:

Snapshot date: The reference date for the statistics produced using Patient Register data is 30 June. The snapshot is taken on the weekend closest to 31 July. This allows approximately a month for patients to re-register, capturing “normal” delays in re-registering with GPs.

Use of student data: Delays in registering moves are most common for young healthy adults, particularly males. A lot of the movement of such patients is associated with study in higher education, and students may be particularly prone to leaving registration at their parental address, as other addresses may be regarded as temporary, despite being away from the home for a number of years. The implications of this can vary from establishment (university) to establishment, as policies will differ, and hence have a geographically differential effect. To adjust for this, Patient Register data are linked at the record (patient) level with data from the Higher Education Statistics Agency (HESA). Adjustments are made for internal migration estimates using this linked data. For the allocation of international migration data to local authority areas both student data and Migrant Worker Scan data are used. Our approach produces higher-quality estimates. However, a student adjustment is not applied to SAPE, where this remains a distortive effect.

Adjustment for babies: Babies under 1 year of age can only appear in the Patient Register snapshot at the end of year. Therefore, for internal migration estimates a special adjustment is made using data from NHS Central Register and using data on 1-year-old children as a proxy. Our Population Statistics Division (PSD) are assessing the inclusion of birth data in their methodological redevelopment work on internal migration.

Data item issues

Name

Methodological redevelopment work on internal migration requires the matching of records from different datasets, 1 of the variables in the new matching method is name. The data available has undergone an anonymisation process (hashing) and as such staff cannot see individual actual names. Typo type errors and alternative name forms can have an impact on this matching (for example, can John Smith be matched to Jonathan Smith?). To address this issue we have developed a range of innovative matching techniques to improve the quality of matching, so that this type of error has a much reduced impact. This development is ongoing to further improve match results.

Address

Addresses can be prone to data entry type errors. We pre-process address data using geo-referencing software to assign the Unique Property Reference Number (UPRN) to each record. The UPRN is used to provide the geographical information needed for statistical production (for example, output area, local authority). The software used is capable of handling a degree of error in the address fields and variation in address form. However, the matching of imperfect addresses is in itself prone to a small degree of error, allocating an address to an incorrect UPRN. The impact of errors in an address is therefore minimised; however, there remains a small percentage of addresses that cannot be successfully geo-referenced and a small error rate. The exact level of non-matches varies from year to year.

Place of residence is also needed to help with matching people from 1 data source to another; this is required for the student adjustment used for internal migration.

Sex

No adjustments are made to address error in recording the sex of young babies. The impact of this issue is considered to be very low.

Flag 4s

Flag 4s (added when a patient registers with a previous address outside the UK) are removed from the record after a patient makes an internal move within the UK. This is a special instance of the snapshot issue. The allocation of international migration estimates geographically is only made on using these data where a more relevant suitable alternative is not available, minimising the impact of this issue. Work is currently under way to develop alternative data from the health systems where it may be possible to capture flag 4s before they are removed from the record following a move.

Flag 4s will remain on a record until a move within the UK is registered, thus data are linked longitudinally and only new flag 4s (where the flag 4 did not exist in the previous year’s extract) are used to allocate estimates of international migration.

6 . Producer’s quality assurance investigations and documentation

Producer quality assurance checks

We are moving increasingly to the collect once, use many times approach to data. As part of this approach, production of statistics is distributed across different teams within our organisation, with common processing undertaken once centrally before there are internal data handovers to teams who take on the production of statistics. This section therefore reflects the common processing that is undertaken, with a separate section for common checks undertaken, and then checking specific to the particular outputs or output areas.

Common processing

Process flow

Checks are made at various stages through the processing of the data. We receive 87 source files; these are held in the secure research environment where they are merged and loaded into SAS (a statistical processing software package). Initial processing takes place and the data are subject to geo-referencing. Further checks take place before 2 versions are created. The first version has a number of variables removed. These are identifier variables such as name and full address. These are removed to reduce the sensitivity of the data so it can be used in the secure production environment. This version is used for the production of population statistics. For the second version, identifier variables are hashed (a 1-way encryption process). When the variables have been hashed the data are made available to researchers. This second version is used for research into future population statistics methodologies. These versions are stored in separate secure environments and security arrangements mean that our staff do not have access to more than one environment at any one time.

Validation checks

On loading of the data onto Statistical Research Environment (SRE) systems a sample of records is examined (“eyeballed”), this is a non-random sample of approximately 10 records from each of the start, middle and end of the data file. A manual validation check of these records is undertaken, which includes that the records supplied are in the expected format. This check will also identify, after loading the data into SAS, if columns and records do not correctly align – for example, the forename is consistently read into the forename slot and does not drop into the surname slot.

Throughout processing SAS produces a log. The log is checked for the words Error, Warning or Note. If any of these words are found the cause is investigated. SAS log files are examined (after each stage of code is run). Examination of the log-file will[1] detect the following types of error:

- character (text) data when numeric data are expected
- fields that are truncated (for example, 10 characters in a field where 8 are expected)
- data that are in the incorrect format (for example, dates not in the expected format)
- unexpected characters (non-standard characters within text, beyond normal punctuation, for example, control codes)

We use matchkeys[2] in demographic research when assessing potential methods. When data are received from different data sources a matchkey is created using values that have been taken from specific variables within a record. For example, date of birth may be combined with name and postcode. The matchkeys are anonymised via a hashing process (which cannot be reversed) prior to use by researchers so that they cannot identify individuals from the information they are using. Checks are undertaken on a small non-random selection of records prior to anonymisation to ensure that the matchkey process has worked correctly. We have published further information about matchkeys.

Sense checks

The manual sample check (described above) is mainly used as a sense check. The sense check looks to see that the data look as expected, for example, names look to be normal names, and not other text; surnames generally look like surnames and not first names (and vice-versa); addresses look to be in the right and consistent order; that fields are appropriately populated; that data do not conform to odd patterns (for example, dates are chronological).

Sense (”eyeball”) checks are also undertaken whenever files are split or merged. These checks are mainly there to check that the correct variables are in each split and merged file. They are also used to check that the local unique identifier is on each file, so that a merge can be successfully made to join the data back up.

Population pyramids are generated in reports for each local authority as part of the standard processing. A check is made to ensure that these have been generated correctly.

An individual’s age at mid-year is derived from their date of birth. Spot checks are applied to a small (non-random) number of records to check these have been correctly calculated. This check is undertaken before hashing as date of birth is 1 of the variables that is hashed.

Consistency checks

When the data are first loaded into SAS a note of the number of records is undertaken. The data are received as 87 separate data files (1 for each NHAIS area). These 87 files are loaded into SAS and merged, after which the total number of records is noted. This number is then used at stages throughout the processing to ensure that the number of records continues to “add up”. That is, the total number of records across file splits and after merges sums back to this total number of records. This check is undertaken after every merge or split during processing.

Processing in secure production environment

Data are exported to a secure production environment before further processing for internal migration, local authority distribution of international migrants and small area population estimates (SAPE). These data include date of birth but are otherwise de-identified.

Processing for internal migration

Internal migration makes use of:

- date of birth: used to determine age at 30 June
- sex
- postcode: used to determine local authority
- acceptance date: to help determine if a move has taken place
- NHS number: used to establish links between 1 year and the next

Checks are performed to determine: the level of “missingness” of these variables; that records have valid entries; and that NHS numbers are not duplicated.

Processing for local authority distribution of international migrants

International migration distribution makes use of:

- Flag 4: used to determine if registration has come from outside of the country
- postcode: used to determine local authority
- date of birth: used to determine age at 30 June
- name: as part of the matching process

Checks are taken to ensure that counts (total and local authority) are similar to the previous year for relevant data items and that proportions are of a similar magnitude. Where checks reveal changes to historical patterns, possible explanations are investigated and the supplier is contacted if necessary to provide further clarity.

Processing for small area population estimates (SAPE)

SAPE use:

- date of birth: used to determine age at 30 June
- sex
- postcode: used to determine output area (a geographical building block)

The SAPE team receive a version of the data after it has been assessed by the team carrying out the work on the local authority distribution of international migrants. The data is received outside of a secure environment and as such has been subjected to disclosure control and only includes output area code, age, sex and count information. Any row with missing data is removed from the dataset. The plausibility of data is then checked at an appropriate level of geography (above output area as numbers at output area are small and prone to fluctuation) by comparing with the 2 previous years. Areas of concern are noted, resulting investigations then determine whether action is warranted.

Corrections to data

Postcodes are cleaned to correct for common errors (for example, changing a zero to a letter O where a zero is invalid, for example, P015 to PO15).

Where an address does not match to a postcode, and an alternative postcode can be derived, the alternative postcode is imputed (the original postcode is retained and can be used as an alternative if required).

The data have a standard set of cleaning applied. Cleaning includes the removal of any punctuation, except hyphens, from names and converting all address fields to uppercase. Names are only used in the statistical matching of data. As the treatment of punctuation in names can differ within system practice and between systems, the consistent removal of punctuation improves the matching reliability, so that for example Darcy will match with D’arcy within automated matching.

Data are not de-duplicated in the preparation stage. Any records with duplicate NHS numbers are included in the data, with the addition of a flag to indicate duplication.

Metrics

Missingness

A missingness report is generated as part of the standard processing. This is checked to see that it has correctly run.

Geo-referencing

There are 4 geo-referencing “passes”. A proportion of the dataset is geo-referenced at each pass. These rates are compared from one year to the next and investigative action undertaken if there is a substantive difference in any of the rates. The overall geo-referencing rate for the 2015 Patient Register is 92%, the same as the previous year.

When running the geo-referencing on 2014 data a drop to 92% from 96% for 2013 data was recorded. Following investigative action it was determined that the drop was as a result of changes in the underlying coding of the geo-referencing, where the suppliers code no longer referenced to a chosen most likely address in some ambiguous matched-cases, but instead produced a failed match. The data were double processed using a previous version of the software and reference files, to allow internal users to assess the impact of the change.

Comparison with other data sources

Security requirements mean that data with identifiers (name, address, date of birth) cannot be compared with other sources, or with older (or newer) versions of the same data, until the data have been hashed. A further security measure of separation of duties means that the team undertaking the pre-processing are not able to undertake analysis on the data once they have been hashed. Analysis is undertaken by separate teams. This substantially limits the degree of comparison possible.

Metrics from one year to another can be compared, such as geo-referencing rates as above. Other comparisons need to take place where only hashed (that is, “encrypted”) data are available.

A comparison with data year on year is made as part of internal acceptance testing by the internal analysis team.

We publish, annually, a tool that enables the comparison of population estimates with a range of sources including the Patient Register. This allows comparison of Patient Register data with the annual mid-year population estimates, births and retirement-age data.

Distortive effects of performance measurements and targets

Over-coverage: See ‘Issues in design and definition of targets’ above for reasons why over-coverage may occur. The GP register has exceeded the estimated size of the England and Wales population every year since 1961; and in 2011 the register exceeded the size of the census-based population estimate by 4.3% (see Comparison of Previous and Improved Methods of Estimating International Immigration at Local Authority Level).

Differential coverage by geography: The patterns and demographics of population movement combined with a tendency of patients, particularly certain types of patients, to delay (re)registration. This has the potential to combine with the list cleaning issues. The combined effect is that list inflation can vary appreciably by geography. In places the effect can be dramatic with the census-based population in around 3% of local authority areas differing by over 10% from the size of the Patient Register in 2011 (see Comparison of Previous and Improved Methods of Estimating International Immigration at Local Authority Level).

Differential coverage by age and sex: Like for geography, the combined effects of factors that influence (re) registration have an impact on coverage by age and sex. In 2011, most of the list inflation is for those aged 20 to 64, and males appear to be over-represented at ages 27 to 68.

Impacts for producing statistics

The NHS Patient Register provides a broad coverage source within England and Wales and covers a large proportion of the population, with a good degree of accuracy for statistical purposes.

However, when used in the production of population statistics the register has a number of issues around selective, variable and differential under- and over-coverage of the population that need to be taken into account in the production of statistics based on the Patient Register.

Patient Register data are a snapshot and this needs to be taken into account when producing change measures.

Name and address information are subject to error, such as transcription, interpretation of handwriting, typos and spelling mistakes. This will impact on matching methods in the future, and appropriate techniques will be adopted to help address the issue.

Risk

The Patient Register, in common with many administrative sources, is subject to change beyond the control of statistical producers that can occur as a result of changes to underpinning rules, classifications, administrative definitions, systems, policy, interpretation and other external effects.

The National Health Applications and Infrastructure Services (NHAIS) is due to close. We are evaluating the Patient Demographic System (PDS) as a replacement for this data and assessing what the resultant impact on statistics might be.

The team that undertakes the pre-processing of the data are in the process of expanding the range of quality assurance that is undertaken, and embedding quality into the coding and procedures used to pre-process the data. Nevertheless the need for improved quality assurance is recognised, and there is a risk that errors can occur. This element of risk has been reduced and will be further reduced. It is unlikely that a major issue will escape detection, although there remains a possibility that an unknown and undetected issue flows through into the data that may have a smaller impact on published statistics.

Quality issues

This section provides a summary of the issues with Patient Register data that can have an impact on the quality of our statistics. In many cases we have developed methods to help mitigate these issues, to make the most out of this powerful data source.

Patient Register Quality issues

Issue	Description	Impact	Mitigation
Missed moves	A range of the quality issues below, including coverage, timelags and snapshot issues, can result in missed or delayed moves.	Where data on a move is missed (or delayed) the internal migration data will not record the move. This in turn means that population estimates and projections will show one more person in one area and one less in another. These issues will sum to zero nationally, but have local impact.	Mitigation is not applied for this issue.
Under-coverage: prisoners	This issue is adjusted for in population estimates, so impact is limited.	Internal migration statistics are defined as excluding prisoners. Population estimates (including small area estimates) have separate adjustments for prisoners.	This issue is adjusted for in population estimates, so impact is limited.
Under-coverage: armed forces	Armed forces and some dependants are not covered in the Patient Register.	This issue is adjusted for in population estimates, so impact is limited.	Migration statistics are defined as civilian migration. Population estimates (including small area estimates) have adjustments for armed forces.
Under-coverage: private patients	The Patient Register records only those registered with NHS GPs	Moves missed of patients registered only with private practices. Those migrating into the UK who register only with private practice are missed, as such there is potential for bias in areas with pockets of great wealth.	Limited – private-only patients will only impact on the number of international migrants if they are captured by the International Passenger Survey. The methodology uses additional data sources, but these may not contain private patients and as a result there may be some bias in the local authority distribution of migrants.
Under-coverage: babies	Babies can be poorly covered in the Patient Register data until they are registered at a local GP practice and their data “settles down”.	The existence and/or basic demographics for babies can mean that they are not matched with data from other sources. Changes of address for babies under 1 year old will not be captured because of the snapshot issue.	A special adjustment is made to estimate moves of babies aged under 1 year old during internal migration processing.
Under-coverage: migrants into the UK who have yet to register	Data for new or returning entrants to the UK will only be captured when they register at a GP which may be after a gap.	This issue can cause inaccuracy in allocating international migrants to local geography and may mean internal moves are missed.	Patient Register data are used in combination with other data to estimate the geographical distribution of migrants.
Under-coverage: “no contact” removals	If there has been no contact with a patient for some time, after attempts to make contact, they can be removed from the Patient Register.	Some patients will be removed from the register, even though they remain resident. This can mean that they are not available for matching, and also some internal moves will be missed (if they re-register at a later date at a new residence).	Mitigation is not applied for this issue.

Over-coverage: duplicate registrations	Duplicate registrations can occur where a) 2 records for the same person are in the dataset with the same NHS number; b) where 2 people are recorded with the same number; or c) the same person is recorded with 2 different NHS numbers.	Duplicate registrations can inflate the register.	Where there are duplicate numbers these are flagged on the dataset. Internal migration, local authority distribution of international migrants, and SAPE use data with same-number duplicate records removed, other duplicates are not corrected for.
Over-coverage: emigrants	Emigrants who leave the UK but do not de-register will remain on the register.	The register can give an inflated picture of the size of the population in an area.	Mitigation is not applied for this issue.
Over-coverage: short stays	Registration on the Patient Register uses a different residence definition to the statistical definition.	The Patient Register can provide an inflated estimate of the number of usual residents. The estimated number of international migrants can be distributed to local level on the basis of inclusion of short stay patients.	Partial – Patient Register data are used in combination with other data to estimate the geographical distribution of migrants. This distribution has no impact on the total number of international migrants, which is estimated separately.
Over-coverage: armed forces	Some armed forces personnel are captured on the Patient Register	Adjustments and definitions are used, based on the approach that armed forces personnel are excluded from the Patient Register data, however some armed forces personnel are included, meaning the definition is not perfectly applied, and potentially double counting in adjustments.	Mitigation is not applied for this issue.
Over-coverage: prisoners	Some prisoners (with a sentence over 6 months) are captured on the register where they have not interacted with prison medical services	Adjustments and definitions are used, based on the approach that prisoners (over 6-month sentence) are excluded from the Patient Register data, however some prisoners are included, meaning the definition is not perfectly applied, and potentially double counting in adjustments.	Mitigation is not applied for this issue.
Shared custody	Shared custody can cause a range of issues.	Shared custody has the potential to cause duplicate registrations, use of different names from other sources, and missed internal moves.	Mitigation is not applied for this issue.
Time lags in (re) registration	Patients can be slow to re-register a change of address and GP practices will process data to different timescales. Young adults may also choose to remain registered at their parental address.	Patient Register data can show lags in address details, the impact of this is different from place to place, by age and sex, and has a particular impact in areas of high concentrations of students. This can affect the timing and age distribution of internal migration moves.	The snapshot used for statistics is taken approximately 1 month after the reference date to allow for some delay in re-registration. Data on students from the Higher Education Statistics Authority is used to supplement Patient Register data in the creation of internal migration estimates.

Care homes (and similar)	A move of a patient into a care home may not result in a re-registration, but a subsequent death recorded with a residence address at the care home. The reverse can also occur.	This can result in error in the statistics with deaths resulting in numbers removed from the wrong location.	Mitigation is not applied for this issue.
Data transcription error	Data are manually transcribed from handwritten forms and subject to transcription errors.	Some data, particularly names and addresses, may contain errors.	The geo-referencing software has been thoroughly tested on Patient Register data and is able to deal with a degree of transcription error.
Flag 4 operation does not exactly match the statistical concept	Flag 4s are used to indicate a previous address abroad. They are removed following an internal move, and may also be used after a residence abroad shorter than the required statistical definition.	Flag 4s do not match exactly the statistical concept for identifying a migrant. They may be used as an indicator, particularly of distribution, but total figures need to be used with care.	Flag 4s are combined with other data to distribute international migrants and are not used to determine totals.
Default date of birth	Where a date of birth is unknown a default date of birth may be used, typically 1 January. 1 January 1900 or 1901 may be used where an age is not known.	Dates of birth of 1 January should be considered with caution and ages relating to 1900 or 1901 are likely to be erroneous.	A separate approach is used to estimate the very elderly population. As the year of birth used (except 1900 or 1901) is likely to be approximately correct, and the number of cases is small, the degree of error in age data is small enough to be disregarded.
Snapshot	The Patient Register data are a snapshot in time, which does not track change.	Where a patient moves more than once during a year only the addresses at the start and end of year are captured – underestimating the number of moves.	Internal migration estimates are calculated with this issue in mind and alternative data are used to help make adjustments for this issue.
		Flag 4s are removed after an internal move (see Flag 4s above)	See above for Flag 4s.
Infrequent audit	The last audit was undertaken in 2009	Infrequent audit may have an impact on the quality of the data and that the over-coverage issues raised above will build.	Measures are built into the statistical methodology to help address over-coverage issues, as described above.
List cleaning	Differential approaches are taken to list cleaning	List inflation can be higher in some areas, and at some times, than at others.	Measures are built into the statistical methodology to help address over-coverage issues, as described above. Internal migration data uses data on moves, and not total list size, so differential list cleaning has a limited impact. However, if a patient is removed from a list, so they are missing at a mid-year point, and subsequently re-registers at a new address, this move is missed.

Disclosure control	Disclosure control techniques are applied to the data before publication	Aggregate data have a very small error due to disclosure control, however the level of error is not large enough to affect statistical utility.	Only the level of disclosure control required to create sufficient uncertainty is used and the technique applied is the most suitable for the methodology. The statistical disclosure control technique is reviewed as part of methods development.

Conclusion

The NHS Patient Register provides a broad coverage source for the England and Wales population, with a good degree of accuracy for statistical purposes.

The source has a number of issues that need to be taken into account in the production of statistics, these include:

- some armed forces and dependants, prisoners and private-only patients are not included on the register
- patients often do not register when they leave the UK, which means the register contains patients who are no longer resident
- patients can be slow to (re)register following a change of address (particularly relating to students and graduates)
- measures to keep lists up to date will vary from GP practice to GP practice

These issues vary by the main demographic variables of age, sex and location. Where possible, we take these issues into account when using this administrative data source as part of the methodologies used for population statistics either in production or when formulating new methods.

The Patient Register is combined with other data using appropriate methods in order to produce reliable statistics relating to population.

For statistical purposes, the register has a number of issues around selective, variable and differential under- and over-coverage of the population that need to be taken into account in the production of statistics.

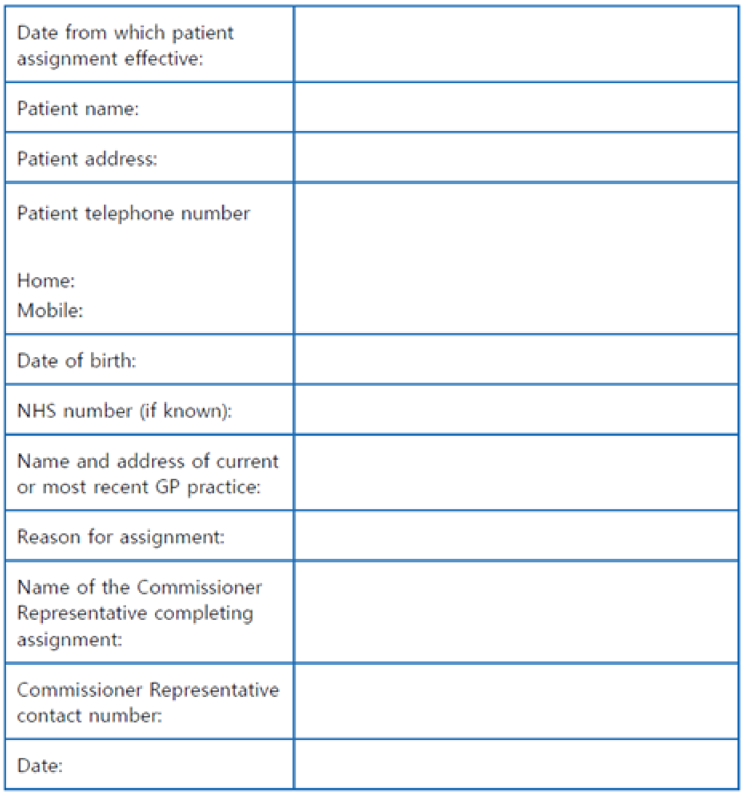
Notes for: Producer’s quality assurance investigations and documentation

1. This is a standard process for SRE datasets and all of the error types listed have been identified at different times across the range of datasets that are acquired for the SRE.
2. Matchkeys are alphanumeric strings generated within processing from other data held on each record. For work on Patient Register data they are based on combinations of important demographic data (such as name, date of birth and sex). They are generated to enable automatic matching of individuals across different data sources.

7. Annex A

Form for assignment of a patient to a practice

Form used for registration of patient.



Date from which patient assignment effective:	
Patient name:	
Patient address:	
Patient telephone number Home: Mobile:	
Date of birth:	
NHS number (if known):	
Name and address of current or most recent GP practice:	
Reason for assignment:	
Name of the Commissioner Representative completing assignment:	
Commissioner Representative contact number:	
Date:	