Office for National Statistics

Article

Administrative sources used to develop the Statistical Population Dataset for England and Wales: 2016 to 2021

Quality overview of the administrative data sources used to develop the Statistical Population Dataset (SPD) version 4.0.

Contact:
Neus Beascoechea Segui, Hannah Howard, Katt Skippon, Ireoluwaposi Onadeko, Christos Chatzoglou, Aysha Khan and Joseph Herson
pop.stats@ons.gov.uk
+44 3000 682506

Release date:
3 March 2023

Next release:
To be announced

Correction

10 November 2023 12:01

We have made a number of updates and minor corrections to the Quantitative Quality Indicators dataset which are recorded in the 'Correction_notice' worksheet within the dataset.

Table of contents

1 . Main Points

- We assessed the quality of each of the administrative sources used to develop the Statistical Population Dataset (SPD) version 4.0 against the five European Statistical System (ESS) quality dimensions.
- This article explores how each of the sources are used to develop the SPD to demonstrate transparency and that suitable data sources have been quality assured.
- We explain the use of our Quantitative Quality Indicators (QQIs) and grading system to provide a comprehensive and comparable assessment of quality for the administrative sources used over time.
- Our assessment shows that the administrative sources fit the five ESS dimensions well overall and demonstrated consistency over time.

2 . Overview of transformation

This research forms part of our population and social statistics transformation programme, which aims to provide the best insights on population, migration and society using a range of administrative data sources. The findings will form part of the evidence base for the National Statistician's Recommendation in 2023 (PDF, 249KB) on the future of population, migration and social statistics in England and Wales.

Administrative data have been acquired in line with the Office for National Statistics' (ONS's) data acquisition policy and data ethics policy to assist with the ONS's research into transforming census and data collection practices. Further information can be found in Section 14: Ethics and security.

3 . Assessing the quality of administrative data

The quality of a data source can be defined as its "fitness for purpose". We focused on data sources used to develop our Statistical Population Datasets (SPDs) that feeds into our Dynamic Population Model (DPM). The administrative data listed also include relevant variables that are required to integrate with other sources in the construction of the Statistical Population Dataset (SPD) and to produce administrative-based population statistics on the size and composition (by age and sex) of the usually resident population for England and Wales.

This includes:

- Personal Demographics Service (PDS)
- Higher Education Statistics Agency student data (HESA)
- English and Welsh School Census (ESC and WSC)
- Individualised Learner Record (ILR)
- Hospital Episode Statistics and Emergency Care Dataset (HES and ECDS)
- Death Registrations
- Benefit and Income Datasets (BIDs)
- Customer Information System (CIS)

We assessed these sources using Quality indicators for phase 1 errors (PDF, 805KB) of the Statistics New Zealand Error Framework and the source stage of the Guidelines for assessing the quality of administrative sources for use in censuses by the United Nations Economic Commission for Europe (UNECE) . This allowed us to evaluate each source against the purpose for which the data were collected by the supplier.

We assessed the quality of each source and provided a grading of the Quantitative Quality Indicators (QQI). These are numerical values that provide evidence for assessing the quality of administrative data received by the Office for National Statistics (ONS). The indicators used in the QQI are mainly derived from guidance from Statistics New Zealand (Phase 1) and our Quality errors framework.

We assessed the quality of each source and compared them with the European Statistical System (PDF, 457KB) (ESS) quality dimensions:

- relevance
- accuracy and reliability
- timeliness and punctuality
- accessibility and clarity
- coherence and comparability

We adapted the ESS quality dimensions to meet the purpose of this article in assessing the quality of each source against their original and intended purpose, rather than when combined to create an output. Considering this, we adapted the relevance dimension to include both the purpose of the source as intended by the supplier, and the coverage.

4 . User needs

The Statistical Population Dataset (SPD), used in combination with other sources as part of the dynamic population model (DPM), will enable the Office for National Statistics (ONS) to meet user needs for providing coherent statistics on the size of the population and changes over time (both nationally and locally).

Users told us that they need these statistics in a frequent and timely manner to enable evidence-based decision making. Our statistics need to be relevant in a rapidly changing society, and robust in their quality. This article covers the underlying sources used to develop the SPD to demonstrate transparency, which is one of the cross-cutting themes of the Code of Practice for Statistics, and that suitable data sources have been quality assured.

5 . Quantitative Quality Indicators (QQI)

Quantitative Quality Indicators (QQI) have been produced for the sources used to develop the Statistical Population Dataset (SPD) version 4.0 for 2016 to 2021. The QQI are conducted on each source individually. We developed a grading system to simplify the QQI outputs into an interpretable and comparable assessment of quality.

The QQI provide evidence for the following quality indicators:

- lag from collection to supply
- coverage
- replicates
- duplicates
- item non-response
- implausible responses
- linkage to the Demographic Index (DI) (as explained in Understanding quality of linked administrative data sources in England and Wales, using the 2021 Census - Demographic Index linkage)

Definitions of the quality indicators, along with their grading are included in the QQI Table. Please note, QQI have not been produced for Benefit and Income Datasets (BIDs) and Customer Information System (CIS) data as they were newer additions to the research.

6 . Personal Demographics Service (PDS) quality assessment

Relevance

The Personal Demographics Service (PDS) contains demographic data for those who have received an NHS Number after interacting with an NHS Service in England, Wales and the Isle of Man. Data are collected from all NHS care providers, including GP practices. The system enables healthcare professionals to bring together patient information quickly and accurately. Therefore, the PDS provides excellent coverage of the resident population of England and Wales, through the interactions they have with NHS services. There is some evidence of undercoverage, as seen in our Administrative data used in Census 2021, England and Wales methodology, for those having private health care and new migrants.

The PDS shows some overcoverage, particularly for more urban local authorities (LA) where there is a higher rate of people moving in and out of the LA. For example, those visiting a LA may temporarily interact with an NHS service and then move out of the area but remain on the PDS (for example, short-term migrants). Similarly, they may not notify their GP when they move until they need to see a doctor again, this can cause overcoverage in one area and undercoverage in another area. We are working with NHS England to better understand overcoverage so that we can address it from both perspectives.

Accuracy and reliability

Data for the PDS are collected through interactions with NHS Services via forms completed online, for example the GMS1 family doctor services registration form, or on paper under the direction of departmental staff (usually NHS staff). Different collection methods, particularly paper forms can lead to an increase in data entry error, therefore reducing the accuracy.

NHS England is developing a new GP registration service that provides all GP practices with an integrated online option for patients. This allows online updates to contact details to improve the accuracy of the PDS and reduce third-party input errors. NHS England also make bulk changes to systematically correct contact and address details. Validation, quality assurance processes and resolution of data quality incidents are performed by the PDS National Back Office (NBO) weekly.

Timeliness and punctuality

An extract is supplied to the Office for National Statistics (ONS) annually between August and September. Since the stock extract references the previous year backwards from 30 July, there is a lag of one to two months. Other updates to data are received by the ONS weekly, ingested monthly and processed quarterly.

Accessibility and clarity

Extracts are supplied by NHS England alongside a data dictionary with information about updates to variables. Processing summary notes are provided if, for example, variables have been derived. Data are supplied with standardised formatting, such as expected variable type and length, aiding accessibility and clarity. These help users to understand the quality of the variables within the dataset. The ONS meets regularly with data experts from NHS England who advise on the quality of the data.

Comparability and coherence

Every NHS number is present on the PDS system and information is collected when there is an interaction with NHS services, it is also not mandatory to supply information for all variables. Therefore, there will be variation in the completeness of data per record affecting comparability and coherence across years.

NHS England provides guidance for patients and NHS services on collecting PDS data. For example, a user guide for completing the family doctor services registration form and support for practices adopting the new GP surgery service. There was an observed decrease in GP registrations and address changes from April 2020 (the start of the coronavirus (COVID-19) pandemic) followed by an increase in the first half of 2021, rising above pre-coronavirus pandemic levels. This may be because of a backlog of people returning to the GP after lockdown, updating their details because of the vaccination programme, or people moving house and changing their address as the property market reopened (in May 2020 in England and in June 2020 in Wales). This could affect the comparability of 2020 and 2021 data against other years.

The Quantitative Quality Indicators (QQI) for this source are provided on Worksheet 1 of QQI table.

7 . Hospital Episode Statistics (HES) and Emergency Care Dataset (ECDS) quality assessment

Relevance

The NHS's Hospital and Episode Statistics (HES) records details of all attendances, appointments and admissions to NHS Hospitals in England. The source allows hospitals to be paid for the care they provide and is used for research and planning health services.

Between 2016 and 2019 HES consisted of three datasets: Accident and Emergency (A&E), Outpatient (OP), and Admitted Patient Care (APC). From 2020, HES only included OP and APC.

The Emergency Care Dataset (ECDS) is the national dataset for urgent and emergency care, replacing Accident and Emergency (A&E) from 2020, as described in the HES processing cycle and data quality checks. ECDS collects information from all hospitals with A&E departments and includes information on attendance. HES and ECDS are therefore relevant for population statistics, providing information about the general population present in England.

HES and ECDS may include records for those who died or emigrated after interacting with the datasets, creating overcoverage. There is potential undercoverage in HES and ECDS for parts of the population that rarely interact with healthcare services, for example young adult males and migrants, or those that frequently record different addresses, for example students.

Accuracy and reliability

Under section 45c of the Statistical and Registration Service Act 2007, all NHS hospitals are legally required to complete HES and ECDS in England.

Data are recorded by healthcare professionals during a patient's interaction with an NHS hospital. Therefore, incorrect values for certain data items may be recorded, particularly during emergency care situations where it may not be easy to gather accurate information. There can also be differences between the data recorded on HES and ECDS compared with the Personal Demographics Service (PDS).

The data are initially collected as part of the Commissioning Dataset (CDS) and are submitted to NHS England who process and return it to the healthcare provider as the Secondary Uses Service (SUS). The data are subject to HES validation checks and ECDS validation checks by NHS England before being supplied to the Office for National Statistics (ONS), increasing the accuracy and reliability of the data.

Timeliness and punctuality

NHS England supplies the ONS with HES and ECDS data monthly and annually. The reference period for the annual business-as-usual supply of HES and ECDS data begins in April. The datasets are then delivered to the ONS in October. This means there is an 18-month lag between the start of data collection for the annual supply and the data being available in the ONS. The potential lag between data collection and supply is offset by the monthly supply, so the data available for analysis are timelier.

Accessibility and clarity

HES and ECDS data are supplied by NHS England, also providing metadata and a HES data dictionary. Supplier provided metadata and data dictionaries improve the clarity of the data, as variables are clearly defined.

Comparability and coherence

Data are collected by hospital staff during a patient's interaction with an NHS hospital. Guidance for using Systematized Nomenclature of Medicine Clinical Terms codes (SNOMED), the standard clinical terminology for the NHS to support recording of clinical information, is provided by NHS England. The use of standardised codes ensures that data can be easily compared across years.

NHS England cleans, standardises and validates data they receive, which improves coherence and ensures it is easily comparable.

In April 2021, the patient identifier variable was changed from Hospital Episode Statistics Identifier (HESID) to Master Person Service (MPS) identifier as part of a wider strategy to move to a common patient identifier across all national patient level datasets, such as the Personal Demographics Service (PDS). While this improves comparability across patient level datasets, it may affect comparability across years within HES.

The coronavirus (COVID-19) pandemic affected the coherence of HES and ECDS data. Read about the impact of COVID-19 on:

- Hospital Outpatient Activity 2020 to 2021
- Hospital Admitted Patient Care Activity 2020 to 2021
- Hospital Accident and Emergency Activity 2020 to 2021

The Quantitative Quality Indicators (QQI) for this source are provided on Worksheet 4 of QQI table.

8 . Death Registrations quality assessment

Relevance

The purpose of the data is to register all deaths occurring in England and Wales. This helps inform policy decisions, monitor the population's health and measure progress against health-related goals. Therefore, Death Registrations provide information about those to remove from our estimates of the usually resident population in England and Wales.

Deaths that occur outside of England and Wales are excluded from the data, this may include usual residents of England and Wales, such as those serving in the armed forces abroad.

Deaths of individuals whose usual residence is outside of England and Wales but whose death occurs in England and Wales are included in total figures but are excluded from all smaller geographies. Read more about the coverage of death registrations in our Mortality Statistics in England and Wales QMI.

Accuracy and reliability

Under the Births and Deaths Registration Act 1874, it is a legal requirement for all deaths in England and Wales to be registered within five days.

Data items for each record largely depend on information supplied by the informant (usually a close relative), a police officer, or a witness in the case of deaths certified after an inquest. The cause of death is recorded by a Medical Certificate of Cause of Death (MCCD) that is completed by a doctor or a coroner. Under the Coroners and Justice Act 2009, guidance is given to doctors completing MCCDs. All variables are subject to validation and quality assurance checks throughout the data journey from the General Register Office (GRO) to the Office for National Statistics (ONS). Read more about the process of the registration of deaths and the quality of mortality data in our User guide to mortality statistics.

Timeliness and punctuality

There are situations where the registration of a death will be delayed beyond the five days, for example, when a death is referred to a coroner for further investigation. Read more on the impact of registration delays on mortality statistics in our Impact of registration delays on mortality statistics in England and Wales: 2020 article.

Data are supplied to the ONS daily when a registration is made through the Registration Online (RON) system. Figures based on final data are supplied by the ONS annually in January. The annual supply has a lag of approximately seven months from the reference year.

Accessibility and clarity

A data dictionary for death registrations contains metadata about records and variables and how these are used for validation. It also contains information about transformation and derivation of variables.

Comparability and coherence

Records are coded to the World Health Organisation (WHO) International Classification of Diseases-10 for cause of death and the Standard Occupation Classification (SOC) for occupation, allowing for international comparisons.

To allow for differences in the age structure of populations age-standardised mortality rates (ASMRs) are used, and therefore valid comparisons can be made among geographic areas, over time and between sexes.

The Quantitative Quality Indicators (QQI) for this source are provided on Worksheet 8 of QQI table.

9 . Higher Education Statistics Agency student data (HESA) quality assessment

Relevance

The Higher Education Statistics Agency (HESA) collects, processes, and publishes data about students enrolled at publicly funded Higher Education (HE) providers in the UK. It is optional for privately funded institutions to submit data to HESA. Therefore, HESA provides excellent coverage of those in public Higher Education, including international students residing in England and Wales.

The following instances are excluded from the dataset at an individual level:

- students studying their entire course outside of the UK
- students not funded for study by distance learning overseas

There are a small number of distance learning students studying outside the UK who are funded (for example, Crown servants overseas); these are included in the HESA student record.

There may be overcoverage as students who drop out may still be included in the dataset. Please see the HESA sampling frame for more information.

Accuracy and reliability

Data are collected by HE institutions through enrolment (made online or by post), the main facilitator of this is University and Colleges Admissions Service (UCAS) applications. A verification service run by UCAS (PDF, 199KB) flags applications for potentially fraudulent activity, missing and misleading information or potential duplicates, for further investigation.

Data supplied to HESA are subject to an extensive quality assurance process with a range of automated HESA validation checks that are run against all submissions.

Address data (postcode) are collected at the start of a student's period of study and may not be updated again. Because of the coronavirus (COVID-19) pandemic, guidance was issued by HESA on how HE providers should collect the required student data. The coronavirus (COVID-19) pandemic increased the likelihood that a student's term time postcode reflected where they intended to reside, rather than where they were actually residing potentially reducing the accuracy of data about their location.

Timeliness and punctuality

Data are collected throughout the academic year (1 August to 31 July) and submitted to HESA in October after the academic year ends. Data are supplied to the Office for National Statistics (ONS) in December of the following year, creating a 17-month lag.

HESA's new Data Futures programme intends to bring the requirements of two existing data collections (Student record and Student Alternative record) from one annual collection to in-year submission, greatly increasing the timeliness of HESA data from the 2024 to 2025 academic year.

Accessibility and clarity

One standardised supply of data is received from HESA to the ONS and there are detailed metadata available online, which improves the accessibility and clarity of the data. The ONS meet regularly with data experts from HESA who advise on the quality of the data.

Comparability and coherence

HESA collect data from several HE providers and provide data collection resources to support the data collection process, which ensures data are coherent. It is mandatory for HE providers to report data to HE funding and regulatory bodies.

The Quantitative Quality Indicators (QQI) for this source are provided on Worksheet 6 of QQI table.

10 . English School Census (ESC) and Welsh School Census (WSC) quality assessment

Relevance

The English School Census (ESC) and Welsh School Census (WSC) collect demographic information on all pupils attending state schools in England and Wales, respectively. Read about types of school in Englandand types of school in Wales.

The data are used to allocate resource and funding, and to assess changes in education policy, so there is an incentive from schools to ensure the accuracy of the data.

The ESC and WSC provide excellent coverage of the population aged 5 to 15 years in England and Wales.

Pupils who are not educated in a state school are excluded from the school censuses. Pupils attending Ministry of Defence schools based overseas are removed from the Office for National Statistics (ONS) supply because their usual residence is outside of England and Wales. Pupils in pupil referral units (PRUs) are not included in the WSC unless the pupil is also registered at a state school.

Pupils who are permanently excluded from school prior to school census day will not be present on the dataset, creating potential undercoverage. However, those pupils excluded on school census day will be present. Further undercoverage comes from children of recent migrants who may not immediately attend school, despite intending to be resident long term. Children of short-term migrants who are attending school will appear as active on the dataset when they may only be attending for a few months, potentially resulting in overcoverage.

Accuracy and reliability

Under Section 537a of the Education Act 1996, all state schools are legally required to complete the English or Welsh School Census.

Given that each school manages their collection of data independently, there may be some geographical variation in terms of accuracy and reliability. For example, the frequency that updates are requested from parents or guardians will vary and so information for some schools will be more up to date, and therefore more accurate than others.

Data are submitted to Department for Education (DfE) via the COLLECT system for England. The system runs automatic validation checks on the data. Schools are required to amend or provide suitable explanations for all errors. Guidance for completing this process within the validation period is provided, helping to further improve the accuracy of the data.

For Wales, data are collected via the school's management information system (MIS), which returns errors and query reports. This includes checks for unexpected characters or logical inconsistencies. Schools are required to resolve all errors and queries. However, variables in the data might still contain oddities that could affect the accuracy of the data. For example, incorrect postcodes, where schools do not know the postcode of a pupil, they are advised to use the postcode of the school and to make sure they have documented it as such.

Read more about accuracy of the data in the Welsh Government services Schools' census results: February 2022 . The headteacher must authorise the return before it is sent to the local authority via the Data Exchange Wales Initiative (DEWI).

For both England and Wales, a summary report with comparisons with the previous year must be checked for accuracy and completeness. This report is approved by the headteacher, and sense checked by the local authority who are responsible for resolving duplications. Validation checks conducted by the supplier increase the accuracy and reliability of the data.

There may be a lag present for the migrant population. This includes attendance lag because pupils may show up later in the year, or changes of details lag as the migrant pupil moves between local authorities.

Timeliness and punctuality

For England, data are collected by schools each term (October, January, and May). The spring term (January) subset of the data are supplied to the ONS annually between May and June of the same year, creating a lag of four to five months.

For Wales, data are collected by schools throughout the year and submitted on School Census Day in January. Data are supplied to the ONS annually between June and July of the same year, creating a lag of six to seven months.

Accessibility and clarity

Data are supplied by the Department for Education (DfE) for England, and the Welsh Government School Statistics Team. See the metadata for both ESC and WSC. The accessibility and clarity of the data are affected because of limited metadata currently being available for the ESC; however, the ONS is working with the DfE to further our understanding of the source and the variables within.

For WSC, there is more than one row per pupil for the Special Education Needs (SEN) table as a pupil may have more than one SEN type. However, pupils are allocated a unique pupil number (UPN), which remains on the school's MIS and is present on the dataset supplied.

Comparability and coherence

ESC uses standardised codes for missing values (not applicable), which ensures that data can be easily compared across years.

The number of variables in the ESC dataset differs between years, for example there are 13 more variables in 2016 compared with 2017 to 2019. This limits comparability between certain variables across years.

Patterns across all checks have been consistent from 2016 and 2017 to 2021, with no noticeable coronavirus (COVID-19) impact on the ESC dataset, aiding comparability and coherence across years.

For WSC, multiple collection methods are used and may vary among schools, for example data may be collected via paper form or online. However, all information is mandatory so it can easily be compared over time. The same Pupil level annual school census (PLASC) guidance is received by all schools to complete the WSC.

The reference period switch from January to April in 2021 to 2022 data changed the pattern of steady increases for WSC, however differences across years have otherwise been reasonably consistent.

For WSC, there have been annual changes in variables, for example in 2019 all care-related items were removed. Some other known quality issues in the WSC dataset include duplicate records such as names appearing twice.

The Quantitative Quality Indicators (QQI) for this source are provided on Worksheets 2 and 3 of QQI table.

11 . Individualised Learner Record (ILR) quality assessment

Relevance

The Individualised Learner Record (ILR) data contain information about learners and the learning undertaken by them in England. Data are collected by learning providers in the Further Education (FE) and Skills sector in England that receive funding through an England based funding model as shown in the Specification of the Individualised Learner Record for 2023 to 2024. The Further Education and Skills sector includes FE colleges, sixth form colleges, training organisations, local authorities, academies, and voluntary and community organisations. ILR therefore provides excellent coverage of the population in further education in England. Data are submitted to Education and Skills Funding Agency (ESFA) within the Department for Education.

The Office for National Statistics (ONS) supply excludes learners born before 1 September 1984, leading to potential undercoverage of mostly older English FE students.

Learners are removed from ILR if they withdraw without completing one episodeof learning (a period of continuous enrolment at a single education provider, for example an individual completing four A-Levels over two years would be a single episode).

Accuracy and reliability

Data are collected via a management information system and submitted to ESFA using the learner data service. Learning providers receive ILR guidance to complete the ILR in accordance with validation rules and schema. This improves the accuracy and reliability because there is consistency across data collection and verification methods.

Data entry errors found in ILR data can be corrected as soon as the error is found if the error was made in the current academic year. Read more about correcting data errors. Once the data are supplied to the ONS, a series of quality checks are conducted.

Timeliness and punctuality

Data are collected from learning providers throughout the academic year, which runs from 1 August to 31 July and are submitted monthly to the ESFA. Data are supplied to the ONS annually between April and June of the following year, creating a lag of 9 to 10 months.

Accessibility and clarity

Data are supplied by ESFA to the ONS without any metadata, instead the ILR standard file specifications and reference data is used to understand how and when data are collected. The accessibility and clarity of the data are affected by the lack of sufficient metadata.

The data are formatted as one record per learner, per learning provider, per year. Learners are assigned a unique learner number (ULN) by ESFA, which follows all their learning instances at all UK providers, improving tracking of learners over time and data integrity.

Standard missingness codes are used, although they differ depending on the variable.

Comparability and coherence

Overall, ILR is mandatory for learning providers to complete, however some variables are optional. Therefore, there will be variation in the completeness of data per record affecting comparability and coherence across years.

The Quantitative Quality Indicators (QQI) for this source are provided on Worksheet 7 of QQI table.

12 . Benefit and Income Datasets (BIDs) quality assessment

Relevance

Benefit and Income Datasets (BIDs) contain information on anyone in the UK in receipt of a benefit, tax credit or state pension, as well as individuals who pay tax through the Pay As You Earn Real Time Information (PAYE-RTI) system. The purpose of the data is to monitor the benefits system and to aid decision making around policy development and evaluation, as described in Uses and users: DWP benefits statistical summary. BIDs are relevant for population statistics, providing information about the working age and pensioner populations present in England and Wales.

As shown in our Admin-based income statistics Quality and Methodology Information (QMI) report, BIDs consist of seven different datasets. Four are supplied by the Department for Work and Pensions (DWP):

- Universal Credit (UC)
- Single Housing Benefit Extract (SHBE)
- Personal Independence Payments (PIP)
- National Benefits Database (NBD)

Three are supplied by His Majesty's Revenue and Customs (HMRC):

- P14 (contains information derived from Pay As You Earn - Real Time Information (PAYE-RTI))
- Tax Credits (TC) - this contains Working Tax Credit, and Child Tax Credit (CTC), (CTC is being replaced by UC)
- Child Benefit (CB)

BIDs exclude those who are unemployed (including students not working) and not:

- in receipt of a pension
- claiming benefits
- part of a household claiming benefits

Self-Assessment is not included in BIDs, so self-employed individuals not paying tax through the PAYE-RTI system (such as carers, agency staff, painters for example) and who have not claimed any benefits will not be present on BIDs.

Accuracy and reliability

Data collection methods differ depending on the benefit being claimed. Applications may be completed online, by paper forms or telephone, if requested. The likelihood of data entry errors is increased by using paper and telephone collection methods.

Staff are required to undergo validation and compliance training to verify whether people are credible against a legal standard and to ensure they comply with the rules and receive their entitlement. This improves the accuracy and reliability because there is consistency across data collection and verification methods.

Data undergo validation checks conducted by DWP or HMRC. DWP conducts internal validation checks for consistency and missingness, for example. Read more about quality assurance principles, standards and checks on the government website.

Validation checks for HMRC supplied datasets include sense checks and comparison across years. Read the Child Benefit Statistics quality report as an example of the statistical processing and quality management of HMRC supplied datasets. Validation checks conducted by the suppliers increase the reliability of the data.

Timeliness and punctuality

BIDs data are not currently provided as a regular, routine supply to the ONS, therefore there is variation in lag. However, the ONS is in negotiation with DWP to secure timely monthly supplies from 2023, with a one to two-month lag.

Accessibility and clarity

Data are supplied by HMRC and DWP with a data dictionary containing information about the variables. Supplier provided data dictionaries improve the clarity of the data, as variables are clearly defined.

Data are obfuscated, meaning they are modified to be obscure or unclear to prevent disclosure of an individual's complete information. Obfuscated data affects the clarity of BIDs data available to the ONS for analysis as some details on records are not accessible. Data include a unique identifier to link to the Customer Information System (CIS) so demographic characteristics can be assigned to BIDs activity.

Comparability and coherence

BIDs datasets vary in terms of type and number of variables, formatting across variables, and expected counts for variables. Data items that are mandatory to complete also differ per dataset. To have a successful claim, individuals are required to provide information such as name, address and for most benefits, National Insurance number. Such variation affects comparability overtime and across different benefit datasets.

Demand for Universal Credit significantly increased in 2020 because of coronavirus (COVID-19). The increase in people interacting with DWP's benefit system potentially impacts the coherence of the data across years. We are still working to understand the impact of coronavirus on BIDs.

13 . Customer Information System (CIS) quality assessment

Relevance

The Customer Information System (CIS) contains demographic information on everyone who has a National Insurance number (NINo) in the UK so that the Department for Work and Pensions (DWP) can easily access information about their customers. DWP, His Majesty's Revenue and Customs (HMRC), the Department of Communities Northern Ireland (DfCNI) and death registrations report into the CIS, and the data is supplied to the Office for National Statistics (ONS) by DWP. Therefore, the CIS has a good coverage of the population in England and Wales.

Those that may not be present on this dataset because they do not have a NINo include:

- children whose parents have not claimed Child Benefit, however Child Benefit claims are excluded from CIS from 2021
- short term migrants
- international migrants who have retired
- migrants with dependents who have not claimed any benefits
- unemployed international students
- those who do not have a right to work in the UK or claim benefits

Accuracy and reliability

Data are collected when a UK citizen turns 16 years and applies for a NINo, or when a parent or guardian claims child benefit and HMRC allocates a reference number to the child, which is converted into a NINo when the child turns 16 years. Migrants need to apply for a NINo to work in the UK, a NINo is allocated if the application is successful.

Data are regularly updated through the DWP and HMRC systems. Death registrations are also used to keep CIS records updated, though risk of overcoverage remains where people move abroad and have not alerted DWP or HMRC.

DWP carry out internal validation checks to quality assure data while preparing a supply to the ONS, for more information see the Quality statement: DWP benefits statistical summary. Validation checks conducted by the supplier increase the reliability of the data.

Timeliness and punctuality

CIS data are not currently provided as a regular, routine supply to the ONS, therefore there is variation in lag. However, the ONS is in negotiation with DWP to secure timely monthly supplies from 2023, with a one to two-month lag.

Accessibility and clarity

CIS is a longitudinal dataset where new supplies replace old ones, limiting accessibility to data.

The ONS receives two tables of CIS data from DWP, with different demographic information on each. Data are obfuscated, meaning they are modified to be obscure or unclear to prevent disclosure of an individual's complete information. Obfuscated data effects the clarity of the data available to the ONS for analysis. For example, linkage error increases when details such as name and date of birth are obfuscated. The tables contain cumulative data at an individual level.

DWP provide the ONS with a data dictionary, which includes information about changes to variables. Supplier provided data dictionaries improve the clarity of the data, as variables are clearly defined.

Comparability and coherence

Data that are mandatory to complete differ depending on how an individual is added to CIS, reducing both comparability and coherence across years. However, there is guidance on the government website for DWP staff allocating NINos, to increase coherence across CIS.

14 . Ethics and security

All the administrative data sources used by the Office for National Statistics (ONS), including the data sources covered in this report are subject to robust controls to ensure that individuals cannot be identified. Furthermore, the ONS complies with all data protection legislation, the Data Protection Act 2018, the Statistics and Registration Act 2007 and the Digital Economy Act 2017. Further information, including the ONS's privacy statement and data protection policy can be found on our data protection page.

The ONS systems and data are secured by design, access and permissions are managed centrally and subject to strict approval processes. Access is restricted to users with the appropriate security clearance and an approved need-to-know business requirement.

15 . Data Sources

Quantitative Quality Indicators produced for administrative sources used to develop Statistical Population Dataset version 4.0
Dataset | Released 3 March 2023
We have produced Quantitative Quality Indicators (QQI) for the sources used to develop the Statistical Population Dataset (SPD) version 4.0 for England and Wales. The QQI are conducted on each source individually. We developed a grading system to simplify the QQI outputs into an interpretable and comparable quality assessment.

16 . Glossary

Administrative data

Collections of data maintained for administrative reasons, for example, registrations, transactions, or record-keeping. They are used for operational purposes and their statistical use is secondary. These sources are typically managed by other government bodies.

Dynamic Population Model

The Statistical Population Dataset (SPD) will be one of the core sources used in the Dynamic Population Model (DPM), which uses statistical modelling techniques and demographic insights alongside a range of data sources to produce coherent and timely estimates of the population and population change.

Observed population

The observed population is the people, units or objects that are actually recorded within an administrative data source, for example the electoral register should be populated with everyone that is eligible to vote however some people do not register to vote so the observed population is lower than the target population.

Overcoverage

Overcoverage occurs when a member of the target population is counted more than once at the same location, more than once at a different location, counted in the wrong location or is incorrectly included.

Statistical Population Dataset

Administrative data is used to approximate the usually resident population within England and Wales.

Target population

The target population is the people, units or objects that should be recorded within an administrative data source, for example the electoral register should be populated with everyone that is eligible to vote.

Undercoverage

Undercoverage occurs when some members of the target population are inadequately represented (population not being present on administrative data, or present in the wrong location).

Usually resident population

We are currently adopting the UN definition of "usually resident" that is, – the place at which a person has lived continuously for at least 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months (United Nations, 2008).

17 . Future developments

We will develop quality measures for all the main sources that we use to research our usually resident population and feed into the dynamic population model (DPM). This includes our coverage-adjusted Statistical Population Datasets (SPD), our international migration estimates and our internal migration estimates. It also includes having quality measures for individual sources that are used.

18 . Related Links

Developing Statistical Population Datasets, England and Wales: 2021
Article| Released 28 February 2023
Aggregate comparisons between the Statistical Population Dataset version 4.0 (v4.0) and Census 2021.

Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage
Article| Released 28 February 2023
Analysis of the Statistical Population Dataset Version 4.0 2021 using a linkage between Census 2021 and the Demographic Index.

Dynamic population model, improvements to data sources and methodology for local authorities, England and Wales: 2011 to 2022
Methodology| Released 28 February 2023
Developments of methods and data used in the dynamic population model.

Developing our approach for producing admin-based population estimates, England and Wales: 2011 to 2016
Article| Released 21 June 2019
Research into developing a new methodology to create population estimates from administrative data. These estimates are not official statistics on the population.

Administrative data used in Census 2021, England and Wales
Methodology| Released 15 September 2022
The administrative data sources that were used in the statistical design for Census 2021, with information on their coverage, accuracy, and timeliness.

Quality of administrative data in statistics
Methodology| Released 28 February 2023
A framework to help assess the quality of administrative data for use in the production of statistics.

19 . Cite this article

Office for National Statistics (ONS), released 3 March 2023, ONS website, article, Administrative sources used to develop the Statistical Population Dataset for England and Wales: 2016 to 2021.