DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 453–473, http://dx.doi.org/10.1515/JOS-2015-0027

Linkage of Census and Administrative Data to Quality Assure the 2011 Census for England and Wales

Louisa Blackwell[1], *Andrew Charlesworth*[2], *and Nicola Jane Rogers*[3]

The 2011 Census for England and Wales made extensive use of administrative data to quality assure the estimates. This included record linkage between census and administrative data. This article describes the role of record linkage in the quality-assurance process. It outlines the operational challenges that we faced and how we resolved them. Record linkage was confined to a sample within 58 carefully selected local authorities. We found characteristic patterns of under- and overcoverage in the National Health Service Patient Register, which we illustrate here with examples. Our findings may be useful in countries that, like England and Wales, do not have a comprehensive population register to draw on and that need to understand issues of coverage in their routinely collected administrative data and the use of these data to estimate populations.

Key words: Record linkage; administrative data coverage; linkage methods.

1. The Role of Administrative Data and Record Linkage in the Production of 2011 Census Estimates for England and Wales

This article describes how administrative data were used to quality assure the 2011 Census. This included record linkage between census and administrative data, which helped us to understand the discrepancies between these data that were found when aggregate-level totals were compared. Providing new insights into patterns of under- and overcoverage in the National Health Service Patient Register, this research also helped us to understand and explain why and how census estimates differ from administrative counts in particular types of local authority. We describe the methods, systems, and processes used for the linkage, and give an overview of our results and the conclusions that we drew from them. We also outline some of the operational challenges that we had to overcome. These challenges largely stemmed from the awkward reality that the research questions to be addressed by record linkage emerged during census processing and thus could not be known in advance. Our approach may be useful for other organisations and National Statistics Institutes that do not have the benefit of national population registers and that seek to understand the representativeness of routinely collected administrative data and their use in estimating the population.

[1] Office for National Statistics (ONS), Digital, Technology and Methodology, Fareham, PO15 5RR Hampshire, UK. Email: louisa.blackwell@ons.gsi.gov.uk
[2] Department for Energy and Climate Change, 3 Whitehall Place, SW1A 2AW London, UK. Email: Andrew.Charlesworth@decc.gsi.gov.uk
[3] Office for National Statistics (ONS), Population and Demography, Fareham, PO15 5RR Hampshire, UK. Email: nicola.rogers@ons.gsi.gov.uk

The 2011 Census for England and Wales aimed to count the entire population, both people and households. Asking the same questions everywhere, the census is an important source of data for comparing different parts of the country. It also underpins nonresponse weighting in a range of key national statistics produced from surveys. Ahead of the census, the Office for National Statistics (ONS) developed a comprehensive address register. This enabled ONS to add addresses and unique codes to every census questionnaire before distribution. These codes enabled ONS to keep track of paper and online returns and helped the central office to direct field staff to high-priority areas. One of ONS's strategic aims was "to maximise overall response rates and to minimise differences in response rates in specific areas and among particular population sub-groups" (Cabinet Office 2008, 23).

No census is perfect and inevitably some people and households were missed. ONS used complex statistical techniques to estimate missed people and households. This involved a coverage adjustment based on a large survey called the Census Coverage Survey (CCS), carried out independently of the census. Record linkage between the census and the coverage survey allowed ONS to estimate the population that the census had missed using Dual System Estimation methods (DSE, for more details see ONS 2012a). The estimation process incorporated a number of quality-assurance checks.

Beyond this, ONS carried out further quality assurance including a comparison of both the census counts and the estimates, which include DSE adjustments for under-coverage, against administrative sources. The aim of the quality-assurance process was to identify where further adjustments were required before the estimates could be finalised. Examples of potential issues included difficulties in data collection, data processing, or the estimation process that could lead to errors in coverage.

Extensive checking against administrative data was unprecedented for the England and Wales Census (see White et al. 2006 for usage in 2001) and involved thousands of comparisons. Where the core checks, carried out on all 348 local authorities in England and Wales, found differences that could not easily be explained, we carried out a supplementary analysis. This comprised two stages: an initial analysis at low geographic levels, and where this did not explain differences, record linkage. Ahead of the 2011 Census, ONS described when the extra quality-assurance work would be required and how it should be prioritised (ONS 2009, 2011a, 2011b). The quality-assurance approach was developed in consultation with academics, statisticians, demographers and users of census data (For further information, see ONS 2009 and 2012b). Record linkage between the administrative sources and the census was only used to investigate and understand discrepancies that could not be resolved at the aggregate level. This was the first time that ONS had used administrative microdata in this way for census quality assurance (ONS 2013a).

The approach was a 'top-down' strategy, driven by an overriding need for efficiency and timely results. The quality-assurance process for the 2011 Census was bounded by operational delivery constraints on the one hand and a desire to publish results in a timely fashion on the other. The 'window' for quality assuring the census estimates, initially local authority by local authority and then at the regional and national levels, demanded strict prioritisation. Record linkage for unresolved data anomalies was a possibility, but only for a limited number of areas. Thus record linkage was only carried out for areas where checks

on the aggregate-level data highlighted the need for more detailed investigation. In this respect the approach has some parallels with 'macroediting' that is used to find and correct errors in survey data by considering first the impact on data aggregates (see, for example Granquist 1991). The need for flexibility and analytic agility shaped the development of the data linkage system and the processes that we used during planning and the live operation.

The primary focus for quality assurance was the main population base for outputs, the usually resident population as of census day (27 March 2011). For 2011 Census purposes, a usual resident is defined as anyone who, on census day, was in England and Wales and had stayed or intended to stay for a period of twelve months or more, or had a permanent address in England and Wales and was outside of England and Wales on census day and intended to be outside for less than twelve months. This article sets out work done by the Census Quality Assurance Data Matching team, which began in 2011 and finished in 2013.

Section 2 of this article describes the administrative sources that were available for record linkage, together with the census information that we used, in addition to census responses. In Section 3 we then discuss the operational challenges that we faced, which were dominated by the need to complete the quality-assurance process quickly in order to publish timely results. Ahead of record linkage, the administrative data were used at aggregate level to address data anomalies that the core quality-assurance processes identified. Section 4 describes what we learnt from the aggregate-level comparisons. Section 5 describes our linkage methods and we present our results for 58 local authorities in Section 6. Our conclusions, in Section 7, aim to assist other National Statistics Institutes planning to make increased use of administrative data for population estimation in a census context. We also describe how ONS is taking forward administrative record linkage to support the 2021 Census.

2. The Data Available for Linkage

The 2007 Statistics and Registration Service Act provided a legal gateway for ONS to access record-level data (microdata) from other government departments for the purpose of population estimation. Through these and other provisions, ONS gained access to the NHS General Practitioner (GP) Patient Register, the School Censuses of England and Wales, the Higher Education Statistics Agency (HESA) Student Records, the Live Births Register, the Deaths Register, Electoral Registers, and Valuation Office Agency data.

Record linkage focused primarily on the NHS Patient Register. The Patient Register includes the general identity details of patients registered with GPs. It is used within the NHS for calculating payments to GPs and for the selection of NHS patients for participation in health-screening programmes. It is one of the largest population databases in operation in England and Wales. The Patient Register was the highest-quality record-level source with the widest population coverage that was available to us at that time. We also anticipated that queries about 2011 Census estimates from key users would be based on local Patient Register counts. In addition to using the Patient Register to quality assure the census counts and estimates, we needed to understand the quality of the Patient Register and its patterns of coverage, relative to the census, to respond to stakeholder queries following the publication of census results.

Quality checks ahead of record linkage confirmed that live births and deaths were reflected accurately in the Patient Register, so these were not included in this linkage exercise.

The census data used in record linkage included: census responses (both households and individuals), including 'dummy form' information which is supplied by enumerators for nonresponding households; the census address register; the census address register history file (ARHF), which contained addresses that were assessed as nonresidential or derelict and therefore not sent a census questionnaire; census 'associated address' records, including responses to the census question 'One year ago, what was your usual address?'; second residence addresses (including students' term-time addresses) and visitors' usual residence; field operation information drawn from the Census Management Information System (CMIS); census questionnaire images. Census questionnaires have been securely destroyed, but ONS is obliged to retain census questionnaire images, which will be made publicly available in 2111.

3. Building a Linkage Methodology and Architecture for Census Quality Assurance and the Imperative for a Flexible Approach

A number of issues and uncertainties demanded a flexible approach to record linkage. Some of the challenges we faced, and their resolution, were:

3.1. Security Risks

A number of physical, technical, statistical, and legal safeguards ensured that the microdata used for linkage were handled securely. Physical safeguards included restricting their use to the census physical safe setting, where security doors ensured that only authorised staff could enter. Technical safeguards included holding and processing microdata within the census IT environment, a closed and monitored system that did not allow users to copy, print or download the data being processed. The linkage design provided statistical protection as most of the linkage was within postcodes used for the CCS, and these are not publicly known. In addition, identifying information such as name, date of birth, postcode, and address were only used to link record pairs and were not stored in analytical datasets. Legal safeguards included the requirement for all staff, including the clerical matchers, to sign the Census Confidentiality Undertaking and Declaration and receive Defence Vetting Agency Security Clearance. The penalty for a breach of data confidentiality could be a prison sentence, and all staff in the matching team signed confirmations that they understood this.

3.2. Uncertain Analytic Requirements

It was impossible to predict all of the issues that record linkage would need to address. The geography or population subgroup under consideration would determine which administrative data should be used. The data architecture therefore had to allow linkage between all or just some sources, with capacity to add new sources if they became available. 'Data architecture' refers to the collection of interlinked tables used to store the results of all address and person linkage. These were held separately for each local

authority to maintain file sizes that were efficient to process. Using local authorities as the basic unit for analysis also reflected the quality-assurance process, which considered and approved the estimates for each local authority in turn.

3.3. Late Availability and Uneven Quality of Data

Only the Patient Register, Valuation Office Agency and census address register were available from the start of the quality-assurance process. Census person data became available as local authorities were processed (mirroring the order in which quality-assurance issues were raised), while CCS and other census information were only available late in the process. HESA, English and Welsh School Census and Births data became available after the quality-assurance process had begun. Electoral Register data were available for most local authorities, but were inconsistently formatted and required substantial cleaning and standardisation. A key requirement for the data-linkage architecture was the ability to incorporate new data if and when they became available.

The linkage algorithms that we used and the sequence of linking different sources had to remain flexible during the operation. For example, the School Census data for Wales were only available at a higher geographical level than for England. Our data tables and record linkage programmes were adapted to reflect this difference. Likewise, the Electoral Register cleaning and preparation revealed missing data for some local authorities. Where this occurred, we requested that records be resupplied and the subsequent delays impacted on the sequencing of local authorities through the linkage process.

3.4. The Requirement for Timely Results

Census quality assurance involved the approval of 348 local authority estimates at a series of Quality Assurance Panels (for more detail see ONS 2009 and 2011a). Where data issues could not be resolved using data at aggregate level, record linkage was used. Our systems and methods were designed to respond quickly to these requests, involving automation where possible.

The Census for England and Wales took place on March 27, 2011. ONS was committed to publishing the results in July 2012. The final agreement to publish the estimates was made by an Executive Quality Assurance Panel, the National Statistician and the Director General, executive ONS management and executive management representation from the Welsh Government. To achieve this, the estimates needed to be quality assured by April 2012. Census estimates were available for assessment by Quality Assurance Panels of ONS and external experts from September 2011. This provided just over six months to approve the estimates for all 348 local authorities, for the regions and at the national level in England and Wales. Within this brief window, record linkage, which is very labour intensive when it is supported by a clerical review of links being made, had to be done in a selective and efficient way.

To reduce the turnaround time required for data linkage results, we linked Patient Register addresses to the census address register in 37 local authorities ahead of the census. These local authorities were mostly areas of high population turnover, taking into account migration patterns since 2001. As census processing got underway, they were prioritised by the expected delivery date for their processed person-level data, in

anticipation of the order they would be considered by the Quality Assurance Panels. Record linkage for each of these local authorities was suspended if they were approved by the Quality Assurance Panel, and new areas not in the original list of 37 were added as new issues arose. These included some local authorities whose estimates fell outside the tolerance bounds set for the core checks (described in ONS 2012b), and where further analysis using aggregate-level data could not resolve the anomalies. By the end of the operation, data for 58 local authorities were linked. Identifying the more challenging local authorities and completing address linkage ahead of the live operation allowed preliminary work to proceed in an intelligent way, and maximised the number of local authorities overall that could be linked.

We included a number of local authorities with stable populations, which pose little enumeration challenge because they have low levels of international and internal migration. These provided a context for the results for more challenging areas. They also validated the linkage methods that we used.

3.5. Keeping the Scale of the Linkage Task at a Manageable Level

Some quality-assurance issues were concerned with small geographic areas or population subgroups, such as students in communal establishments or babies under the age of one. Where issues were generalised across the population, linkage typically focussed on the postcodes used for the CCS (for more details, see Abbott 2009). The CCS is a sample of approximately one per cent of the country carried out after the main census and is used to create the census estimate. The CCS uses a selection of postcodes within Output Areas (OAs), which are re-enumerated independently from the census field operation. The CCS selects a sample of OAs, stratified by local authority and a national 'hard-to-count' index. Output Areas (OAs) are the lowest geographical level at which census estimates are provided. They are built from adjacent postcodes. OAs cover 40–250 households and 100–600 people and postcodes have an average of 15 households. The 'hard-to-count' index is a proxy measure for census nonresponse (for further details, see ONS 2012a). The CCS re-enumerates approximately half of the postcodes within the selected OAs and contains more postcodes in areas where the census response rate was expected to be lower. Administrative data linkage within these clusters of postcodes provided a strategic sample that constrained the scale of the record-linkage task and also provided CCS data as an additional data source for comparison against the administrative data. Crucially, by using this sample we were able to provide record linkage and analysis for a greater range of local authorities.

3.6. Ensuring Quality and Consistency in Record Linkage

The quality of record linkage, both automatic and clerical, was monitored and managed through two processes. The first involved a continuous feedback loop of linkage best practice for the clerical matching team. An example of this was the accumulation of knowledge and experience in ethnically-specific naming conventions and variations. The second involved an expert matcher's review of linkage decisions, using both a random sample and having two matchers complete the same linkage. Systematic discrepancies were addressed through further training and review.

3.7. Complexity of Linkage and Storing Results at Both Individual and Address Levels

Storing the results was complicated by the large number of sources used, the two levels at which linkage took place (addresses and individuals) and the reality of one-to-many links for both addresses and individuals. These complications meant that extra care was necessary to analyse the linkage results. One-to-many links for addresses arose from less precise recording of addresses, for example in the Patient Register. This typically involved subdivisions within buildings (for example 'Flat 1') being omitted from a Patient Register address. Thus a number of addresses in the census, referenced in more detail, could link to a Patient Register 'shell' address. One-to-many person-level links arose from multiple enumerations of individuals in the census (discussed more fully in ONS 2012c). In addition, the linkage process allowed unlinked addresses to be linked as a result of person-level linkage, for example where capture errors (a typical example was where data scanning read marks on the paper questionnaire as characters) produced address differences that confounded the address linkage the first time around.

4. What We Learnt from Comparing Different Sources at Low-Level Geographies

Core checks, applied to all local authorities, included checking estimates by age, sex and other key variables against a range of aggregated administrative and survey sources. Where the core checks identified data anomalies, supplementary checks were carried out. These involved exploring the data at a low geographical level, mainly Output Area (OA) or above. Some checks were at postcode level.

In most cases, supplementary analysis resolved apparent data anomalies. The anomalies tended to arise as a result of two main problems, the first of which is the time lag that is inherent in many of the administrative sources. People's circumstances change (for example they move house), and there is a delay before this is captured in their administrative data. The failure of most administrative systems to capture reliable, timely information on migration leads to inflated datasets containing invalid records. A second problem that we found was a degree of subjectivity in addressing. Administrative systems vary in the level of detail or accuracy used to record where people live. This was more problematic where people live in subdivided properties. Typically we found that the census information was more timely and accurate. An exception was the addressing for student halls of residence, where the census sometimes captured the administrative building that census forms were sent to for onward distribution, whereas HESA data captured students' dwellings with greater geographic accuracy.

Supplementary analyses included comparisons between the census and Patient Register at person and household level. For example, a discrepancy between census and Patient Register counts in Westminster found one area where there were several thousand more patient registrations than census individuals. Analysis by age found that the excess patient registrations were mostly of student age. Further investigation revealed that this area contained a medical centre attached to a London university. The address for this centre was wrongly given as the home address of many students registered with the practice.

In areas with high concentrations of students, the number of patient registrations often exceeded census counts for young adults. Further investigation revealed that Patient

Register counts implied that student halls were filled beyond their published capacities (see Figure 1), with the ratio of registrations to published capacities frequently higher than one. Further analysis of the date that these patients were registered confirmed that former residents had almost certainly moved on but not updated their NHS records, either because they had not yet reregistered with a new GP or had left England and Wales.

We also compared census counts and estimates against, among others: Patient Register counts of under 1's and those in the Register of Live Births; School Census counts of ethnic groups; lists supplied by local authorities of addresses containing 'annexes', along with Valuation Office Agency information and Patient Register counts; Patient Register and School Census counts for addresses within holiday parks; international migrants as defined by Patient Register records with 'flag 4' status, given to new registrations from abroad.

To understand and explain a substantial difference between census counts and council tax records, we found one area where the census found fewer than twenty households, yet the council tax data showed several hundred more. This was explained by a large block of flats that had been almost completely emptied for demolition.

5. Linkage Methods

Figure 2 summarises our record linkage processes. Linkage involved exact automated linkage, score-based automatic linkage (using similarity scores), clerical resolution of candidate pairs generated by the automatic systems, and a clerical search for residual records. This was a unique exercise carried out to validate the census.

5.1. Data Preparation

Each administrative dataset was standardised and cleaned, including removing duplicates, checking and aligning variable formats, checks for coding inconsistencies, and checking the number of unknown or missing values for each variable (1.1 in Figure 2). Electoral Registers were the most resource intensive to prepare. Maintained and supplied by

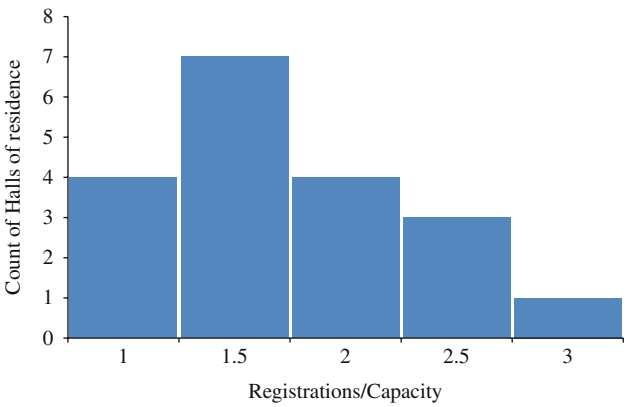

Fig. 1. Ratio of patient registrations and published capacities in student halls, in a sample of halls of residence in one university town

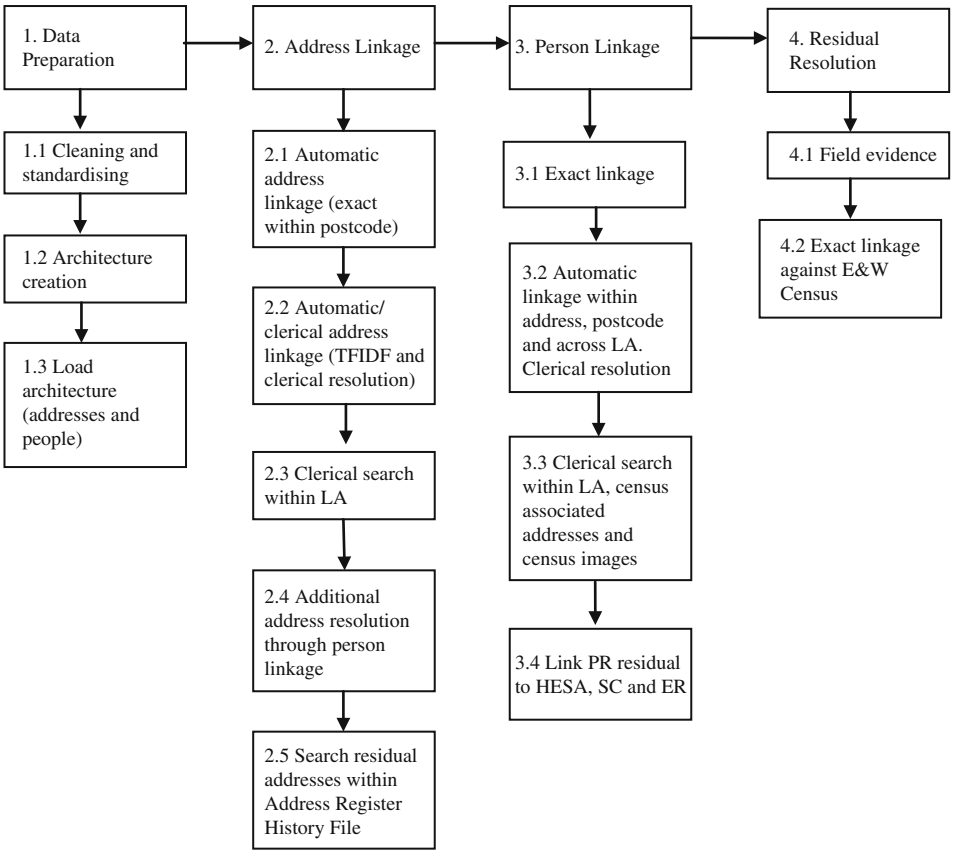

Fig. 2. 2011 Census quality-assurance record-linkage process. Abbreviations: LA (local authority), PR (Patient Register), SC (School Census), ER (Electoral Register), E&W (England and Wales)

individual local authorities, these registers were held in a wide range of formats. Some of the standardisation could be automated, while some rare and unique differences required manual correction.

We attached geography codes to addresses using the software package 'Matchcode', supplied by Capscan (for more detail see http://www.capscan.com/).

We aligned the administrative sources to census definitions where possible. For example, HESA data record all students on a course at an institution within an academic year, regardless of the course duration, so individuals may have multiple instances within an institution in the same academic year. To align these data with census definitions, we used a subset of HESA records for those aged 18 and over with start dates before and end dates (or continuing) after March 27, 2011 (census day). Rules to prioritise multiple records were applied to select just one record for linkage. See ONS (2012d) for more information on the challenges of aligning definitions.

The final two stages of data preparation, 'Architecture creation' and 'Load architecture' (boxes 1.2 and 1.3 in Figure 2) refer to the creation of interlinked data tables where we stored the data for linkage and the linkage results.

5.2. Address Linkage

Addresses in the Patient Register, Electoral Register, Valuation Office Agency data, and the English School Census were linked with those in the census address register within CCS postcode clusters in selected local authorities.

Exact linkage (2.1 in Figure 2) used only flat number/property subdivision/house name, house number and road, and finally postcode. Variables with low discriminatory power such as town were excluded as they could only introduce error.

A second stage (2.2 in Figure 2) used 'Term Frequency Inverse Document Frequency' (TFIDF) linkage, which assigns a weight to each pair of words in a pair of addresses, depending on how commonly the words within the addresses appear in each of the datasets (Li et al. 2010). TFIDF linkage used all available address elements. Common terms within the address, such as 'town', calibrate and weight the less frequent ones. Linked records incorporating 'Hill Street' would have a lower weight than 'Segensworth Road', due to the rarity of 'Segensworth'. Scores for each address are weighted according to the number of words included in the address. The best-scoring candidate match for each address was referred for clerical review and confirmation.

A third stage (2.3 in Figure 2) involved a clerical matcher searching for an address match, firstly within the given postcode and then across the local authority as a whole.

Inaccuracies in recording addresses led to some addresses being falsely unlinked. Some of these addresses were subsequently linked through person linkage (2.4 in Figure 2). Where individuals living in unlinked addresses were linked, a check was made to see if these were falsely unlinked addresses due to data discrepancies.

Finally, in addition to searching for matches within census data, the Address Register History File (ARHF) was also checked (2.5 in Figure 2). The ARHF contained addresses that had not been sent a census questionnaire, for example because they were commercial addresses or known to be derelict buildings.

5.3. Person Linkage

Individuals within the Patient Register were linked to census records. Unlinked patient registrations were then searched for within the Electoral Register, School Census, and HESA data to assess the strength of their presence in administrative data. Our linkage strategy was deliberately designed to maximise linkage rates while minimising false links.

As with address linkage, the first stage of person linkage (3.1 in Figure 2) was exact linkage using forename initial, the first three characters of surname and full date of birth (dd/mm/yyyy).

A number of automatic linkage strategies followed (3.2 in Figure 2), firstly using the results from address linkage. Within linked addresses, the linkage criteria were relaxed to forename initial or a SPEDIS value of less than 100, first three characters of surname or SPEDIS value of less than 100 and two of the three date-of-birth elements matched. SPEDIS measures how close the spellings of two words are. It is a function within the SAS statistical analysis software package. The lower the score, the better the match. This relatively high threshold allowed potential matches to be referred for clerical resolution (scores of 101–200 were disallowed).

Within matched postcodes, linkage criteria were the first three characters each of forename and surname and two of three elements of date of birth. When searching more widely for CCS postcode cluster records within a local authority, forename, surname, date of birth, and sex all needed exact matching.

There then followed rules-based linkage techniques (see Li et al. 2006). Firstly, within local authorities, individuals with the same day and year of birth and sex were linked using month of birth, exact forename and surname with a qgram threshold of 0.4 or above. Qgrams measure the level of agreement between groups (in our case, pairs) of characters within the two strings being compared (the code for the qgram comparison is available from ONS upon request). The second strategy required exact surname matching and forenames with a qgram threshold of 0.4 or above.

All exact matches were recorded without further scrutiny. For individuals linked within linked addresses, those with name discrepancies, where sex was uncoded and where there was error in dates of birth were referred for clerical confirmation. All matches within postcodes and local authorities were reviewed clerically, as were duplicate matches and those identified through the rules-based linkage strategies.

Unlinked patient registrations were searched for clerically, firstly across the local authority and secondly through 'associated address' information (3.3 in Figure 2). This involved matching against census respondents who gave the Patient Register address as their usual address one year ago, as a second residence or as a usual residence for visitors.

To identify census matches missed because of potential data-scanning error, census form images were checked.

Where linked individuals were in addresses that were unlinked, these were referred for clerical review. In this way, addresses that either were recorded very differently between sources or contained scanning error were resolved (2.4 in Figure 2). Clerical matchers were able to carry out free text searches on name and address and any combination of day, month and year.

5.4. Residual Resolution

Any patient registrations that remained unlinked at the end of this thorough linkage process were searched for within the other administrative sources: the Electoral Register, School Census, and HESA data (3.4 in Figure 2).

To further resolve unlinked records, we used evidence from the census field operation (4.1 in Figure 2). Where there was no response to the census, enumerators classified addresses according to evidence they could find in the field. Thus we were able to classify unlinked records as having an address that appeared to be a second home, having an address that was occupied but the occupants were refusing to comply with the census, or as clearly vacant.

A final person linkage stage involved searching, using exact matching, across England and Wales as a whole (4.2 in Figure 2).

Table 1 provides examples of linkage rates achieved through exact, rules-based, and clerical methods. It highlights the limitations of exact linkage. Inconsistencies between names on the Patient Register and on the census form arose for a number of reasons including inconsistencies in recording names, such as abbreviations ('William' and 'Bill' for example), middle names given as forenames, inconsistent translations from

Table 1. Patient Register to 2011 Census linkage rates at each processing stage

Local authority	LA with a stable population Aylesbury Vale	Metropolitan LA Birmingham	Inner London LA Lambeth
Total number of patient registrations in the sample	2,732	21,313	10,532
% Exact linkage (3.1 in Figure 2)	54.0	50.3	34.3
% Rules–based linkage with clerical resolution (3.2 in Figure 2)	13.8	18.3	19.0
% Clerically linked (3.3 in Figure 2)	21.0	13.0	10.0
Final linkage rate	88.8	81.5	63.3

non-English (such as Chinese or Russian) characters into the English alphabet, and scanning error, among others.

6. Linkage Results for 58 Local Authorities

Record linkage proceeded on a local authority by local authority basis. Areas were selected for record linkage either because they were high migration areas where the different sources were most likely to diverge, because the Quality Assurance Panel had identified data anomalies and wanted further analysis, or because we had identified them as a useful benchmark against which to compare more challenging areas. As the number of local authorities with linked patient registrations grew, it became clear that a typology of local authorities was visible in the data.

Inevitably, not all records can be linked. Firstly, some census respondents are not registered with an NHS GP. Examples include new arrivals to England and Wales who are yet to register with a GP; those who have moved to a new area and not updated their GP registration; people using private health care rather than the NHS; those covered in the NHS outside of the GP system, such as prisoners or members of the armed forces.

Secondly, although the census aimed to capture the entire population on census night, some people were missed (the 2011 Census person response rate was 94 per cent and overcoverage was estimated at 0.6 per cent). ONS estimates the extent of undercoverage (at six per cent) using the CCS and DSE or Dual System Estimation. In terms of the administrative record linkage carried out for census quality assurance, the individuals that the census and the CCS missed could appear as unlinked patient registrations.

There is also an issue of synchronicity between the datasets. The census provides a snapshot of the population of England and Wales on census night, March 27, 2011. The Patient Register extract was taken on April 23, 2011. The gap between these reference dates was to allow people moving house to register with a GP in their new area. However, some people take longer for this so there will always be some disagreement between the sources, even in areas with relatively little population turnover (Smallwood and Lynch 2010). Moreover, if people who leave the country do not inform their GP that they are going, they remain on the register until the local health authority cleans them off the list.

Population groups that are absent or over-represented on the Patient Register produce characteristic differences in the demographic profiles for local areas. Area characteristics also shaped the patterns of coverage, as we show below for university towns. As a further example, Richmondshire and Forest Heath local authorities are home to large military bases and here the 2011 Census estimate exceeds the Patient Register count by 15 and 14 per cent, respectively. Among males aged 16–64, this rises to 37 and 15 per cent. Kensington and Chelsea have 2011 Census estimates that are six per cent higher than the Patient Register count for those aged 65 and over, reflecting a concentration of private healthcare users here (see ONS 2012e).

Powys, where the 2011 Census estimated 133,000 usual residents, had the highest Patient Register linkage rate of 93.7 per cent. This left 104 unlinked patient registrations and 231 unlinked census records *within our sampled postcodes*, where the total number of patient registrations was less than 2,000. Areas with stable populations typically had linkage rates above 85 per cent. Areas with higher levels of population turnover had Patient Register linkage rates of between 75 and 85 per cent. Linkage rates below 75 per cent typically occurred in London, and were lowest in the Inner London boroughs, where population turnover and international migration are at their highest. Kensington and Chelsea had the lowest linkage rate, with fewer than two thirds (60.5 per cent) of patient registrations linked to the census. However, comparisons of unlinked records and the coverage adjustment in each area (not shown here) provided further confidence in the census estimates. Linkage rates are summarised in Table 2.

6.1. Local Authorities With Stable Populations

Figure 3 shows the unlinked Patient Register and census records for males in a local authority with a stable population. Areas with high record-linkage rates were those with low levels of internal and international migration. Unlinked patient registrations tended to be higher for working-age people. There were more unlinked census records (dashed lines in the graphs) than unlinked patient registrations (dotted lines). This was true for most of the local authorities where linkage rates were high. Even in these areas where the two sources were most closely aligned, there were more unlinked records for men than for women (not shown here).

Patient Register records appear to be less accurate for men, who visit their GPs less frequently. This leads to longer time lags in updating NHS registrations when men move house than when women do, and the result is that Patient Register entries refer to people, men in particular, who no longer live in the area. This is more problematic in local authorities with less stable populations, such as inner-city areas, which people migrate to for work or study purposes.

6.2. Inner London

The discrepancy between the census and the Patient Register was greatest in Inner London. Figure 4 shows the linkage results for males in an Inner London local authority. For men between the ages of 25 and 44, the Patient Register had more unlinked records than records that linked to the census.

Table 2. Summary of person linkage rates

	April 2011 Patient Register			2011 Census		
	Average linkage rate* (%)	Highest linkage rate (local authority) (%)	Lowest linkage rate (local authority (%)	Average linkage rate* (%)	Highest linkage rate (local authority) (%)	Lowest linkage rate (local authority (%)
All 58 local authorities	79.7	93.7	60.5	81.2	93.6	63.9
Local authorities with stable populations	88.3	93.7	80.9	87.5	93.6	77.8
Metropolitan areas excluding Inner London	81.0	85.6	76.9	80.0	88.1	73.5
Inner London	68.3	75.2	60.5	72.3	79.2	63.9

*This is measured across all local authorities or local authorities in this category and is the total of the individual linkage rates divided by the number of local authorities in that category. For each local authority, the Patient Register or census linkage rate is the number of *linked* registrations or census records as a percentage of the *total number* of registrations or census records.

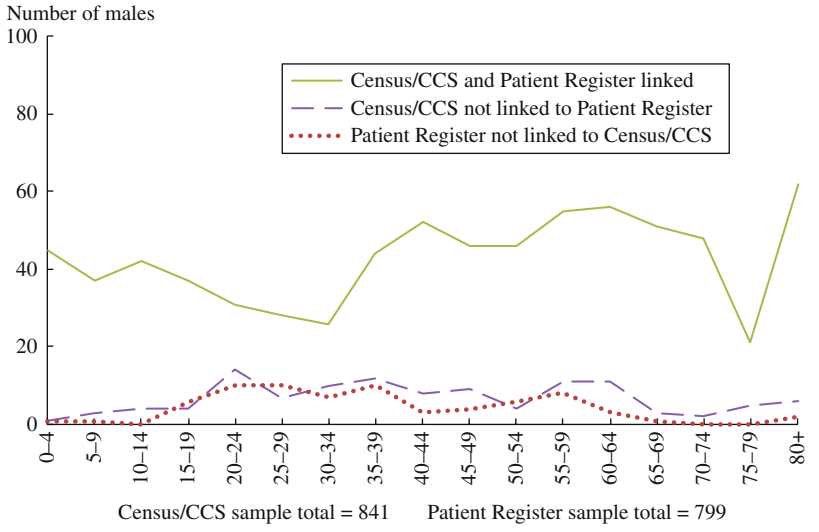

Fig. 3. Census and Patient Register linkage results for males in an area with a stable population, by age

6.3. University Towns

Another pattern that we found in some areas with high proportions of students is illustrated in Figure 5. Here, there were more unlinked census records for men aged 20–24 than were linked to the Patient Register. Thus over half of the men in this age group that the census or CCS captured were different to those on the local Patient Register. Many students leave their patient registrations at their home (parental) addresses. After (eventually) registering with a GP at their term-time address, men in particular are slow to update their addresses on the Patient Register at their new address when they move away. The extent of this

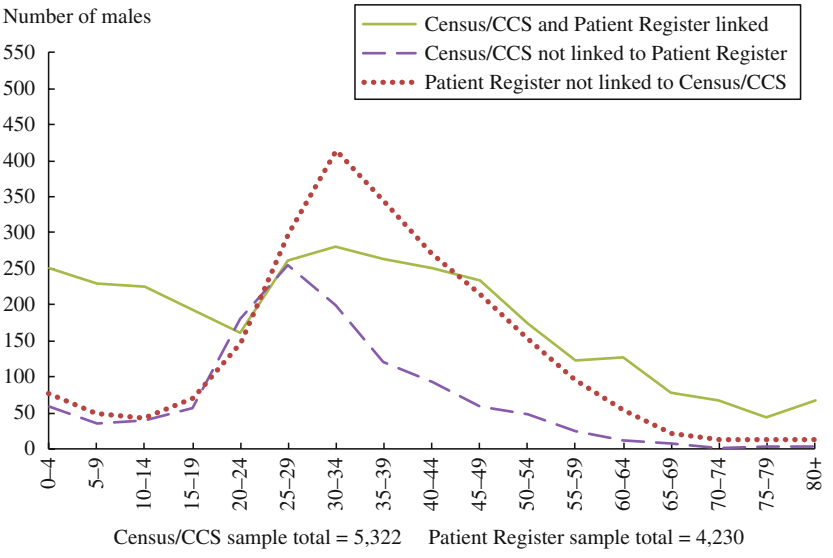

Fig. 4. Census and Patient Register linkage results for males in Inner London, by age

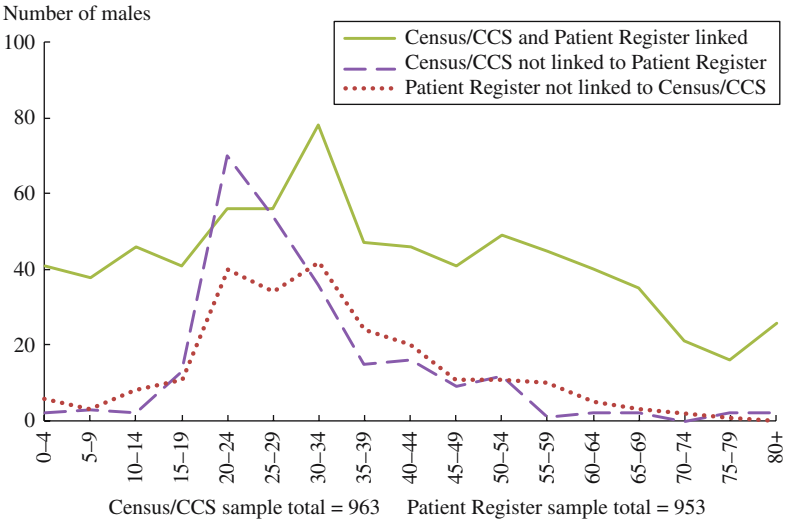

Fig. 5. Census and Patient Register linkage results for males in a university local authority, by age

disagreement between the sources cannot be deduced from the comparison of totals (Figure 6). Since the totals are similar, the comparison masks the problem that they include many people who are unique to each source.

6.4. International Migration and Excess Patient Registrations

People born outside the UK are more likely than the UK born to leave. If people do not de-register before they emigrate, their registration remains active until it gets cancelled in periodic Health Authority (HA) list-cleaning operations. The delay could cause an overcount, which contributes, for example, to the excess patient registrations seen in areas with large populations from overseas, including overseas students.

Flag 4 in the Patient Register denotes new registrations from abroad. Figure 7 compares, for each local authority, the proportion of patient registrations that did not link to the census against the proportion of unlinked records that have 'flag 4' status on the Patient Register. The local authorities fall into four groups:

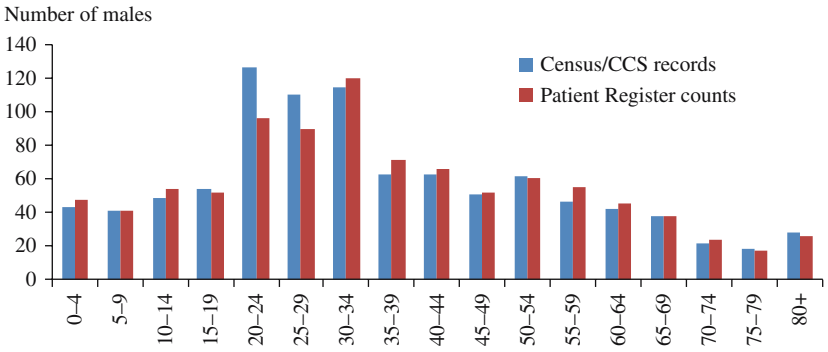

Fig. 6. Census and patient register totals for males in a university local authority, by age

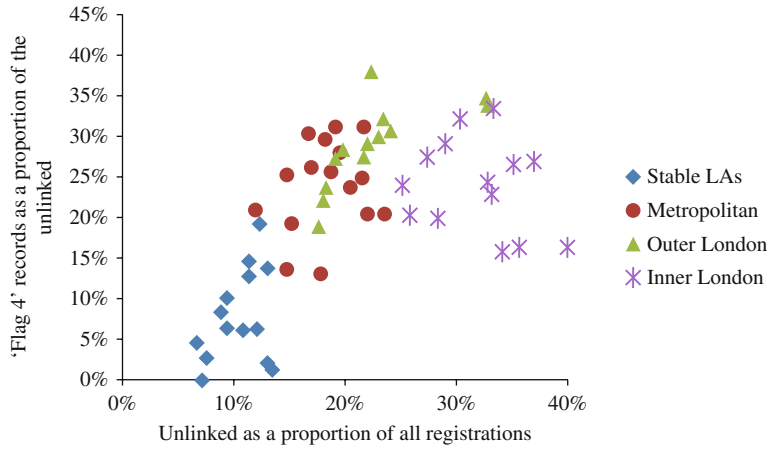

Fig. 7. Unlinked patient registrations and the proportion of international migrants within them, within local authorities. Note: **LAs with stable populations included:** *Aylesbury Vale, Blaenau Gwent, Ceredigion, Cheshire East, Knowsley, Maidstone, Mendip, Mid Devon, New Forest, Powys, Richmondshire, Tendring, Torbay.* **Metropolitan areas:** *Birmingham, Blackburn with Darwen, Bradford, Cambridge, Cardiff, Colchester, Leeds, Leicester, Liverpool, Manchester, Newcastle upon Tyne, Nottingham, Oxford, Sefton, Sheffield, Slough, Southend-on-Sea.* **Outer London:** *Barking and Dagenham, Barnet, Brent, Croydon, Ealing, Enfield, Harrow, Hounslow, Kingston upon Thames, Merton, Newham, Richmond upon Thames, Waltham Forest.* **Inner London:** *Camden, City of London, Greenwich, Hackney, Hammersmith and Fulham, Haringey, Islington, Kensington and Chelsea, Lambeth, Lewisham, Southwark, Tower Hamlets, Wandsworth, Westminster.*

- LAs with stable populations and with the least unlinked registrations, of which very few were new registrations from abroad.
- Metropolitan areas outside of London, with higher proportions of unlinked records. Among the unlinked records, between a quarter and a third were new registrations from abroad.
- Outer London local authorities, which were similar to the other metropolitan areas but tended to have higher proportions of unlinked patient registrations. (These were combined with 'Metropolitan Areas excluding Inner London' in Table 2.)
- Inner London local authorities with the highest proportions of unlinked patient registrations. Here, the proportions that were new registrations from abroad were similar to local authorities in Outer London and other metropolitan areas.

The combined evidence from the local authorities we analysed suggests that in areas with excess patient registrations, fewer than half were new registrations from abroad. Thus both internal migration and international migration are associated with excess patient registrations.

In all local authorities we found that unlinked patient register records were more likely than linked ones to be new registrations from abroad.

7. Conclusions

The 2011 Census for England and Wales used administrative data to quality assure census counts and estimates to a degree that was hitherto unprecedented and involved record

linkage between census and administrative data for the first time. In the process, it revealed patterns of differential coverage in routinely collected administrative records, notably the NHS Patient Register.

The availability of rich sources of administrative data provided an unprecedented opportunity to assess and possibly enhance the quality of the census estimates, but posed some serious operational challenges. The scope for using the administrative sources was very time limited and difficult to plan ahead. The research questions to be answered by record linkage (and by extension, the administrative sources to be used) were not known in advance. In order to provide timely evidence for census quality assurance, flexibility was the key:

- Flexibility to hold and link new and upcoming sources as the census operation progressed
- Flexibility to exploit and incorporate the full range of census information as it emerged, including enumerators' 'dummy' returns for nonresponding households or derelict properties
- Flexibility to switch data linkage effort between areas as required by the Quality Assurance Panels
- Separation of person and address linkage so that linkage to the census address register would give us a head start before person-level census data were available and the quality-assurance process was fully underway
- Flexible analytical resources so that different and multiple sources could be used, as aggregates and as microdata, to address research questions that were not known in advance.

The CCS was important for the data linkage task because it provided a strategic sample that constrained the scale of record linkage and augmented the data available for analysis.

Out of 348 local authorities, we linked data in 58. These were the most challenging areas, together with some with stable populations against which we could benchmark our results. We needed high-quality linkage as the Patient Register was used by local authorities as a comparator for census estimates. We extended our linkage strategy to incorporate rules-based linkage. The use of clerical matchers was key to the success of this approach. We found inevitable discrepancies between the census and Patient Register, due to time lags in updating address information and definitional and coverage shortfalls in NHS Patient Registrations.

We found overcount in the Patient Register in areas with high levels of internal and international migration, and higher levels of overcount for men than for women. Because women visit their GPs more frequently, we speculate that they are more likely to be recorded at their current address. In some university towns, the 2011 Census data appear to be more accurate and timely for the student population than the Patient Register, as a result of a tendency for undergraduates to remain registered at their parental address. Analysis of unlinked records provided further confidence in the census estimates.

The Census Quality Assurance Panel recommended to the National Statistician that census estimates for all 348 local authorities could be published, but in the course of the quality-assurance process there were minor adjustments based on comparisons against administrative data. Even though the comparisons did not lead to any substantial

amendments to census estimates, the process was very worthwhile. Firstly, it increased our confidence in the 2011 Census processes and resulting estimates; secondly, our understanding of the coverage and quality of administrative sources was greatly enhanced through both the aggregate-level comparisons and through record linkage. This provided valuable and transparent evidence to address queries that were raised about the census estimates following their publication. Thirdly, our experience of carrying out the linkage and analysis has helped to shape and inform the use of administrative data for population estimation.

7.1. Beyond 2011

In May 2010, the UK Statistics Authority asked ONS to begin a review of the future provision of population statistics in England and Wales in order to inform the government and Parliament about the options for the next census. In response, the ONS set up the Beyond 2011 Programme to undertake this work. The Programme has undertaken extensive research into and consultation on new approaches to counting the population and reviewed practices in other countries. A key focus of this work has been research into making better reuse of administrative data. The research culminated in the National Statistician making her recommendation in March 2014 (ONS 2014), which was subsequently accepted and endorsed by the Board of the UK Statistics Authority and supported by the government in July 2014.

Three key strands of work have been identified to take forward the National Statistician's recommendation:

- **2021 Census Operation** – research, development, implementation and operation of a 2021 online Census and Census Coverage Survey. At this early stage we anticipate that special attention will need to be given to online collection, the modernisation of our field processes, and making better use of administrative data.
- **Integrated Population Statistics Outputs** – integration of census, administrative and survey data to produce outputs. In this case, attention is focusing on taking forward work on data linkage, considering how administrative data can be used both to enhance the census and produce new or improved outputs.
- **Beyond 2021** – research into the shape of the census and population statistics system beyond 2021. This longer-term work will look at proposals for the future of the census and population statistics beyond 2021, including research into the potential need for new surveys after 2021 and the benchmarking of new methods.

The programme involves working with large quantities of personal information relating to everyone in England and Wales, obtained from a range of administrative sources. It is recognised that the planned approach of linking multiple administrative sources might elevate the associated risks relating to the privacy of data concerning people and households. To mitigate this risk, ONS has decided to anonymise administrative data prior to linkage to ensure that high levels of anonymity and privacy are maintained. This has resulted in developing a new method for linking anonymous data (more details are provided in ONS 2013c). Further information on the policy for safeguarding data during the research phase of the Beyond 2011 Programme can be found in ONS (2013b).

8. References

Abbott, O. 2009. "2011 UK Census Coverage Assessment and Adjustment Strategy." *Population Trends* 137: 25–32. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/2011-uk-census-coverage-assessment-and-adjustment-methodology—article-from-population-trends-137.pdf (accessed 16 July, 2015).

Cabinet Office December. 2008. "Helping to Shape Tomorrow- the 2011 Census of Population and Housing in England and Wales." UK: The Stationary Office.

Granquist, L. 1991. "Macro-Editing – A Review of Some Methods for Rationalising the Editing of Survey Data." *Statistical Journal of the United Nations Economic Commission for Europe* 8: 137–154.

Li, B., H. Quan, A. Fong, and M. Lu. 2006. "Assessing Record Linkage Between Health Care and Vital Statistics Databases Using Deterministic Methods." *BMC Health Services Research 6*. Doi: http://dx.doi.org/10.1186/1472-6963-6-48.

Li, D., S. Wang and Z. Mei. 2010. "Approximate Address Matching," In proceedings of the International Conference on *P2P, Parallel Grid, Cloud and Internet Computing (3PGCIC)*, 4–6 November, 2010. Doi: http://dx.doi.org/10.1109/DGCIC.2010.43, 264–269, Institute of Electrical and Electronic Engineers. New York: USA.

Office for National Statistics. 2009. *2011 Census Data Quality Assurance Strategy*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/2011-census—data-quality-assurance-strategy.pdf (accessed 16 July, 2015).

Office for National Statistics. 2011a. *2011 Census – Methodology for Quality Assuring the Census Population Estimates*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/2011-census—methodology-for-quality-assuring-the-census-population-estimates.pdf (accessed 16 July, 2015).

Office for National Statistics. 2011b. *Guidance on Core to Supplementary QA*. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/guidance-on-core-to-supplementary-qa.pdf (accessed 16 July, 2015).

Office for National Statistics. 2012a. *The 2011 Census Coverage Assessment and Adjustment Process*. Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/index.html (accessed 16 July, 2015).

Office for National Statistics. 2012b. *Quality Assurance of 2011 Census Population Estimates*. Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-assurance/index.html (accessed 16 July, 2015).

Office for National Statistics. 2012c. *Overcount Estimation and Adjustment*. 2011 Census: Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release–quality-

assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf (accessed 16 July, 2015).

Office for National Statistics. 2012d. *Beyond 2011: Exploring the Challenges of Administrative Data*. Methods and Policies Report (M2). Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011—exploring-the-challenges-of-using-administrative-data.pdf (accessed 16 July, 2015).

Office for National Statistics. 2012e. *Comparisons Between 2011 Census Estimates and the GP NHS Patient Register Adjustment*. 2011 Census: Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/how-census-compares-with-other-data-sources/index.html (accessed 16 July, 2015).

Office for National Statistics. 2013a. *Results From Using Routinely-Collected Government Information for 2011 Census Quality Assurance*. 2011 Census: Methods and Quality Report. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-assurance/index.html (accessed 16 July, 2015).

Office for National Statistics. 2013b. *Beyond 2011: Safeguarding Data for Research: Our Policy*. Methods and Policies Report (M10). Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-safeguarding-data-for-research-our-policy–m10-.pdf (accessed 16 July, 2015).

Office for National Statistics. 2013c. *Beyond 2011: Matching Anonymous Data*. Beyond 2011 M9 Methods & Policies. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-matching-anonymous-data–m9-.pdf (accessed 16 July, 2015).

Office for National Statistics. 2014. *The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority*. Available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation–and-recommendations/national-statisticians-recommendation.pdf (accessed 16 July, 2015).

Smallwood, S. and K. Lynch. 2010. "An Analysis of Patient Register Data in the Longitudinal Study - What Does it Tell Us About the Quality of the Data?" *Population Trends* 141: 151–169.

White, N., O. Abbott, and G. Compton. 2006. "Demographic Analysis in the UK Census: a Look Back to 2001 and Looking Forward to 2011." In Proceedings of the American Statistical Association, Survey Research Section. Alexandria, VA: American Statistical Association.

Received January 2014
Revised February 2015
Accepted March 2015