

International Journal of Population Data Science

Journal Website: www.ijpds.org



Estimating households and populations from primary care electronic health records: comparison with Office for National Statistics Census 2021 aggregated estimates

Marta Wilk^{1,*}, Gill Harper¹, Nicola Firman¹, Chris Dibben², Rich Fry³, and Carol Dezateux¹

Submission History

Submitted:	10/12/2024
Accepted:	22/07/2025
Published:	24/11/2025

¹ Centre for Primary Care, Wolfson Institute of Population Health, Faculty of Medicine and Dentistry, Queen Mary University of London, Yvonne Carter Building, 58 Turner Street, London, E1 2AB

² The University of Edinburgh, School of Geosciences, Drummond St, Edinburgh EH8 9XP

³ Swansea University, Medical School, Singleton Park, Swansea, SA2 8PP

Abstract

Introduction

Up-to-date, high-quality estimates of population and households are essential for planning the provision of local and central infrastructure.

Objectives

We aimed to derive estimates of population size, and household numbers and size on Census date (21/03/2021) using north-east London primary care Electronic Health Records (EHR) and calculate levels of their agreement with the publicly available official Census 2021 estimates to assess if health data have the potential to be used to create reliable statistics.

Methods

We compared EHR and Census population estimates by sex, age, local authority, and IMD quintile, and EHR and Census household estimates by number, size, and local authority. We estimated 95% Limits of Agreement between EHR and Census household and population estimates using the Bland and Altman method. In sensitivity analyses, we excluded people with no General Practice encounter within 12 months and compared the adjusted population's size to Census estimate.

We compared EHR and administrative Statistical Population Dataset (SPD) to Census population estimates by sex and age, and EHR and Admin-based Occupied Address Dataset (ABOAD) to Census household estimates by local authority and household size.

Results

EHR population estimate was 2,130,965, i.e. 7.1% higher than Census of 1,990,087. EHR household estimate was 658,264, i.e. 9.1% lower than Census of 724,045. The estimate of population with recent GP encounter was 11.6% lower than the Census estimate.

Compared to Census, both SPD and EHR overcounted population of males (10.7%, 7.9% respectively) and females (3.6%, 2.7% respectively). Both ABOAD and EHR had undercounted households compared to Census (-7.3%; -9.1% respectively).

Conclusions

Reliable, up-to-date populations and households estimates can be derived from health records. High residential mobility increases the complexity of deriving these estimates. Excluding people without GP encounters does not improve agreement with Census. Future work will focus on comparing Census and EHR estimates using individual-level data.

Keywords

Population estimates; household estimates; electronic health records; data representativeness; census comparison study

*Corresponding Author:

Email Address: marta.wilk@qmul.ac.uk (Marta Wilk)

Introduction

Up-to-date, high-quality estimates of population and household numbers, together with household size and composition, are used by local and central governments for planning the provision of schools, housing, health and social care services, and other infrastructure, allocation of funds based on relative populations, and by government and non-government bodies for deriving the regional statistics including those on health, economy and education.

Currently, population and household estimates for England and Wales are obtained from the national Census, planned and carried out by the Office for National Statistics (ONS) every ten years. Aggregated data are published by small geographic areas and by selected population characteristics including sex, age, ethnic group, and household composition [1, 2]. The Census data collection and processing methodology aims to provide accurate, high-quality statistics on population and household numbers and characteristics [3].

The population estimates are updated between consecutive Censuses by ONS, by producing mid-year population estimates through the cohort component method and using the previous Census as a benchmark [4]. However, these include only limited demographic information and do not include information on households, and they can become less accurate in the years further away from the previous Census [5, 6]. There is therefore interest in producing more up-to-date population and household estimates from linked routine administrative data in the United Kingdom [7, 8] as well as internationally [9–12].

Primary care electronic health records (EHRs) document individual patient care, enabling continuity and high quality of healthcare as well as population health management [13]. National Health Service (NHS) primary care EHRs in England and Wales cover the entire registered patient population, are updated in near real-time, and contain a breadth of demographic and health information. Although recorded primarily for administrative reasons, they are commonly used for secondary purposes, including research, local intelligence, and planning and tracking population changes [14]. While records are for individuals, they can be linked to identify households [15] enabling household-level analyses [16–20]. Hence primary care EHRs represent a source of information on populations and households which can be timelier than the decennial Census.

Our overall objective was to assess whether primary care EHRs can be used to estimate populations and households reliably to support planning, business intelligence and research. We derived population and household estimates from the primary care EHRs for an ethnically diverse and disadvantaged population in north-east London (NEL) and compared these with published aggregate estimates from the Census 2021 and to administrative multisource estimates produced by the ONS for the same geographies. We examined the level of agreement between the EHR and Census 2021 aggregate estimates of population and residential household counts with and without exclusion of patients, without a record of a recent consultation with their general practitioner (GP) in the primary care EHR. We also compared primary care EHR population and household estimates with similar estimates obtained from the Statistical Population Dataset (SPD) [7] and Admin

Based Occupied Address Dataset (ABOARD) [21] for equivalent geographies.

Methods

Study design

Cross-sectional study

Data sources

Primary care electronic health records

We downloaded pseudonymised primary care EHRs from the Discovery Data Service (DDS) [22], which receives patient-level electronic health records in near-real time from all NEL general practices, for patients registered on 21st March 2021 with all general practitioners (GPs) in Barking and Dagenham, Hackney, Havering, Newham, Redbridge, Tower Hamlets, and Waltham Forest, across seven local authorities in NEL. Data were downloaded on 28th February 2024.

Office for National Statistics Census 2021

We used the Census 2021 population estimates [23] of usual residents by sex and by single year of age [24], and the Census 2021 household number and size estimates [25], both published at a local authority and Lower layer Super Output Area (LSOA) geographies [26].

The Census data was collected by the Office for National Statistics (ONS) with 89% people invited to complete an online questionnaire and 11% a paper version. The AddressBase Premium gazetteer [27], a system storing all official addresses in Great Britain, was used to ensure all households in England and Wales were invited to take part in the Census. The data were collected by ONS. Results of the Census Coverage Survey, a short separate survey conducted eight weeks after the Census, were used to identify and adjust for the number of people and households not counted, counted multiple times or counted in the wrong place. Full details of the methods used to derive the estimates are given in the “Quality and methodology information (QMI) for Census 2021” report [28]. Data were downloaded on 12th May 2023.

Linked administrative data

We used population estimates from the Statistical Population Dataset (SPD) [7], a dataset obtained through linking administrative records. The SPD was created to develop methods to derive the estimate of the usual resident population for small area geographies more frequently than the current national decennial Census. It links information at the person level from multiple datasets including those held by the NHS, the Department for Work and Pensions, the Higher Education Statistics Agency (HESA), and the Department for Education (full list in Appendix 1, section 11.1), and uses a set of inclusion rules, including appearance in a core administrative dataset and recent activity in one of the datasets. The SPD is produced independently every year, which reduces errors from previous years being carried forward [7, 29]. We

used SPD Version 4.0 by sex and age, with a reference date of 30th June 2021 [30]. More detail on these rules can be found in “Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016”, “Developing Statistical Population Datasets, England and Wales: 2021” and “Statistical Population Dataset version 4: Research to Date and Future Developments” [7, 29, 31]. Data were downloaded on 16th June 2023.

We used the Admin-Based Occupied Address Dataset (ABOARD) [21], a dataset produced by the ONS. It links the SPD which provides the base population, to the NHS Patient Demographic Service (PDS) and the English School Census, which provides Unique Property Reference Numbers (UPRNs) to assign people to a residential address, and to remove those without a UPRN [21]. It uses data from the Ministry of Justice and HESA to remove people living in communal establishments such as prisons or educational institutions. It estimates the number of residents and the number of occupied addresses in England and Wales [21]. We used the ABOARD dataset created for the year 2021, available at local authority level and by household size [8]. Data were downloaded on 5th December 2023.

Except for primary care EHRs, these datasets are published and openly accessible. We downloaded estimates for England and Wales and restricted datasets to the same geographic areas as for the primary care EHRs.

The population and household definitions by data source are summarised in Table 1.

Study sample

We identified 2,236,001 patients from the primary care EHR who, on 21st March 2021, were currently fully registered with one of the GPs in the selected geographical areas. We excluded people who had died before or were more than 100 years old on 21st March 2021, those residing outside the selected geographic area, and those without a high-quality pseudonymised unique address identifier RALF (Residential Anonymised Linkage Field) or residing in commercial properties. This resulted in a population estimate of 2,130,965 people in residential households or communal establishments for comparison with other data sources. This sample of 2,130,965 people was used to derive the population estimates and for all analyses comparing EHR population estimates with those from Census 2021 [1, 2] and Statistical Population Dataset [7] population estimates.

We created the second sample used for the analysis of households. This was done by excluding people living in communal establishments as those do not count as residential households according to the household definitions in all data sources.

We used Unique Property Reference Numbers (UPRN) [21], assigned to the GP-recorded alphanumeric addresses using the ASSIGN (Address MatchInG to Unique Property Reference Numbers) address-matching algorithm [13] and pseudonymised with Open Pseudonymiser [23] into Residential Anonymised Linkage Fields (RALFs) [37] using a study-specific SALT or encryption key. We identified residential households by linking people with the same RALF derived from their registered patient address. We excluded people

living in communal establishments. We included all households irrespective of their size. The final primary care residential households sample comprised 658,264 households containing 2,108,818 people (Figure 1). This sample was used to derive the household estimates and for all analyses comparing the estimates of the households obtained from EHR with those from Census 2021 [1, 2] and Admin-based occupied address dataset [21].

Primary care electronic health record-derived population and household covariates

We categorised the ages of household members on 21st March 2021 in whole years by ten-year intervals, and – for the Bland-Altman method [38, 39] – five-year intervals. We defined sex as a categorical binary variable (male, female) using the recorded value in the EHR. We categorised household sizes in two ways: firstly, to align with ONS Census 2021 definitions (sizes of one to seven with increments by one, and eight or more people), and secondly, to align with ONS ABOARD definitions (sizes one to four with increments by one, and five or more people). We used the 2019 Index of Multiple Deprivation (IMD) [40], the official measure of relative deprivation in England, produced for LSOA-level geographies [41] to categorise households into five quintiles, with quintile 1 being the most, and quintile 5 the least, deprived. We categorised people according to the presence of long-term conditions recorded in the primary care EHR (Supplementary Table S2). A borough/local authority is an administrative unit for local government with the following seven categories: Barking and Dagenham, Hackney, Havering, Newham, Redbridge, Tower Hamlets, and Waltham Forest.

Statistical analyses

We compared EHR estimates of the population of people in residential households and communal establishments, ($n = 2,130,965$ people; Figure 1), with Census 2021 population estimates by sex, age, local authority, and IMD quintile, and EHR household residential household estimates ($n = 2,108,818$ people in 658,264 households; Figure 1), with Census 2021 household estimates by number, size, and local authority. We defined an overcount and undercount in populations and households when the EHR, SPD or ABOARD estimates were respectively greater or less than the estimates derived from the Census. We estimated the 95% Limits of Agreement (LoA) between EHR and Census 2021 household and population estimates using the Bland and Altman method [38, 39]. This method is used to assess the level of agreement between two different methods of measurement, when the true values are unknown. It calculates the mean difference and standard deviations of the difference, which gives the range within which 95% of differences are expected to lie. We estimated the 95% LoA for a log-transformed number of people and a number of households by Middle layer Super Output Area (MSOA) – a geographical boundary representing between 2,000 and 6,000 households and a resident population between 5,000 and 15,000 persons [42] – and for log-transformed number of females and males by age group.

We performed sensitivity analyses by excluding people with no record of a recent clinical encounter with their GP 12,

Table 1: Population and Household definitions by data source

Data source	Population definition	Household definition
ONS Census 2021	The usual resident population, defined as <i>“anyone who on Census Day, 21 March 2021 was in the UK and had stayed or intended to stay in the UK for a period of 12 months or more, or had a permanent UK address and was outside the UK and intended to be outside the UK for less than 12 months”</i> [32]. The Census enumeration is address based, with people counted in their place of usual residence, permanent address or residence where they spend most of their time [33]. The enumerated people include residents of households or communal establishments, or people sleeping rough [33]. Non-residential addresses are not included in Census enumeration [33].	Household estimates published from Census 2021 include households defined as <i>“one person living alone, or a group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room, or dining area”</i> [32]. Sheltered accommodation and caravans serving as usual residence are included in household estimates [32], but people living in managed residential accommodation in a communal establishment do not count as household residents [33].
Primary care electronic health records	Patients who on 21 st March 2021 were fully registered with one of the general practices in the selected geographical areas ¹ . People living in non-residential properties were not enumerated. See supplementary file, Table S1 for further details of residential properties.	A household is defined as a group of individuals who on 21 st March 2021 shared the same pseudonymised UPRNs ² derived from their registered patient address of residential properties, including sheltered accommodation.
Statistical population dataset	Estimated population of usual residents of England and Wales [7, 34] from multiple linked administrative datasets. Enumerates are the people active during the 12 months prior to the reference date – 30 th June 2021 in one or more sources in linked datasets [31].	N/A
Admin-based occupied address dataset	People included in the Statistical Population Dataset, linked to addresses using the address identifiers – UPRNs ² , excluding those for whom UPRNs ² were not found, and those residing in communal establishments including prisons and educational institutions, and other non-household addresses [21].	Occupied addresses are defined as residential addresses with at least one usual resident. Usual residents are assigned from the administrative population dataset – Statistical Population Dataset [7] to addresses using the address identifiers – UPRN ² obtained from the Personal Demographic Service [35] and English School Census [36]. Properties for which UPRNs were not found, communal establishments and other non-household addresses are excluded.

¹Local authorities in north-east London: Barking and Dagenham, Hackney, Havering, Newham, Redbridge, Tower Hamlets, and Waltham Forest.

²UPRN Unique Property Reference Number Study sample.

36, or 60 months before 21st March 2021, and examined the effect on levels of agreement in population estimates by age. Clinical encounters included GP consultations (in person, telephone, video and email), clinic consultations, administrative encounters (notes, letters, documents, reports and other), case reviews and repeat prescriptions. This included consultations during the COVID-19 pandemic. We tested the relationship between the clinical activity in the last 12, 36 and 60 months and a number of long-term conditions (LTCs).

We compared EHR, SPD and Census population estimates by sex and age, and EHR, ABOAD and Census household estimates by local authority and household size.

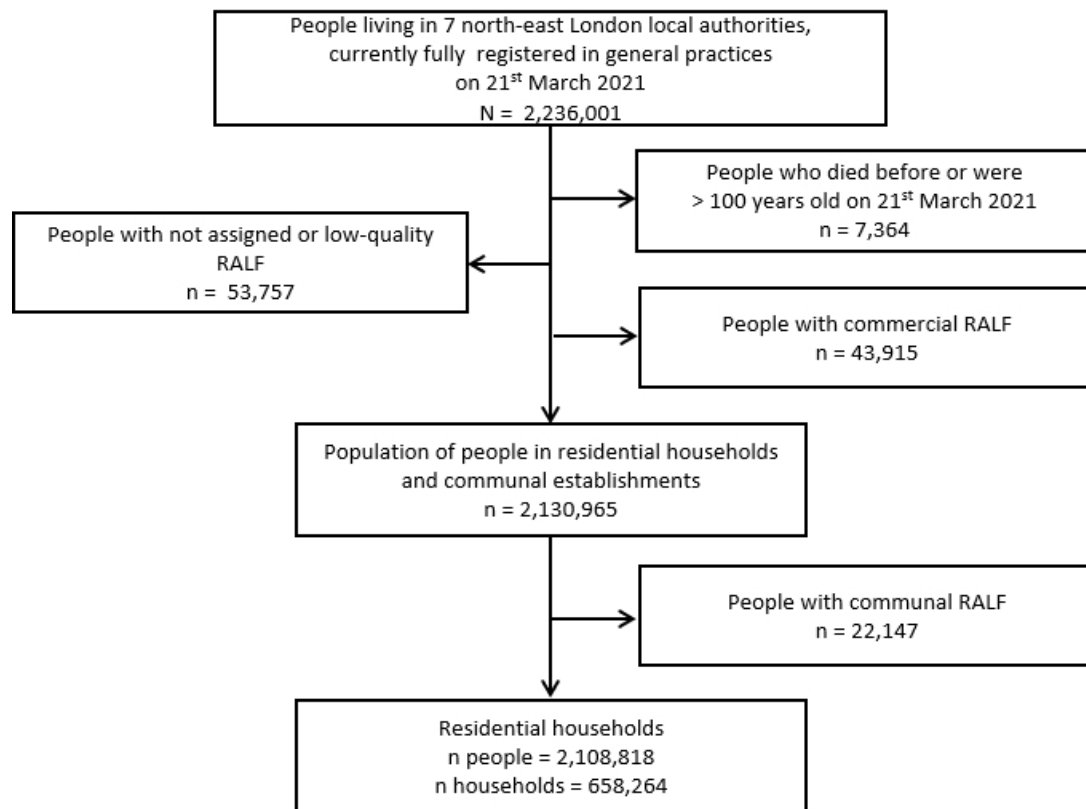
Results

Population estimates

Overall, the population estimate of people living in residential or communal establishments in NEL derived from primary care EHRs was 2,130,965, 7.1% greater than the Census 2021 estimate of 1,990,087 people (Figure 2). The difference in estimates varied by local authority, being lowest in Havering (2.2%) and greatest in Newham (13.2%).

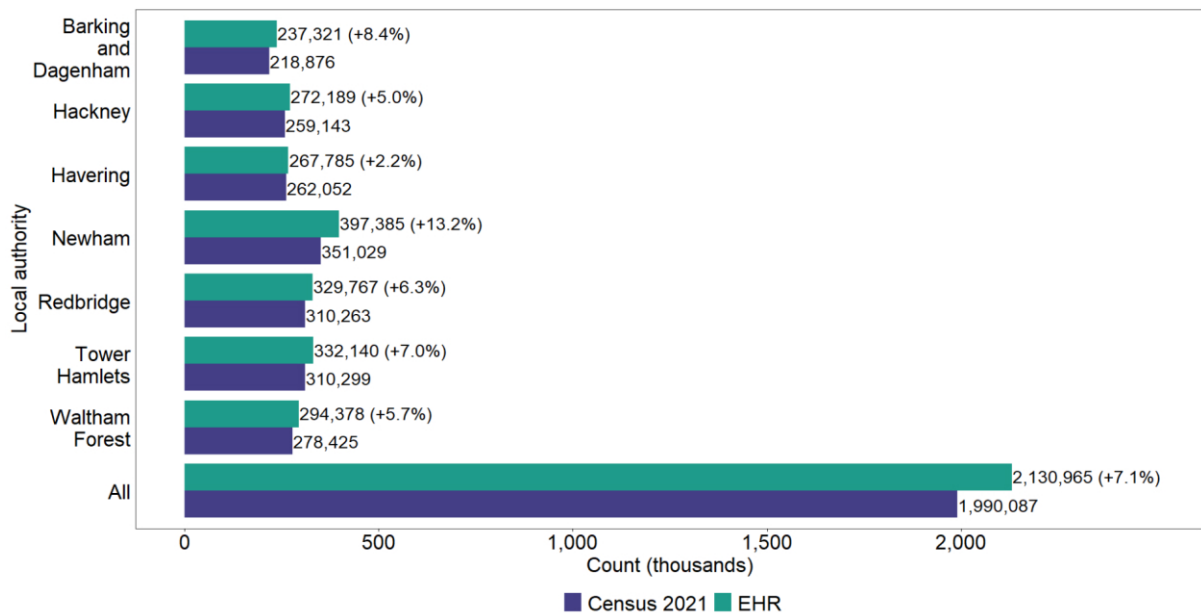
Overall, agreement in population estimates was closer for the female (3.6%) than for the male population (10.7%) (Figure 3, Table S3, Appendix 1). This varied by age group,

Figure 1: Flow diagram showing the derivation of the study population and household sample



RALF – Residential Anonymised Linkage Fields.

Figure 2: Population size estimates obtained from north-east London primary care electronic health records (EHR) and Census 2021 for the whole study sample and by local authority



Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

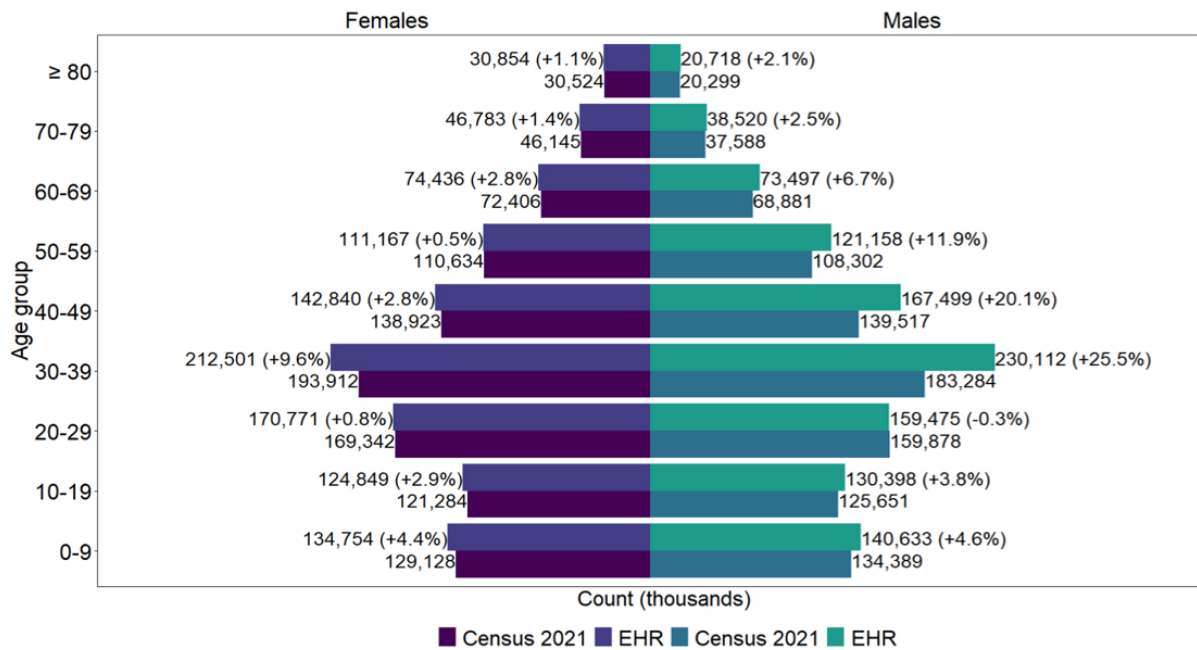
with people aged 30-39 years of either sex showing the greatest (female: 9.6%; male: 25.5%), and people aged 20-29 years the smallest, difference (female: 0.8%; male: -0.3%) (Figure 3).

Overall, the difference between EHR and Census 2021 estimates lay between -3% and 9% for 95% of the age groups

for the female population and -8% and 27% for 95% of the age groups for the male population, using the Bland and Altman method (see Appendix Figures S1a and S1b).

Agreement between EHR and Census population estimates varied by deprivation quintile, with the greatest difference seen

Figure 3: Comparison of the population size obtained from north-east London primary care electronic health records (EHR) and Census 2021 estimates by sex and age group



Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

for the population living in the two most deprived quintiles, where the majority (68%) based on the EHR population estimate lived (Figure 4). In 95% of MSOAs, the difference between Census 2021 and EHR estimates lay between -9.3% and +24.0% (Figure S2).

Household estimates

EHR of residential households estimates were on average 9.1% lower than Census 2021 estimates, ranging from -6.3% in Barking and Dagenham to -10.9% in Hackney (Figure 5).

The majority (78%) of EHR-derived households comprised fewer than four members, with 3.0% and 4.7% respectively being households of seven, or eight or more, members. EHR estimates of household size were lower than Census 2021 estimates across all household sizes for households of four or less, and greater than Census estimates for households with five or more, members (Figure 6), with the greatest percentage overcount observed for household sizes of seven and eight or more members. This was consistent across all local authorities (see Appendix Figure S4).

Sensitivity analyses

The population estimate from EHR was on average 11.6% lower than the Census 2021 estimate when people with no record of a recent clinical encounter with their GP in the 12 months prior to 21st March 2021 were excluded (Figure 7). This was most marked in the 10-19 and 20-29 year age groups. EHR population counts were slightly higher than Census 2021 estimates when excluding people with no clinical activity in the preceding 36 (1.3%) or 60 months (4.3%), with less marked variation by age (Figure 7, Appendix Table S4).

A record of a recent clinical encounter with a GP in the 12, 36, or 60 months prior to 21st March 2021 was much higher for adult patients (≥ 18 years) with a long-term condition than those without, suggesting that exclusion of people without a record of a recent clinical encounter would exclude many patients without long term conditions. (Figure 8).

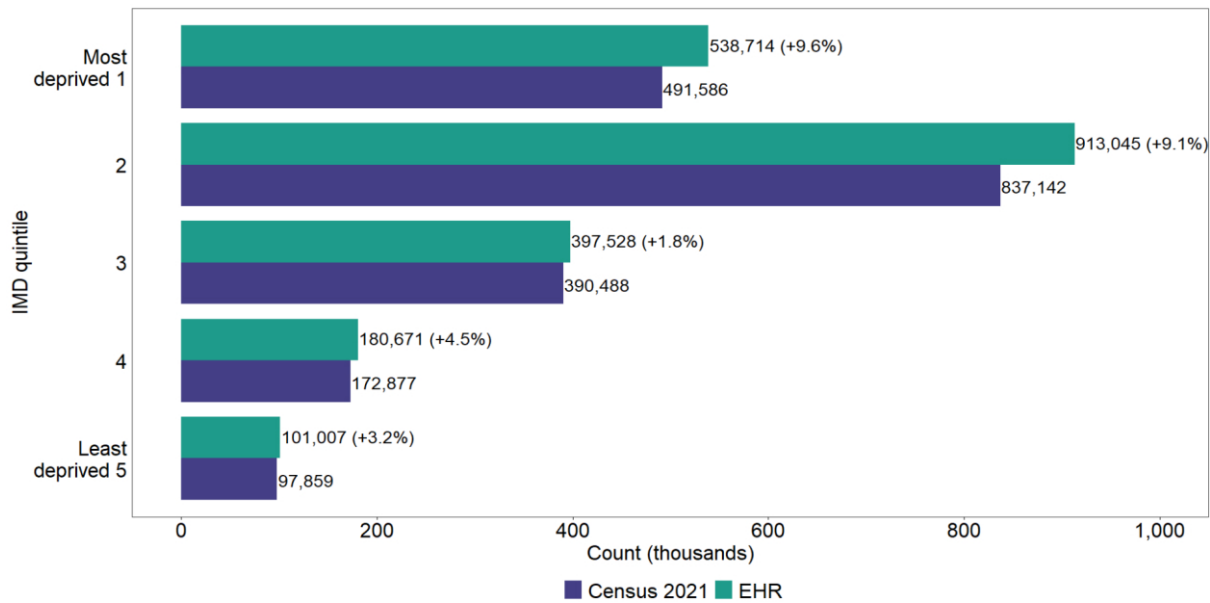
Comparison of electronic health record and Statistical Population Dataset population estimates with Census 2021 estimates

The population estimates derived from EHRs were similar to those derived from SPD and both were higher than the Census 2021 population estimates overall (Figures 9 and 10, Supplementary Table S3) and higher for males than for females (males: 10.7%, 7.9% respectively; females: 3.6%, 2.7% respectively).

Comparison of electronic health record and Admin-based Occupied Address Dataset estimates of household numbers with Census 2021

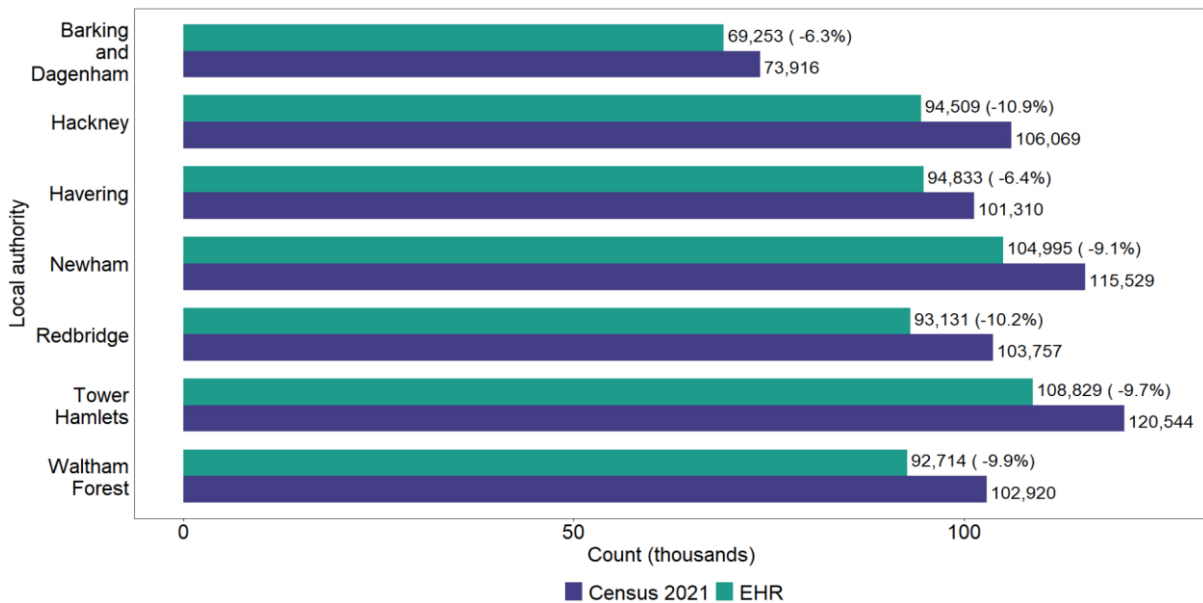
Both EHR and ABOAD demonstrated similar levels of agreement with Census 2021 in the number of estimated households/occupied addresses (ABOAD -7.3%; EHR -9.1%) (Figure 11). Relative to the Census, both ABOAD and EHR household sizes showed a similar pattern of undercounting households of four people or less and overcounting households of five people or more (Figure 12, Supplementary Table S5).

Figure 4: Comparison of the population size obtained from north-east London primary care electronic health records (EHR) and Census 2021 estimates by Index of Multiple Deprivation (IMD) quintile



Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

Figure 5: Estimates of household number obtained from north-east London primary care electronic health records (EHR) and Census 2021



Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

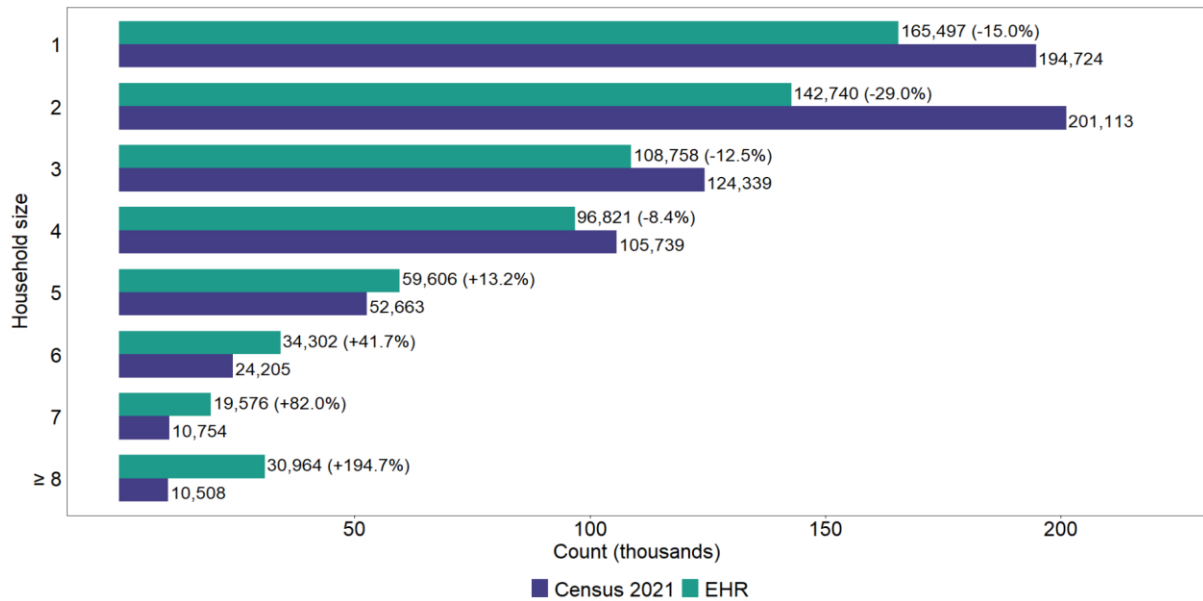
Discussion

Summary of main findings

We developed a dynamic method of deriving population and household estimates from primary care EHRs in an urban, multi-ethnic and disadvantaged London region. Population estimates derived from EHRs on the Census date were 7.1% higher than those derived from the Census, and this was most marked for men aged 30-39 years. The highest agreement between estimates from EHRs and the Census was in the people aged 20-29 years, but this close agreement

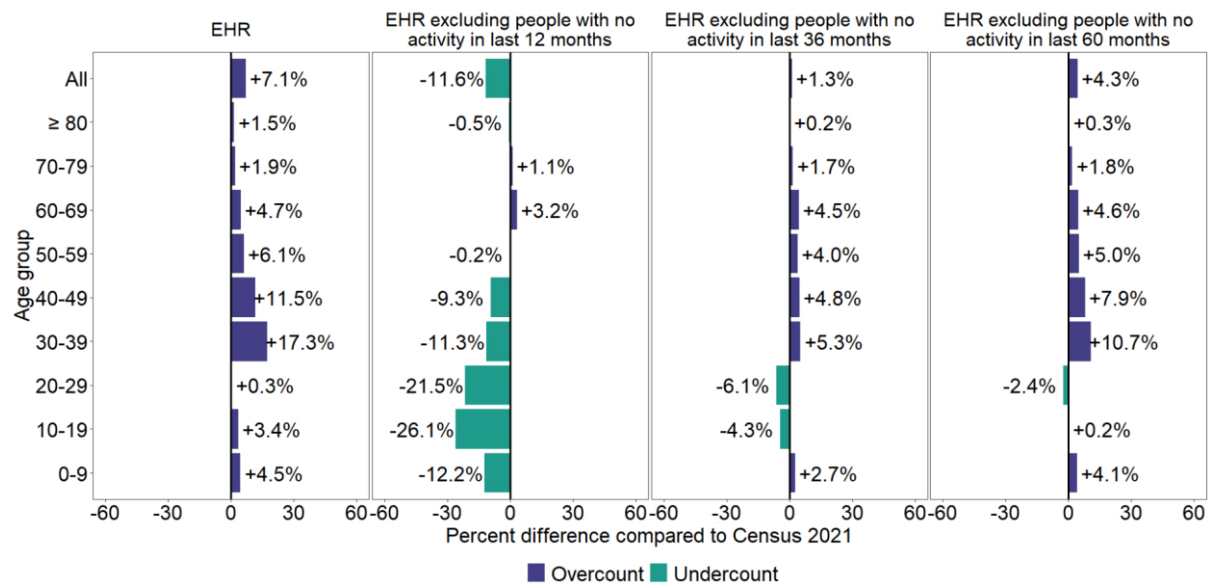
may be partially attributed to the overcount of people with incorrect GP registrations (e.g. those who left the area without registering), being balanced out but the undercount of local residents who did not register at GP as they did not need to use GP services. Overall, there was better agreement in the estimated female than male population. In this highly disadvantaged region, EHR population estimates were closer to Census estimates in the less deprived areas. Exclusion of people with no recorded clinical activity in the EHR did not reduce these observed differences. In contrast, household estimates derived from EHRs were around 9% lower than those derived from the Census, especially for households of four

Figure 6: Estimates of household number by household size obtained from north-east London primary care electronic health records (EHR) and Census 2021



Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

Figure 7: Percent difference between the population size obtained from north-east London primary care electronic health records (EHR) baseline dataset and datasets excluding people with no record of a recent clinical encounter within 12, 36 or 60 months and Census 2021 estimates by age group



Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

or fewer, which comprise the majority of households in this region.

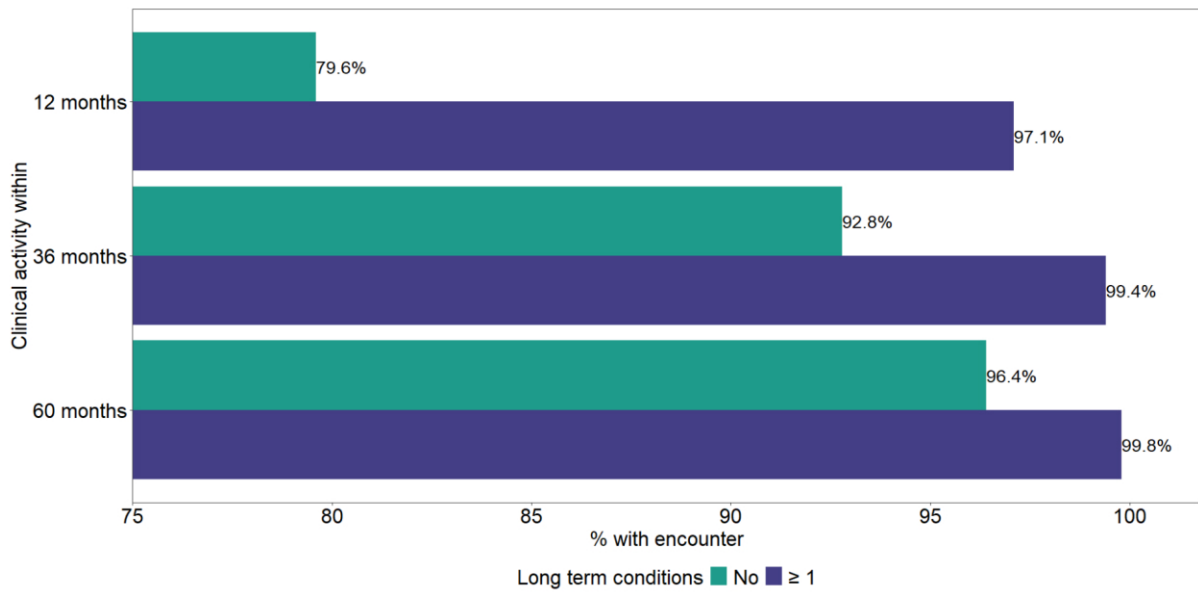
There has been a considerable effort to evaluate the accuracy of population and household estimates derived from administrative data as a supplement for the ONS decennial Census which is less frequent and more costly. Our relatively simple method of estimating populations and households produced estimates similar to those derived from the SPD and ABOAD datasets both of which include linked data from

multiple administrative datasets, including from the NHS PDS which records addresses of all NHS registered patients in England.

Comparison with other estimates

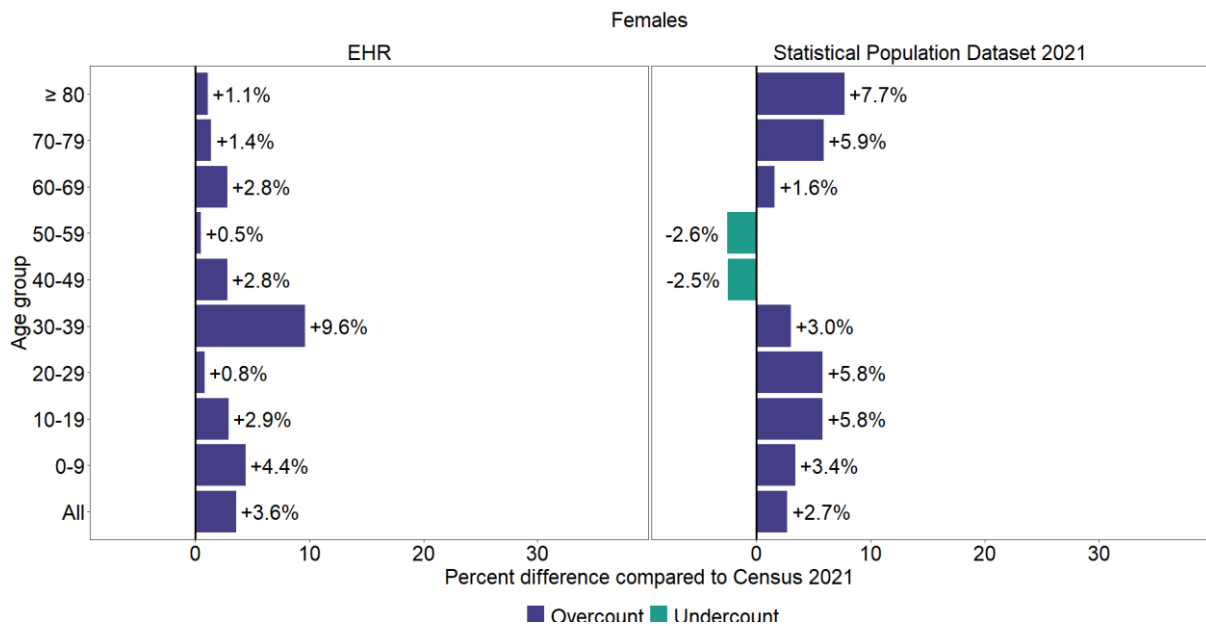
We compared the population counts obtained in this research to the estimates produced on 1st April 2021 from the primary care registration database held by the NHS Application and

Figure 8: Percentage of patients registered at the general practices (GP) with record of a clinical encounter 12, 36 and 60 months prior to 21st March 2021 by the presence or absence of long term conditions



Percentages are $((\text{EHR estimates} - \text{EHR estimates of active patients}) / \text{EHR estimates}) \times 100$. EHR – electronic health records.

Figure 9: Percent difference between the female population size obtained from north-east London primary care electronic health records (EHR) and SPD 2021 and Census 2021 estimates by age group

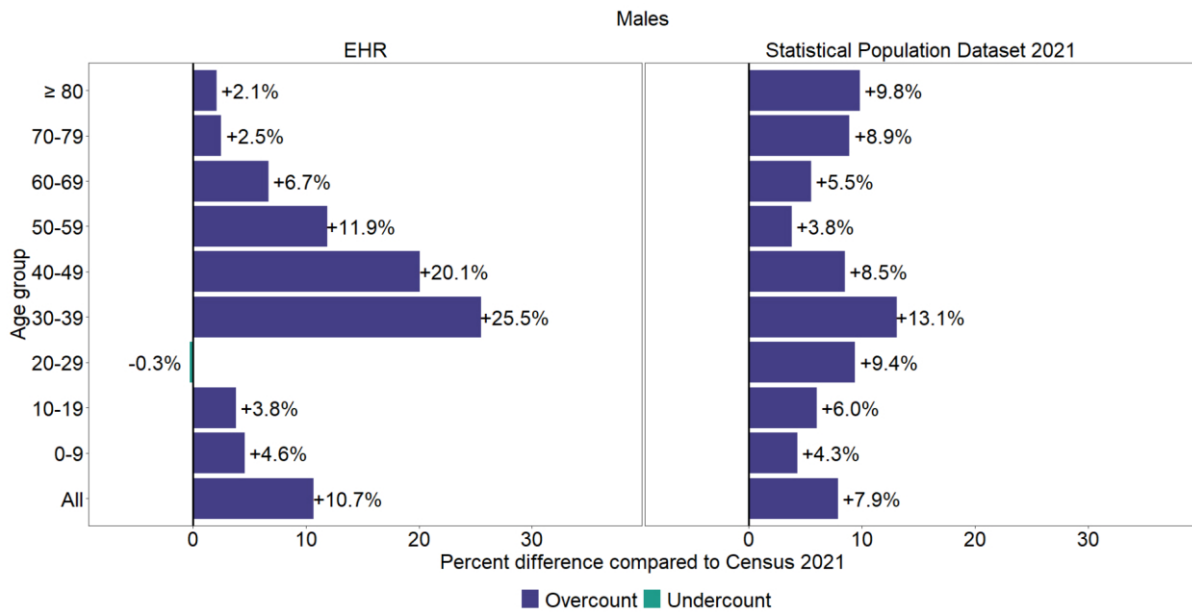


Percentages are $(\text{EHR or SPD estimates} - \text{Census estimates}) / \text{Census estimates} \times 100$.

Infrastructure Services system (NHAIS) [43]. This system, which is being decommissioned in 2024, holds a snapshot of patients registered with GPs for each NHS region in England and is used primarily to process payments to GPs based on list size. While quality-assured, it is recognised that its accuracy is impacted by many factors including NHS reorganisations, such as occurred in April 2021, and the impact of the recent COVID-19 pandemic on patient registrations and patient list updating by GPs [44]. In 2021 our EHR population estimate of 2,130,965 people (Figure 1), was overall 7.6% lower than that of 2,304,999 people derived

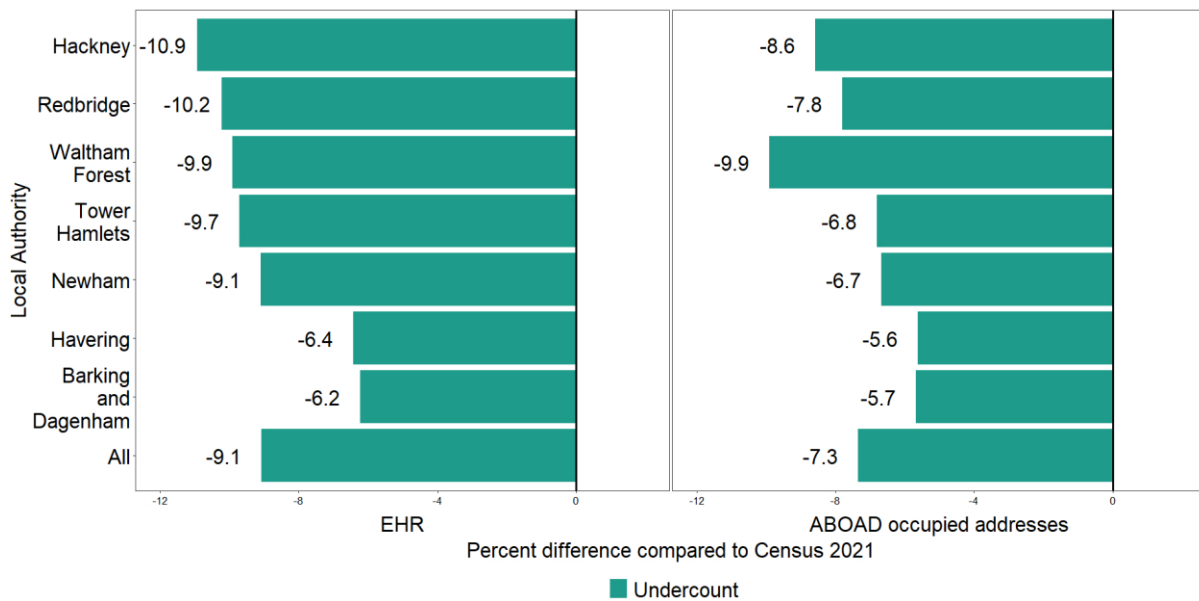
from the NHAIS dataset provided in the Spotlight report [45] for the equivalent geography, but much closer (−3.0%) when comparing the number of currently fully registered patients by the local authority (2,236,001 patients compared to 2,304,999 respectively). The NHAIS database has also been compared to Census 2021 estimates, which suggested a much higher discrepancy of (−15.8%) between the Census 2021 and NHS list with respective derived populations of 1,990,200 and 2,304,999 for the NEL area [45]. Much of this discrepancy was attributed in this report to GP list inflation related to patients moving out of the area and failing to

Figure 10: Percent difference between the male population size obtained from north-east London primary care electronic health records (EHR) and SPD 2021 and Census 2021 estimates by age group



Percentages are $(\text{EHR or SPD estimates} - \text{Census estimates}) / \text{Census estimates} \times 100$.

Figure 11: Percent difference between the number of households estimated from electronic health records (EHR) and admin-based occupied addresses (ABOAD) and Census 2021 by local authority



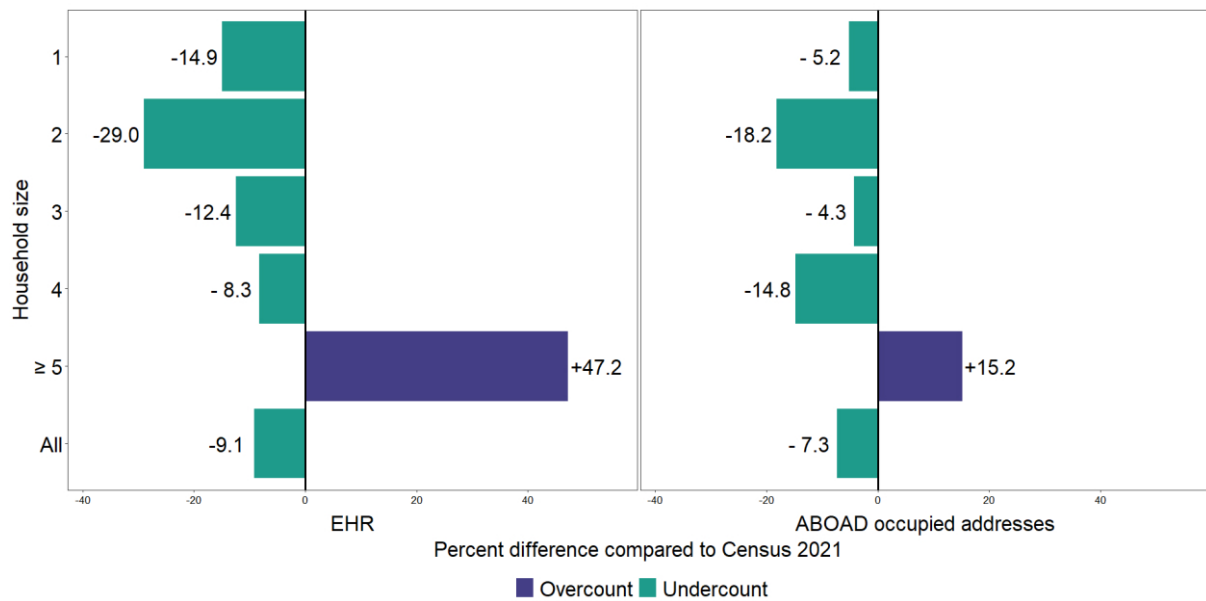
The values in EHR and Census 2021 estimates were rounded to a hundred in line with ABOAD published values. Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

de-register or be de-registered by their GPs. We were able to examine the effect of recency of clinical activity on population estimates and did not find these improved levels of agreement with the Census estimates to any extent. Furthermore, we took care to define residential households and their populations as closely as possible to the definitions used in the Census extracts. This may account for differences between the EHR estimates and Census being smaller than those found by the Spotlight report (7.1% versus 15.8% overcount respectively).

Interpretation

There are several factors in addition to those mentioned in the previous section which could affect the extent to which our population and household estimates for this region agreed with those reported by the Census 2021. These arise from methodological differences in defining the eligible populations and households, as well as wider sociodemographic factors affecting population stability and inclusion in the respective datasets.

Figure 12: Percent difference between the male population size obtained from north-east London primary care electronic health records (EHR) and SPD 2021 and Census 2021 estimates by age group



The values in EHR and Census 2021 estimates were rounded to a hundred in line with ABOAD published values. Percentages are $((\text{EHR estimates} - \text{Census estimates}) / \text{Census estimates}) \times 100$.

The level of the alignment of the GP-registered population with the residential population for a given geography depends on the proportions of people who do or do not live in that geography and are registered or not registered with a GP providing services in that geography. We were careful to include only patients who resided in the NEL region and excluded those with addresses outside of the region, though we did not have information on those living in the region and using GP services outside of NEL.

Differences may also arise due to the lack of comparable definitions of populations and households in the different data sources. Census 2021 data are collected primarily to obtain population statistics while patient registrations are recorded in the EHR to provide health care. Given these differences, the Census 2021 definitions of population and households cannot be directly applied to EHR data. While we were not able to test the intended length of stay at the address to confirm the usual residence rule within the EHR database, we excluded patients with temporary registrations. We were also unable to identify rough sleepers from the primary care data.

The Census 2021 definition of a household as “one person living alone, or a group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room, or dining area” cannot be verified in the EHR. We identified households by assuming that people who share the same RALF at the same point in time comprise a household. This in turn depends on the ability to assign a UPRN to the recorded patient address. The primary care records in the Discovery Data Service (DDS) allocate UPRNs in real-time using the validated ASSIGN algorithm [46]. We excluded 2.4% of patients (53,757 out of 2,236,001) in our population due to no or a poor quality match of their address to a UPRN (Figure 1).

Differences may also arise as a consequence of non-participation in the Census and statistical assumptions and methods used to adjust for non-response. Almost all local authorities in NEL had Census 2021 response rates lower than the average for England and Wales (97%) and three had response rates lower than the average for London (95%) including Hackney (93%), Newham (94%) and Tower Hamlets (94%), with levels of response especially low in younger people aged 20-24, and in Tower Hamlets (90%), in Redbridge (93%) and in Hackney (91%) [47, 48]. There was also uncertainty about the number of households estimated from the Census 2021 in Tower Hamlets, Barking and Dagenham and Waltham Forest, with higher estimates than expected when compared to alternative sources [49]. Residential mobility might explain some of the variation in levels of agreement by the local authority observed in our study. Data from the Trust for London suggests that the inner east London boroughs had among the highest levels of residential mobility in London between 2015 and 2020 [50].

Our findings when we compared EHR estimates to those derived from multisource ONS administrative datasets showed similar levels of agreement with Census 2021. Earlier work combining local administrative datasets with EHRs suggests that multiple administrative data sources may improve enumeration [14]. However, this requires complex linkage which may introduce errors or biases due to linkage and is more expensive than a single source. These biases may arise from issues in the data or the matching algorithms and may result in missed or false matches, or over- and under-representation of some groups in linked data. These biases may be hard to measure and, even when quantified, may not provide information on how linkage bias affects the analyses using linked data [51, 52]. Deriving the estimates from the single EHR source eliminates the risk of linkage bias. A further

advantage of the single EHR source is that only pseudonymised identifiers are used to identify individuals and households.

Local authorities need to have up-to-date estimates of their populations and households for planning purposes. For example, between the Census 2011 and Census 2021, the population of Tower Hamlets increased by 56,200, an increase of 22.1% [53]. Primary care EHRs provide a suitable, simple and dynamic method to obtain these estimates between decennial Census dates with a high level of geographic and demographic granularity. As EHRs also include health information, estimates derived from EHRs can provide detailed information on the health status of populations for local authorities and researchers. The method is cost-effective as it utilises already collected pseudonymised data, is simple to use, and can be used to produce estimates in near-real time, including information on household age, sex and ethnic composition. Regional primary care datasets such as the DDS have the potential to be used to create such estimates for local authorities and researchers.

Strengths and limitations

Strengths of our study include the use of a high-quality, validated algorithm [15] to allocate UPRNs to patient addresses, enabling encryption and safe use of address data for research without compromising patient privacy and confidentiality. We used a published and transparent method to identify household members at a point in time through the shared address identifier RALF [54]. We compared derived population and household estimates in an ethnically diverse mobile population with high levels of deprivation and lower than average Census participation which may limit generalisability to other geographies. We tested the effects of list inflation through sensitivity analyses. Our assessment of patients' engagement with general practice included administrative encounters (for example letters) which could be associated with patients who are still registered with the GP but who no longer reside in the area. However, we were unable to disaggregate these from face-to-face phone or video consultations. While we could not create definitions of populations and households that were entirely comparable to those used in the Census, we excluded patients who were temporarily registered and excluded patients with low quality or no UPRN matches. We were only able to compare our estimates with Census estimates using aggregated numbers rather than on an individual person-level basis, limiting a more granular analysis of the differences between the two sources of data.

Implications for research and practice

Further analyses to compare the Census and EHR estimates using individual-level data are planned as part of the ADRUK-funded Healthy Households project [55]. These will overcome the limitations of using aggregate data and will include methods that estimate the household population size from independent sources. The generalisability of our findings in NEL using the Discovery Data Service could be tested in other regions with similar regional primary care datasets. This would enable lower cost dynamic population and household estimations for planning purposes.

Conclusions

Our study reporting methods to estimate populations and households from primary care EHRs contributes evidence to initiatives to complement the decennial Census by using single or multisource administrative data. As a single source of data, primary care EHRs compare well with more complex multisource methods. It highlights the complexities of deriving population and household counts from administrative data, especially in neighbourhoods with high residential mobility.

Authors contribution

Carol Dezateux obtained funding for the study.

Carol Dezateux and Marta Wilk conceptualised and designed the analyses.

Marta Wilk carried out the literature search, conducted the analyses, and generated tables and figures.

Marta Wilk and Carol Dezateux wrote and revised the initial manuscript.

Marta Wilk, Gill Harper, Nicola Firman, Chris Dibben, Rich Fry and Carol Dezateux contributed to the development of the methodology, and interpretation of analyses and reviewed and revised the final manuscript.

All authors were involved in writing the paper and had final approval of the submitted and published manuscript.

The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Acknowledgements

This work was funded by Barts Charity ref: MGU0419 and MGU0504. It was also supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council investment (part of UK Research and Innovation) [Grant number: ES/X00046X/1]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Statement on conflicts of interest

None declared

Ethics statement

Data access was approved by the NEL Discovery Programme Board on behalf of the data controllers of primary care EHRs. As only routinely acquired de-identified data were analysed no research ethics committee approval was required by the Health Research Authority. Access to general practice data is enabled by data sharing agreements between the DDS and GP data controllers. EPC data are publicly available.

Only aggregated patient data are reported in this study.

Data availability statement

The data controllers have not permitted onward sharing by the study authors of EHRs data used for this research. Office for National Statistics Census 2021, SPD and ABOAD datasets are publicly available.

References

- Office for National Statistics, 2022. About the census. [Available from: <https://www.ons.gov.uk/census/aboutcensus/aboutthecensus>].
- Office for National Statistics, 2022. Official census and labour market statistics. Census 2021. [Available from: https://www.nomisweb.co.uk/sources/census_2021_ts].
- Office for national Statistics, 2023. Design for Census 2021. [Available from: <https://www.ons.gov.uk/census/planningforcensus2021/censusdesign/designforcensus2021>].
- Office for National Statistics, 2016. Methodology used to produce the national population projections. [Available from: [https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies/methodologyusedtoproducethenationalpopulationprojections#:~:text=Summary%20of%20cohort%20component%20method,-Population%20\(year%20x&text=For%20each%20age%2C%20the%20starting,those%20born%20during%20the%20year](https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies/methodologyusedtoproducethenationalpopulationprojections#:~:text=Summary%20of%20cohort%20component%20method,-Population%20(year%20x&text=For%20each%20age%2C%20the%20starting,those%20born%20during%20the%20year)].
- Office for National Statistics, 2022. Estimates of the population for the UK, England, Wales, Scotland and Northern Ireland. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>].
- Office for National Statistics, 2023. Rebasing of mid-year population estimates following Census 2021, England and Wales. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/rebasingofmidyearpopulationestimatesfollowing/rebasingofmidyearpopulationestimatesfollowingcensus2021englandandwales>].
- Office for National Statistics, 2023. Developing Statistical Population Datasets, England and Wales: 2021. [Available from: [https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingstatisticalpopulationdatasetsenglandandwales/2021#:~:text=The%20Statistical%20Population%20Dataset%20\(SPD\)%20aims%20to%20approximate%20the%20usually,small%20areas%20with%20admin%20data](https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingstatisticalpopulationdatasetsenglandandwales/2021#:~:text=The%20Statistical%20Population%20Dataset%20(SPD)%20aims%20to%20approximate%20the%20usually,small%20areas%20with%20admin%20data)].
- Office for National Statistics, 2023. Population and migration statistics transformation, occupied addresses, England and Wales. [cited December 2023]. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationandmigrationstatisticstransformationoccupiedaddressesenglandandwales>].
- Central Statistics Office, 2023. Irish Population Estimates from Administrative Data Sources, 2021. [cited Nomeber 2023]. Available from: <https://www.cso.ie/en/releasesandpublications/fp/fp-ipeads/irishpopulationestimatesfromadministrativedatasources2021/#:~:text=Based%20on%20the%20Irish%20Population,migration%20patterns%20over%20recent%20decades>].
- United States Census Bureau, 2023. Real-Time 2020 Administrative Record Census Simulation. [Available from: <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/evaluate/eae/2020-admin-record-census-simulation.html>].
- Australian Bureau of Statistics, 2021. Administrative data snapshot of population. [Available from: <https://www.abs.gov.au/statistics/people/population/administrative-data-snapshot-population-and-housing-experimental-population-data/latest-release#:~:text=National,were%2065%20years%20and%20over>].
- Stats NZ, 2023. Experimental administrative population census. [Available from: <https://www.stats.govt.nz/experimental/experimental-administrative-population-census/>].
- NHS England, 2023. Purpose of the GP electronic health record. [Available from: <https://www.england.nhs.uk/long-read/purpose-of-the-gp-electronic-health-record/#uses-of-electronic-health-records>].
- Harper G, Mayhew L, 2015. Using Administrative Data to Count and Classify Households with Local Applications. *Applied Spatial Analysis and Policy*, 433-62. <https://doi.org/10.1007/s12061-015-9162-2>
- Harper G, Stables D, Simon P, Ahmed Z, Smith K, Robson J, et al., 2021. Evaluation of the ASSIGN open-source deterministic address-matching algorithm for allocating Unique Property Reference Numbers to general practitioner-recorded patient addresses. *International Journal of Population Data Science*. <https://doi.org/10.23889/ijpds.v6i1.1674>
- Forbes H, Morton CE, Bacon S, Mcdonald HI, Minassian C, Brown JP, et al., 2021. Association between living with children and outcomes from covid-19: OpenSAFELY cohort study of 12 million adults in England. *BMJ*, n628. <https://doi.org/10.1136/bmj.n628>
- Nafilyan V, Islam N, Ayoubkhani D, Gilles C, Katikireddi SV, Mathur R, et al., 2020. Ethnicity, Household Composition and COVID-19 Mortality: A National Linked Data Study. <https://doi.org/10.1101/2020.11.27.20238147>

18. Mikolai J, Keenan K, Kulu H, 2020. Intersecting household-level health and socio-economic vulnerabilities and the COVID-19 crisis: An analysis from the UK. *SSM - Population Health*, 100628. <https://doi.org/10.1016/j.ssmph.2020.100628>
19. Harper G, Lyons J, Fry R, Akbari A, Ahmed Z, Lyons R, et al., 2019. Quantifying multi-morbidity in an ethnically-diverse inner city population: the health burden of households. *International Journal of Population Data Science*. <https://doi.org/10.23889/ijpds.v4i3.1289>
20. Stafford M, Deeny SR, Dreyer K, Shand J, 2021. Multiple long-term conditions within households and use of health and social care: a retrospective cohort study. *BJGP Open*, BJGPO.2020.0134. <https://doi.org/10.3399/bjgpo.2020.0134>
21. Office for National Statistics, 2023. Population and migration statistics transformation in England and Wales, technical topic guide: 2023. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/populationandmigrationstatisticstransformationinenglandandwalestechnicaltopicguide2023#occupied-address>].
22. Discovery Health and Care Data Service, 2022. Discovery East London. [Available from: https://www.discoverydataservice.org/Content/Regional_programmes/Discovery_East_London/Discovery_East_London.htm].
23. Office for National Statistics, 2022. Population and household estimates, England and Wales: Census 2021. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationandhouseholdestimatesenglandandwales/census2021unroundeddata>].
24. Office for National Statistics, 2022. Official census and labour market statistics. Census 2021. TS009. [Available from: <https://www.nomisweb.co.uk/datasets/c2021ts009>].
25. Office for National Statistics, 2022. Official census and labour market statistics. Census 2021. TS017. [Available from: <https://www.nomisweb.co.uk/datasets/c2021ts017>].
26. Office for National Statistics, 2021. Census 2021 geographies. [Available from: [https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies#:~:text=3.-,Middle%20layer%20Super%20Output%20Areas%20\(MSOAs\),MSOAs%20fit%20within%20local%20authorities](https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies#:~:text=3.-,Middle%20layer%20Super%20Output%20Areas%20(MSOAs),MSOAs%20fit%20within%20local%20authorities)].
27. Ordnance Survey, 2021. AddressBase. Postal Addresses Matched to Unique Property Reference Numbers. Vector Map Data. [Available from: <https://www.ordnancesurvey.co.uk/business-government/products/addressbase>].
28. Office for National Statistics, 2023. Quality and methodology information (QMI) for Census 2021. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/qualityandmethodologyinformationqmiforcensus2021#methods-used-to-produce-the-data>].
29. Office for National Statistics, 2022. Statistical Population Dataset version 4: Research to Date and Future Developments. [Available from: <https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2023/05/EAP180-Statistical-Population-Dataset.pdf>].
30. Office for National Statistics, 2023. Statistical Population Dataset version 4.0 2021. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/datasets/statisticalpopulationdatasetversion402021>].
31. Office for National Statistics, 2019. Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21>].
32. Office for National Statistics, 2022. Household and resident characteristics, England and Wales: Census 2021. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/householdandresidentcharacteristicsenglandandwales/census2021>].
33. Office for National Statistics, 2021. Output and enumeration bases: residential address and population definitions for Census 2021. [Available from: <https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/outputandenumerationbasesresidentialaddressandpopulationdefinitionsforcensus2021#place-of-residence>].
34. Office for National Statistics, 2023. All data related to Developing Statistical Population Datasets, England and Wales: 2021. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingstatisticalpopulationdatasetsenglandandwales/2021/relateddata>].
35. Office for national Statistics, 2023. Personal Demographics Service data. [Available from: <https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/censusanddatacollectiontransformationprogramme/futureofpopulationandsocialstatistics/datasourceoverviews/personaldemographicsservicedata>].
36. UK Government, 2022. The School Census – what you need to know. [Available from:

- <https://educationhub.blog.gov.uk/2022/10/07/the-school-census-what-you-need-to-know/>].
37. Healthy Households, 2024. How it works: Household Analysis. [Available from: <https://healthyhouseholds.org.uk/>].
 38. Bland J, Martin, Altman D, G, 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 135-60. <https://doi.org/10.1177/096228029900800204>
 39. Bland M, J., Altman D, G, 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 307-10. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)
 40. Noble M, Wright G, Smith G, Dibben C, 2006. Measuring Multiple Deprivation at the Small-Area Level. *Environment and Planning A: Economy and Space*, 169-85. <https://doi.org/10.1068/a37168>
 41. Ministry of Housing Communities & Local Government. The English Indices of Deprivation 2019, Frequently Asked Questions. 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/853811/loD2019_FAQ_v4.pdf
 42. Office for National Statistics, 2022. Census 2021 geographies. [Available from: <https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies>].
 43. NHS Digital, 2021. Patients Registered at a GP Practice April 2021. [Available from: <https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/april-2021>].
 44. Office for National Statistics, 2016. Patient Register: quality assurance of administrative data used in population statistics, Dec 2016. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016>].
 45. NHS Digital, 2022. Comparing the number of patients registered with a GP Practice in England to the ONS Census 2021. [Available from: <https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/august-2022/spotlight-report-august-2022#list-size-inflation>].
 46. Harper G, Stables D, Simon P, Ahmed Z, Smith J, Robson JP, et al., 2021. Allocating Unique Property Reference Numbers to general practitioner recorded patient addresses using the ASSIGN deterministic address-matching algorithm : Cross-sectional evaluation in a large ethnically-diverse UK population. under review *International Journal of Population Data Science*.
 47. Office for National Statistics, 2022. Compare age-sex estimates from Census 2021 to areas within England and Wales. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/compareagesexestimatesfromcensus2021toareaswithinenglandandwales>].
 48. Office for National Statistics, 2023. Measures showing the quality of Census 2021 estimates. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/measuresshowingthequalityofcensus2021estimates>].
 49. Office for National Statistics, 2022. Maximising the quality of Census 2021 population estimates. [Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/maximisingthequalityofcensus2021populationestimates#collecting-the-data>].
 50. London Tf, 2024. Residential Mobility in London - recent and general trends. [Available from: <https://trustforlondon.org.uk/data/residential-mobility-london-recent-and-general-trends/>].
 51. Shaw RJ, Harron KL, Pescarini JM, Pinto Junior EP, Allik M, Siroky AN, et al., 2022. Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses. *European Journal of Epidemiology*, 1215-24. <https://doi.org/10.1007/s10654-022-00934-w>
 52. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al., 2017. Challenges in administrative data linkage for research. *Big Data & Society*, 205395171774567. <https://doi.org/10.1177/2053951717745678>
 53. Office for National Statistics, 2023. How life has changed in Tower Hamlets: Census 2021. [Available from: <https://www.ons.gov.uk/visualisations/censusareachanges/E09000030>].
 54. Harper G, Firman N, Wilk M, Marszalek M, Simon P, Stables D, et al., 2024. Determining households at a point in time from unique property reference numbers assigned to patient addresses recorded in general practitioner electronic health records. *International Journal of Population Data Science*. <https://doi.org/10.23889/ijpds.v9i1.2379>.
 55. Administrative Data Research UK, 2024. Healthy Households. [Available from: <https://www.adruk.org/our-work/browse-all-projects/healthy-households/#:~:text=This%20project%20will%20create%20a,researchers%20in%20using%20this%20data.>].
 56. NHS Digital, 2023. Quality and Outcomes Framework (QOF) business rules. [cited November 2023]. Available from: <https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/business-rules>].

Abbreviations

ABOAD:	Admin-based occupied address dataset	LSOA:	Lower Layer Super Output Area
ASSIGN:	AddreSS matchInG to unique property reference Numbers	MSOA:	Middle Layer Super Output Area
EHR:	electronic health records	NEL:	north-east London
EPC:	Energy Performance Certificates	ONS:	The Office for National Statistics
GP:	general practice/practitioner	RALF:	Residential Anonymised Linkage Field
		SPD:	Statistical Population Dataset
		UPRN:	Unique Property Reference Number



Supplementary appendix 1

Data sources used to develop the Statistical Population Dataset version 4 (SPDv4.2)

- Patient Demographics Service (PDS) – National Health Service
- Customer Information System (CIS) – Department for Work and Pensions
- Higher Education Statistics Agency (HESA)
- English School Census (ESC) – Department for Education
- Welsh School Census (WSC) – Welsh Government
- Pay As You Earn (PAYE) and Tax Credits data, National Benefits Database (NBD), Housing Benefit (SHBE) data, Child Benefit data – Department for Work and Pensions and HM Revenue & Customs
- Hospital Episode Statistics and Emergency Care Dataset (HES) – National Health Service
- Individualised Learner Record (ILR) - Department for Education [7, 29, 31]

Table S1: Codes and descriptions of Unique Property Residence Numbers (UPRNs) classified as residential addresses

UPRN* Code	UPRN* Description
R	Residential
RD	Dwelling
RD01	Caravan
RD02	Detached
RD03	Semi-Detached
RD04	Terraced
RD06	Self-Contained Flat (Includes Maisonette / Apartment)
RD07	House Boat
RD08	Sheltered Accommodation
RD10	Privately Owned Holiday Caravan / Chalet
RH02	HMO Bedsit / Other Non Self Contained Accommodation
U	Unclassified
UC	Awaiting Classification
UP	Pending Internal Investigation
X	Dual Use

* UPRN – Unique Property Reference Number.



Table S2: We identified sixteen long term conditions (LTCs): asthma, atrial fibrillation, cancer, chronic kidney disease, chronic obstructive pulmonary disease, coronary heart disease, dementia, depression, diabetes, epilepsy, heart failure, hypertension, learning disabilities, mental health, peripheral arterial disease and stroke or transient ischemic attack using standardised codesets defined in the Quality and Outcomes Framework [56] (Appendix, Table S2), and grouped conditions into the following clusters 1) asthma; 2) cancer; 3) chronic obstructive pulmonary disease; 4) mental health: (dementia, depression, epilepsy, learning disabilities, serious mental illness); 5) cardiometabolic conditions (atrial fibrillation, chronic kidney disease, coronary heart disease, heart failure, hypertension, diabetes, peripheral arterial disease, stroke, transient ischaemic attack). We derived a categorical variable to describe patients with none, one, or two or more LTCs clusters.

Cluster	Long term condition	Code description (code cluster) and extraction rule
Asthma	Asthma	Asthma code (AST_COD) excluding patients with a more recently resolved code (ASTRES_COD). Including Asthma Treatment medication code (ASTTRT_COD) recorded during last 12 months
Cancer	Cancer	Cancer code (CAN_COD) recorded on or after 2003-04-01
Chronic obstructive pulmonary disease	Chronic obstructive pulmonary disease	Chronic obstructive pulmonary disease code (COPD_COD) excluding patients with a more recent resolved code (COPDRES_COD). For patients who have a previous resolved code (COPDRES_COD), COPD code (COPD_COD) after the latest resolved code (COPDRES_COD).
Mental health	Dementia	Dementia code (DEM_COD)
	Depression	Patients aged ≥ 18 years, and Depression code (DEPR_COD) recorded on or after 2006-04-01 excluding patients with a more recent resolved code (DEPRES_COD)
	Epilepsy	Epilepsy code (EPIL_COD) excluding patients with a more recent resolved code (EPILRES_COD). Including Epilepsy Drug meds code (EPILDRUG_COD) recorded within last 6 months
	Learning disabilities	Learning Disabilities code (LD_COD)
	Serious Mental Illness	Mental Health code (MH_COD) excluding patients with a more recent remission code (MHREM_COD), and including Lithium meds code (LIT_COD) recorded within last 6 months excluding patients with a more recent Lithium Stopped code (LITSTP_COD) <i>This category combines the QOF Mental Health and QOF Prescribed Lithium registers.</i>
Cardiometabolic conditions	Atrial fibrillation	Atrial Fibrillation code (AFIB_COD) excluding patients with a more recently resolved code (AFIBRES_COD)
	Chronic kidney disease	Patients aged ≥ 18 years, with Chronic kidney disease code (CKD_COD) excluding patients with a more recent resolved code (CKDRES_COD) or a more recent CKD 1 or 2 code (CKD1AND2_COD)
	Coronary heart disease	Coronary heart disease code (CHD_COD)
	Diabetes	Patients aged ≥ 18 years, and Diabetes code (DM_COD) excluding patients with a more recent resolved code (DMRES_COD). Selecting most recent code to define type (DMTYPE1_COD, DMTYPE2_COD)
	Heart failure	Heart Failure code (HF_COD) excluding patients with a more recent resolved code (HFRES_COD)
	Hypertension	Hypertension code (HYP_COD) excluding patients with a more recent resolved code (HYPRES_COD)
	Peripheral arterial disease	PAD code (PAD_COD)
	Stroke or transient ischemic attack	Stroke code (STRK_COD) or TIA code (TIA_COD)

Table S3: Population estimates from EHR baseline dataset, SPD dataset and Census 2021

Age group	Sex	Census 2021	EHR	SPD4
0-9	Female	129,128	134,754	133,570
10-19	Female	121,284	124,849	128,360
20-29	Female	169,342	170,771	179,225
30-39	Female	193,912	212,501	199,775
40-49	Female	138,923	142,840	135,465
50-59	Female	110,634	111,167	107,750
60-69	Female	72,406	74,436	73,585
70-79	Female	46,145	46,783	48,885
≥80	Female	30,524	30,854	32,860
All	Female	1,012,298	1,048,955	1,039,475
0-9	Male	134,389	140,633	140,150
10-19	Male	125,651	130,398	133,250
20-29	Male	159,878	159,475	174,835
30-39	Male	183,284	230,112	207,365
40-49	Male	139,517	167,499	151,395
50-59	Male	108,302	121,158	112,375
60-69	Male	68,881	73,497	72,665
70-79	Male	37,588	38,520	40,925
≥80	Male	20,299	20,718	22,295
All	Male	977,789	1,082,010	1,055,255

Table S4: Population estimates from EHR baseline dataset and the datasets excluding persons without clinical activity within 12, 36 and 60 months

Age group	Census 2021 Number	EHR		EHR patients active within 12 months	
		Number	Difference to Census 2021	Number	Difference to Census 2021
0-9	263,517	275,387	11,870	231,461	-32,056
10-19	246,935	255,247	8,312	182,457	-64,478
20-29	329,220	330,246	1,026	258,296	-70,924
30-39	377,196	442,613	65,417	334,405	-42,791
40-49	278,440	310,339	31,899	252,613	-25,827
50-59	218,936	232,325	13,389	218,439	-497
60-69	141,287	147,933	6,646	145,749	4,462
70-79	83,733	85,303	1,570	84,634	901
≥ 80	50,823	51,572	749	50,549	-274
All	1,990,087	2,130,965	140,878	1,758,603	-231,484

Age group	Census 2021 Number	EHR patients active within 36 months		EHR patients active within 60 months	
		Number	Difference to Census 2021	Number	Difference to Census 2021
0-9	263,517	270,726	7,209	274,421	10,904
10-19	246,935	236,374	-10,561	247,332	397
20-29	329,220	308,973	-20,247	321,424	-7,796
30-39	377,196	397,233	20,037	417,675	40,479
40-49	278,440	291,677	13,237	300,368	21,928
50-59	218,936	227,780	8,844	229,877	10,941
60-69	141,287	147,632	6,345	147,720	6,433
70-79	83,733	85,171	1,438	85,206	1,473
≥ 80	50,823	50,929	106	50,971	148
All	1,990,087	2,016,495	26,408	2,074,994	84,907

Table S5: Household estimates from EHR and ABOAD administrative datasets and Census 2021 The values in EHR and Census 2021 estimates were rounded to a hundred in line with ABOAD published values

Household size	EHR	ABOAD	Census 2021
1	165,500	184,500	194,600
2	142,700	164,500	201,100
3	108,800	118,900	124,200
4	96,800	90,000	105,600
≥5	144,400	113,000	98,100
All	658,200	670,900	723,600

Local authority	EHR	ABOAD	Census 2021
Barking and Dagenham	69,300	69,700	73,900
Hackney	94,500	97,000	106,100
Havering	94,800	95,600	101,300
Newham	105,000	107,800	115,500
Redbridge	93,100	95,600	103,700
Tower Hamlets	108,800	112,300	120,500
Waltham Forest	92,700	92,700	102,900
All	658,200	670,700	723,900

Figure S1a: Ratio of female population by age group from electronic health records (EHR) and Census 2021 against their average. LoA – 95% limits of agreement

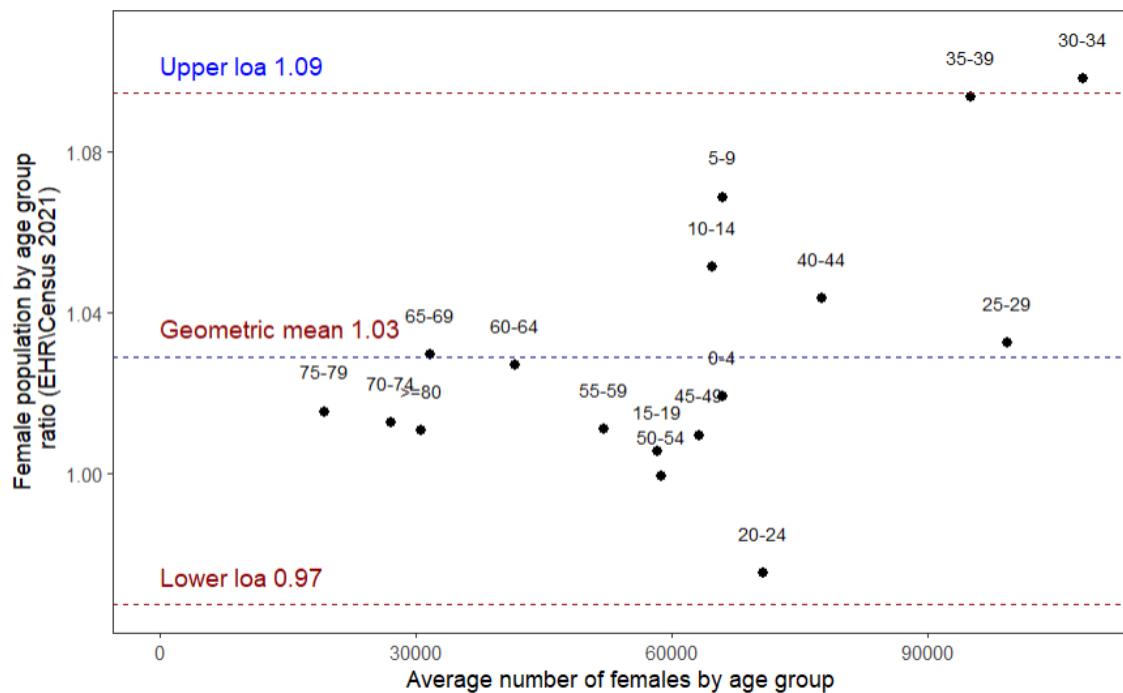


Figure S1b: Ratio of male population by age group from electronic health records (EHR) and Census 2021 against their average. LoA – 95% limits of agreement

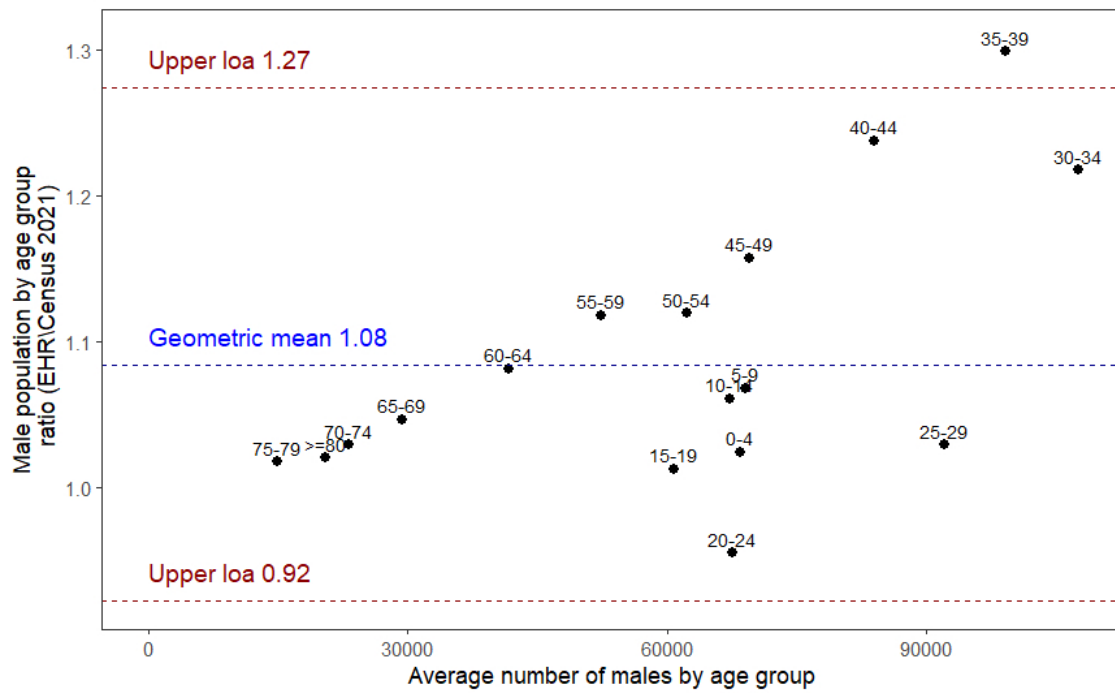


Figure S2: Ratio of population by Middle Statistical Output Area (MSOA) from electronic health records (EHR) and Census 2021 against their average. LoA – 95% limits of agreement

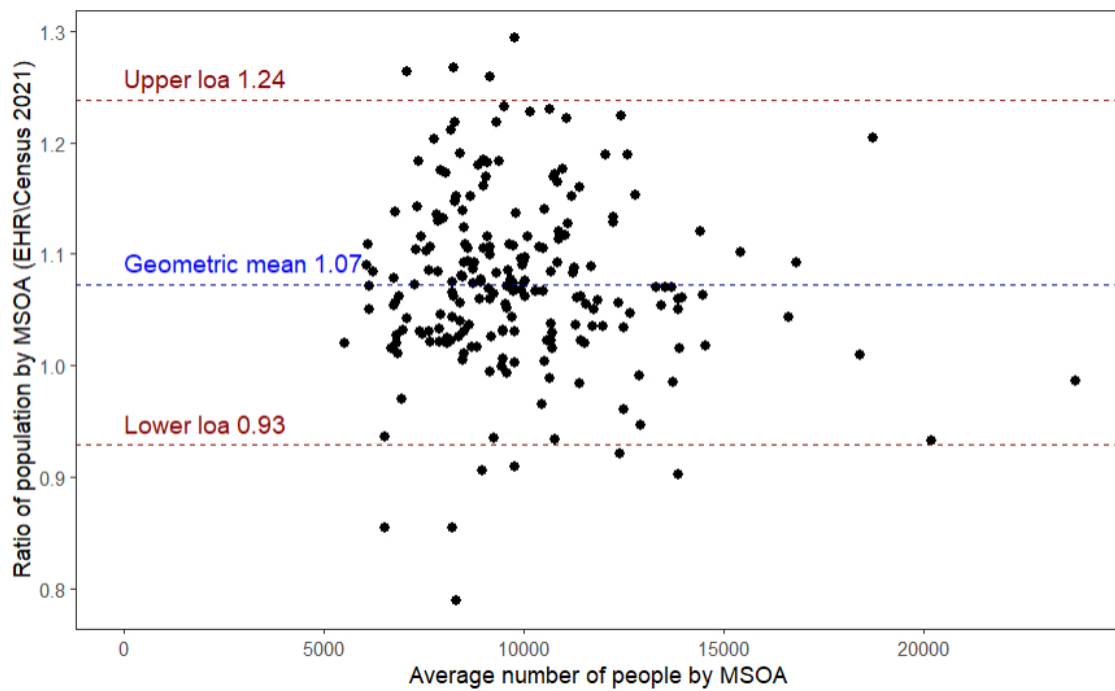
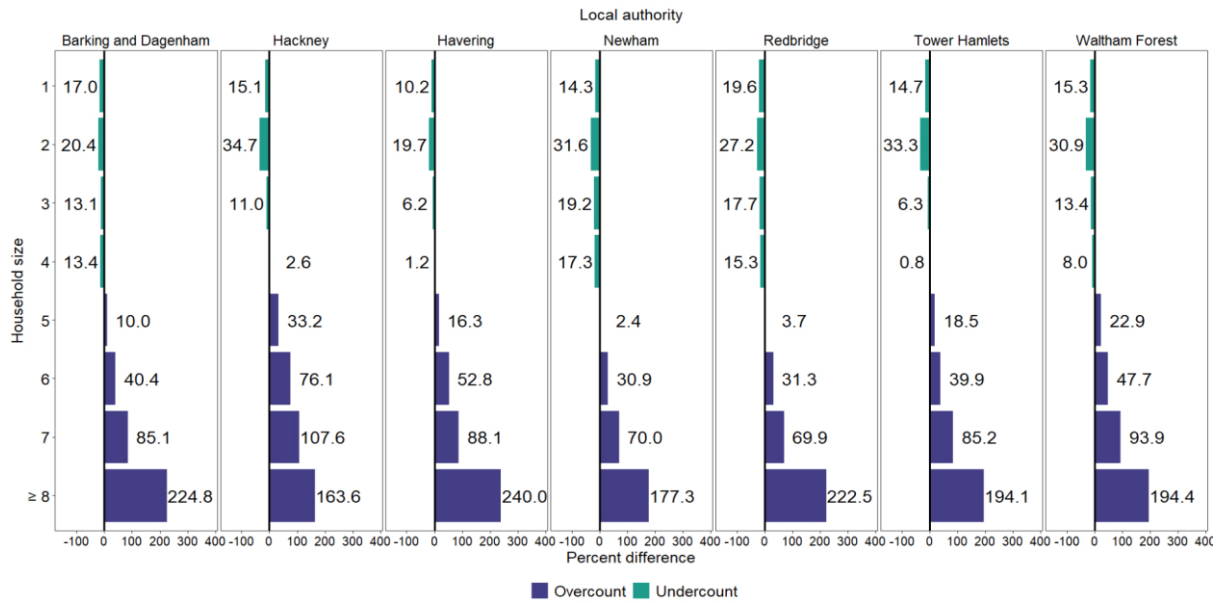


Figure S3: Percent difference between the number of households estimated from north-east London primary care electronic health records (EHR) and Census 2021 estimates by household size and local authority



Percent difference calculated as the difference between EHR count and ONS estimate divided by ONS estimate.

Figure S4: Ratio of households number by Middle Statistical Output Area (MSOA) from electronic health records (EHR) and Census 2021 against their average. LoA – 95% limits of agreement

