International Journal of Population Data Science

Journal Website: www.ijpds.org

Determining households from patient addresses and unique property reference numbers in general practitioner electronic health records

Gill Harper[1,*], Nicola Firman[1], Marta Wilk[1], Milena Marszalek[1], Paul Simon[2], David Stables[2], Richard Fry[3], Kelvin Smith[1], and Carol Dezateux[1]

Submission History

Submitted:	14/12/2023
Accepted:	27/02/2024
Published:	09/04/2024

[1]Clinical Effectiveness Group, Wolfson Institute of Population Health, Queen Mary University of London, London, UK
[2]Endeavour Health Charity, UK
[3]Population Data Science, Swansea University, UK

Abstract

Introduction
Households are increasingly studied in population health research as an important context for understanding health and social behaviours and outcomes. Identifying household units of analysis in routinely collected data rather than traditional surveys requires innovative and standardised tools, which do not currently exist.

Objectives
To design a utility that identifies households at a point in time from pseudonymised Unique Property Reference Numbers (UPRNs) known as Residential Anonymised Linkage Fields (RALFs) assigned to general practitioner (GP) patient addresses in electronic health records (EHRs) in north east London (NEL).

Methods
Rule-based logic was developed to identify households based on GP registration, address date, and RALF validity. The logic was tested on a use case on the household clustering of childhood weight status, and bias in success of identifying households was examined in the use case cohort and in a full population cohort.

Results
92.1% of the use case cohort was assigned a household. The most frequent dominant reason (55.3%) for a household not assigned was that a person had no valid household RALFs available across their patient registration address records. Other reasons are having none or multiple valid household RALFs, or not being alive at the event date.

In the use case, children not assigned to a household were more likely to attend schools in City & Hackney and living in the third most deprived quintile of lower super output areas.

88.9% of the population cohort was assigned a household. Patients not assigned to a household were more likely to be aged 18 to 45 years, living in City & Hackney, and living in the second quintile of most deprived lower super output areas.

Conclusions
We have developed a method for deriving households from primary care EHRs that can be implemented quickly and in real-time, providing timely data to support population health research on households.

Keywords
households; electronic health records; unique property reference numbers; patient data

*Corresponding Author:
Email Address: g.harper@qmul.ac.uk (Gill Harper)

https://doi.org/10.23889/ijpds.v9i1.2379

Introduction

Households are increasingly being used as a unit of analysis in research aimed at understanding the social context and wider determinants of health. Traditional definitions of households and sources of household level data have been from censuses and surveys. Demand to create households from routinely collected data and Big Data reflects the growth in exploiting linked administrative data reflecting their rich information content, speed, frequency, efficiency and lower cost of use, relative to surveys. In the absence of established gold standard methods to harness this data for various purposes, such as creating household units of analysis, new methods are continually required.

Recorded customer or patient addresses within routinely collected data can be used as a proxy for a household where a household is defined as persons who share the same address or residence at the same point in time. Representing addresses in routinely collected data with Unique Property Reference Numbers (UPRNs) [1, 2] – the unique identifier for every addressable location in Great Britain - provides a standardised property address label to support efficient identification of shared addresses across multiple persons and data.

In the UK, UPRNs are now a mandated standard within the public sector, and in 2019, the Public Sector Geospatial Agreement [3] gave more than 5,000 public sector organisations unlimited access to Ordnance Survey data, including UPRNs. We have previously reported the ASSIGN algorithm [4] which we developed to assign UPRNs to general practitioner (GP) patient addresses in National Health Service (NHS) Electronic Health Records (EHRs) in near real-time. This has been implemented in the Discovery Data Service (DDS) covering patients registered with GPs in north-east, south-east and north-west London.

The Secure Anonymised Information Linkage (SAIL) databank (which has worked with UPRNs in their data since 2012) [5], Harper and Mayhew [6], and the Office of National Statistics (ONS) Administrative Data Census (ADC) team [7, 8], were early adopters of UPRN household methods, assigning populations created from linked administrative government and health data to UPRNs to represent occupants of households. The ONS define these as 'occupied addresses'. The latter two methods require each occupant to have a UPRN assigned to their recorded address, and exclude occupants of communal establishments.

Subsequent research has utilised UPRNs on recorded addresses in health data to create households using varying methods. Lloyd *et al* [9] identified household occupants from patients currently registered with a GP using pseudonymised UPRNs in the English Master Patient Index. Similarly, the SAIL databank created households from encrypted UPRNs and address registration dates for individuals registered with a GP practice in Wales [10]. Stafford et al [11] linked a local sample of EHRs to local authority household composition records by UPRN for one London borough to represent households.

The 2019 Coronavirus (COVID) pandemic saw increased momentum in a UPRN approach to creating households for population health purposes when a rapid response was required to understanding COVID and households. Household members became a focus for transmission and outcome risk [12–15].

In existing research, there has been a lack of detail and justification for how methods have been devised for creating households from administrative data. Only the SAIL databank [9] and Lloyd *et al* [8] have described the additional rules and criteria to select household relevant UPRNs, but approaches have not been consistent.

While the GP patient register alone may not capture all correct and current household residents and may bias who is omitted or incorrectly included [16, 17], GP patient registration data provides the greatest coverage of large regional and national populations given that the UK NHS is free at the point of use and is routinely updated.

We report a transparent and reproducible approach to identifying household occupants solely from information available from routine primary care EHRs available for all registered patients, developed by Queen Mary University of London (QMUL) and Endeavour Health Charity and supported by ADR UK (Administrative Data Research UK) [18]. Our overarching aim is to exploit the availability of current real-time and historical UPRNs in routine primary care EHRs for a variety of research purposes centred around identifying members of a household at a specific point in time. We illustrate this method through an indicative use case examining clustering of household child weight status. The use case requires a method to reliably identify the UPRN for the household residence of each member of the study population at a specified point in time, and to include all household occupants at that point in time.

Methods

Data source

The north-east London (NEL) DDS includes EHRs for patients registered with all general practices providing primary care services to the entire geography covered by seven NEL boroughs. At the time of data extraction for this analysis this included 277 general practices. Each GP publishes individual level data (identifiable to approved users, otherwise de-identified), directly from their electronic patient record enterprise system on a daily basis into the DDS and this is provided in deidentified format as a subscriber database. Data was provided in a de-identified format for this study.

Patient registration data

The GP patient EHR contains demographic and registration information including the dates when patients were initially registered (enrolled) with a general practice (start date) and when they deregistered (end date), their age and sex.

Person/patient relationship

A person, recorded as a pseudonymised NHS number, may have multiple patient registrations across time in the NEL DDS system. Each patient registration has a unique ID.

Patient address data

Patients provide their place of residence address to the general practice when they register. In England, practices usually have

a catchment area for eligibility to register and a patient's address confirms their eligibility. Patients are required to advise their general practice of any change of address. Presently general practices do not validate the patient address quality or accuracy when they are provided to them.

The DDS creates address records for patients from the information provided by the GP EHR clinical systems, namely Egton Medical Information Systems (EMIS) and SystmOne (The Phoenix Partnership [TPP]). These clinical systems only hold one current address per patient at any one time. However, with each daily update, if there has been a change of address, NEL DDS records the current date as the end date of the previous address, and the start date of the new address and it retains the previous address. Any change in the address string will trigger a new address record. The address end date is null if it is the current address record. Both the start and end dates are null if it is the only and current address record associated with the registration. One of three address types are assigned based on the NHS GP clinical system Fast Healthcare Interoperability Resources (FIHR) national standard value [19]: 'home address', 'temporary address' or 'old address'.

There were some instances of data corruption: multiple address records containing null or overlapping start and end dates and address time periods not nesting exactly into registration time periods. No pre-cleaning of the raw data was undertaken, therefore the algorithm deals with the data in this state.

Every address record in DDS is allocated a UPRN from the Ordnance Survey Great Britain property gazetteer database AddressBase Premium [20] in near real-time using the ASSIGN algorithm [4]. This is a quality-assured and validated address-matching algorithm with a 98.6% match rate (based on a population of 1.8 million adults registered with a GP in north east London) and high sensitivity and positive predictive value. The UPRN is pseudonymised into a Residential Anonymous Linking Field (RALF) [21] using study-specific encryption keys to preserve patient anonymity and confidentiality. Pseudonymisation is necessary because UPRNs (and in some cases their associated addresses and geographic locations) are publicly available open data. DDS also retains for each UPRN match a set of metadata about the match (created in ASSIGN) or about the dwelling (taken from AddressBase Premium).

Household definition

We define households as comprising one or more people registered as living at the same residence at the same point in time, regardless of relationship, and subject to individual and RALF eligibility rules.

Event date

The event date - the point in time used to define a person's place of residence - can be fixed, i.e. the same date for each person (such as 21st March 2021, the England and Wales Census date), or variable, i.e. different for each person (such as the date of a specific clinical diagnosis, vaccination, or measurement). These dates could be sourced from within the primary care record or provided from external third-party data sources.

Use case

We tested the method in a study, reported elsewhere [22], to examine household clustering of childhood obesity. In this example, dates of school measurements of height and weight varied for each child and were provided by a third-party – local authority public health departments - under a data processing agreement.

We linked school measurement records for 126,829 children participating at 4–5 or 10–11 years of age in the school-based annual National Child Measurement Programme (NCMP) [23] in state-maintained primary schools from four NEL local authorities: Tower Hamlets (2015–2019 school years), City & Hackney (2013–2019), Newham (2014–2019), Waltham Forest (2013/14 and 2015–2019) to GP patient registrations in the DDS. We identified all households with NCMP participants. The household match rate and reasons for non-matches were calculated.

Bias

We compared proportions of demographic variables of the use case cohort with and without a household assigned to examine bias. We did this also for a larger cohort of a full population of 1,374,495 patients of all ages registered with a GP Practice in Tower Hamlets, City & Hackney, Newham and Waltham Forest Clinical Commissioning Groups (CCGs) that had been run through the method to assign a household at England and Wales Census day 21$^{\mathrm{st}}$ March 2021.

Logic

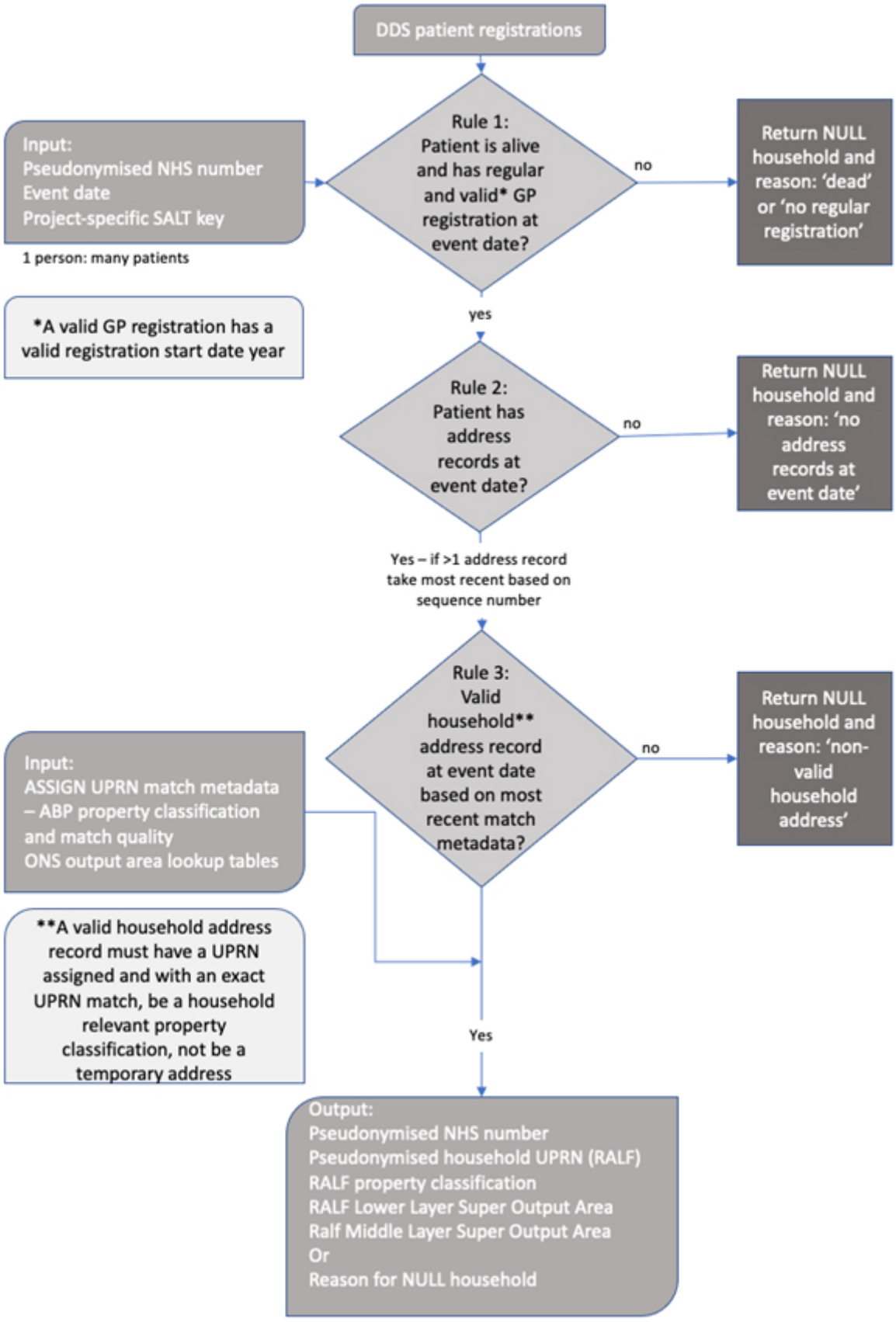
The logic went through a number of iterations. Coding was harmonised across R, Stata and Microsoft SQL Server (MS SQL), with the results from each version compared to identify any disparities. This helped inform the final version of the logic, which was simplified and informed by intelligence from the team, incorporating specific features of the DDS data structure and data quality. The final version was coded in MS SQL and Python (see Supplementary Appendix 1 for Python code) and is summarised in Figure 1.

The logic requires a file containing the pseudonymised NHS number for every person in the cohort, the event date of interest, and the project SALT key, a tool that hashes and encrypts the identifiable NHS number and UPRN so that they are pseudonymised and non-identifiable. The project SALT key is input here so that it is used to create the RALF in the output.

Rule 1 scans and extracts every patient registration that the DDS holds for each person in the cohort. It requires the patient to be alive on the event date and to have a regular i.e. non-temporary GP registration, and for that registration to have valid registration dates. Invalid registration dates are implausible: dates from before the NHS existed, dates in the future, or administrative dummy dates as a proxy for unknown dates. We excluded temporary registrations which imply a person is not a long-term occupant of the household.

The event date was allowed to be after but not including the registration start date, and earlier than but not including the date of death to allow for date range exclusivity in how the data is recorded by the DDS.

Figure 1: Flowchart of household RALF at event date logic ABP = AddressBase Premium



Rule 2 determines if an address record exists for a patient registration at the event date if the address start date is earlier or equal to the event date or is null, and if the address end date is later or equal to the event date or is null. This factors in that in the DDS the address record start and end date can be null.

If there are multiple address records associated with a patient registration, these are assessed in order of recency, determined by the record sequence ID. When the most recent address record is found to exist at the event date, no further address records are assessed. This single address record is passed on to Rule 3.

Rule 3 ensures that the RALF relates to a valid residential household. It uses the UPRN match metadata to check that a UPRN has been assigned, the UPRN is an exact match to the patient address (and not an approximate match), and that the UPRN has a household relevant property classification. 'Temporary' address types are excluded.

If an address record has multiple UPRN match metadata associated with it due to being run through the ASSIGN address-matching algorithm multiple times, the most recent match metadata is chosen.

The logic outputs, for each person, either a null household RALF and the reason why, or the household RALF found at the event date. The property classification from AddressBase Premium, and the Lower layer Super Output Area (LSOA) and Middle layer Super Output Area (MSOA) of the RALF from ONS lookup tables [24] are also provided to approved users within statistical disclosure control standards.

The RALF is encrypted a second time if the output was approved by the DDS data controllers for third party uses (research, planning or health intelligence).

If a person has more than one patient registration that returns a household RALF at the event date, these will exist in the output as multiple rows per pseudonymised NHS number.

Results

Performance

Performance was improved by applying single indexes to the AddressBase Premium UPRNs, and hosting the database and the client in the same CPU memory space. Approximately 849,000 records were processed per minute if there was no requirement to output the reason for a NULL household RALF. If the reason is required, then approximately 157,000 records were processed per minute.

Use case

Each revised version of the logic was run on a test dataset relating to the use case. This comprised 126,829 children with an NCMP measurement and a NEL GP registration ever. Manual checks were made at each iteration, and the final version of the logic identified a household RALF for 116,801 (92.1%) children.

There are up to four non-mutually exclusive reasons why 10,025 of the cohort were not assigned a household RALF across all their address records and patient registrations. In Table 1, these four reasons have been numbered and ranked from 1 to 4, with the higher numbers 'trumping' lower numbers. If a person had multiple address records that were not assigned a household RALF for a combination of all four possible reasons, reason 1 is the highest rank and would be assigned overall.

The most frequent dominant reason for a household RALF not to be assigned in the use case cohort is that none of the RALFs referred to valid households (55.3%). An address is a valid household if a UPRN is assigned with an exact UPRN match, and it has a household relevant property classification, and not be a temporary address.

The proportion of the 10,025 children in the cohort without a household RALF with each combination of these four reasons across their address records is given in Supplementary Appendix 2. The most frequent combination at 35% is to have no valid household RALFs and either is not alive or has no regular GP registrations at the event date.

We examined demographic biases in household RALF assignment (Table 2) and noted where there was a greater than 3% difference in proportions with and without a household RALF. Children of South Asian ethnic group, who participated in the NCMP in 2018, attending schools in Tower Hamlets and living in LSOAs in the second quintile of the IMD were more likely to have household RALF assignment and children attending schools in City & Hackney and living in LSOAs in the third quintile of the IMD were more likely to not have household RALF assignment.

Population cohort

A similar examination of demographic biases in household RALF assignment for a fuller cohort of the NEL GP registered EHR population as at England and Wales Census date 21st March 2021 is given in Table 3. 88.9% of the cohort were assigned a household RALF. Demographic variables used for bias are slightly different between Tables 1 and 2 due to their different sources.

A greater than 3% difference in proportions with and without a household RALF was found for people in the cohort aged under 18 years old, of South Asian ethnic group, and living in the first quintile of most deprived LSOAs who were more likely to have household RALF assignment. Patients aged 18 to 45 years, living in City & Hackney, and living in LSOAs in the second quintile of the IMD were more likely to not have household RALF assignment.

Discussion

Key findings

A method to identify occupants of a household at either a fixed or variable point in time using information from routine primary care EHRs has been developed and implemented in the DDS subscriber database held by the Clinical Effectiveness Group for research and development purposes. The logic is transparent and reproducible in other coding environments.

Using this method, we assigned households to 92.1% of members of a cohort of children participating in the NCMP. The most frequent dominant reason for a household RALF not to be assigned in the use case cohort is that none of the RALFs referred to valid households (55.3%).

Bias was found in household RALF assignment success. In the use case cohort, children of South Asian ethnic group, who participated in the NCMP in 2018, attending schools in Tower Hamlets and living in the second quintile of most deprived

Table 1: Summary of dominant reason a household RALF was not assigned for persons in the use case cohort

Main reason for NULL household RALF	Reason rank	Frequency	%
Multiple different valid household RALFs	1	423	4.2
No valid household RALFs	2	5,548	55.4
No address records at event date	3	963	9.6
Not alive or no regular registrations at event dates	4	3,091	30.8
Total		10,025	100

RALF = Residential Anonymous Linking Field.

Table 2: Proportional differences in demographic variables for persons in use case cohort with a household RALF and those without a household RALF

		With household RALF n = 116,804		Without household RALF n = 10,025		
		n	%	n	%	Difference (%)
Sex	Female	57,389	49.1	4,882	48.7	0.4
	Male	59,415	50.9	5,143	51.3	−0.4
Ethnic group from NCMP	Black	20,580	17.6	1,952	19.5	−1.9
	Mixed and other	21,539	18.4	2,034	20.3	−1.8
	South Asian	35,590	30.5	2,383	23.8	**6.7**
	White	27,248	23.3	2,612	26.1	−2.7
	Not Stated or Null	11,847	10.1	1,044	10.4	−0.3
School year of NCMP measurement	Reception	60,694	52.0	5,412	54.0	−2.0
	Year 6	56,110	48.0	4,613	46.0	2.0
Year of NCMP measurement	2013	1,117	1.0	161	1.6	-0.6
	2014	10,196	8.7	1,101	11.0	−2.3
	2015	16,323	14.0	1,679	16.7	−2.8
	2016	25,752	22.0	2,487	24.8	−2.8
	2017	27,139	23.2	2,088	20.8	2.4
	2018	24,699	21.1	1,674	16.7	**4.4**
	2019	11,578	9.9	835	8.3	1.6
Local authority of school	City & Hackney	25,991	22.3	2,955	29.5	**−7.2**
	Newham	39,589	33.9	3,650	36.4	−2.5
	Tower Hamlets	22,554	19.3	726	7.2	**12.1**
	Waltham Forest	28,670	24.5	2,694	26.9	−2.3
IMD 2019 quintile of child's home LSOA (1 = most deprived, 5 = least deprived)	1	64,648	0.1	5,201	0.2	−0.1
	2	43,785	55.3	4,173	51.9	**3.5**
	3	6,833	37.5	499	41.6	**−4.1**
	4	1,066	5.8	94	5.0	0.8
	5	309	0.9	38	0.9	0.0
	Null	163	0.3	20	0.4	−0.1

Differences greater than 3% in bold. NCMP = National Child Measurement Programme, IMD = Index of Multiple Deprivation, LSOA = Lower layer Super Output Area.

LSOAs were more likely to have household RALF assignment. Children attending schools in City & Hackney and living in the third quintile of most deprived LSOAs were least likely to have household RALF assignment.

We assigned households to 88.9% of a larger population cohort. Patients aged under 18 years old, patients of South Asian ethnic group, and patients living in the first quintile of most deprived LSOAs were more likely to have household RALF assignment, and people aged 18 to 45 years, living in City & Hackney, and living in the second quintile of most deprived LSOAs were least likely to have household RALF assignment.

Table 3: Proportional differences in demographic variables for persons in Census day 2021 population cohort with a household RALF and those without a household RALF

		With household RALF n = 1,222,339		Without household RALF n = 152,156		
		n	%	n	%	Difference (%)
Sex	Female	593,994	48.6	70,682	46.5	2.1
	Male	628,345	51.4	81,474	53.5	−2.1
Age	Up to 18 years old	265,223	21.7	26,198	17.2	**4.5**
	18 to 45 years old	620,394	50.8	87,801	57.7	**−6.9**
	45 to 65 years old	248,295	20.3	28,218	18.5	1.8
	65 years and older	88,427	7.2	9,939	6.6	0.6
Ethnic group from EHR	Black	169,701	13.9	18,325	12	1.9
	Mixed and Other	126,162	10	19,413	12.7	−2.7
	South Asian	337,907	27.6	36,744	24.1	**3.5**
	White	506,286	41.4	65,970	43.4	−2
	Not Stated or Null	82,283	7.1	11,704	7.8	−0.7
Local authority of patient address	City & Hackney	263,443	21.7	39,161	26	**−4.3**
	Newham	369,917	30.3	44,407	28.9	1.4
	Tower Hamlets	309,397	25.3	36,957	24.3	1
	Waltham Forest	278,946	22.6	30,781	20.2	2.4
	Other	636	0.1	850	0.6	−0.5
IMD 2019 quintile of home LSOA (1 = most deprived, 5 = least deprived)	1	355,659	29.1	38,285	25.2	**3.9**
	2	656,144	53.7	90,328	59.4	**−5.7**
	3	157,634	12.9	17,694	11.6	1.3
	4	40,201	3.3	4,328	2.8	0.5
	5	12,065	0.9	671	0.4	0.5
	Null	636	0.1	850	0.6	−0.5

Differences greater than 3% in bold. EHR = Electronic Health Record, LSOA = Lower layer Super Output Area.

Strengths and limitations

The methodology was able to draw upon routinely collected primary care EHRs for a whole population with near real-time UPRN assignment. While the logic is specific to the architecture of the NEL DDS, it is generalisable and can be adapted to other health record systems allocating UPRNs to patient addresses. We have presented a transparent account of the rules used and reasons for exclusion of addresses or individuals. The outputs of our code enable researchers to understand reasons for a non-match and any associated biases.

The utility can be implemented quickly and in real-time, providing frequent granular data on households, overcoming reliance on the decennial census with aggregated outputs.

We were not able to link the GP patient registration data to any other population dataset to improve the completeness of ascertainment of the population; for example by identifying household members not registered with a GP or by removing people who had moved but not updated their addresses or changed GP. Accuracy of the GP registration address records rely on the quality of the address given by the patient and changes in address being recorded and updated by the practice in a timely manner.

We applied stringent criteria to select only those UPRN matches and property types that were indicative of a household, however we were not able to benchmark and validate the results against any gold standard household occupant dataset.

If a RALF was excluded at Rule 3 as a non-valid household because there was no property classification for the UPRN, there may be geographical bias in which local authorities have higher proportions of property classification missing in their local property gazetteers that feed into AddressBase Premium. Therefore, the bias would be sourced from the geography, not the person. This will be further explored in future work.

Caveats to be considered by users are that by using electronic health record data, we do not know the relationships between the household occupants. This may or may not be a disadvantage, depending on the application. Also, to understand how patient addresses that the method is run on are sourced and maintained in the database. In this case, the results are subject to some DDS address data quality issues. Data flow into the DDS began in 2014 therefore the system holds only address records at that point in time and address changes since then. Address records will exist for registrations that ended pre-2014, due to the patient leaving or dying, but this will only be the current address at the time of leaving or

dying. Therefore, determining the household RALF in DDS for event dates before 2014 is less reliable.

Where the results contain multiple different valid household RALFs at an event date for a person in a cohort, it is up to the user to decide the most appropriate course of action depending on their purposes.

Implications

The household RALF utility has the flexibility to be used for any fixed or variable event date and creates households in a standardised way that was not previously available. It is currently challenging for researchers to identify individual households in a robust way and link other housing and property information to them, resulting in a lack of research-ready data on outcomes within and between different types of households and how they change over time. The scale and coverage of EHR data offers the potential to create households for larger populations and in a more timely manner longitudinally than is found in the more traditional longitudinal household surveys that researchers have previously had available to them such as the Understanding Society UK Household Longitudinal Survey [25] or the NHS Health Survey for England [26].

This utility will contribute to meeting that challenge and enable important population health research by providing the means to create a household unit of analysis to study the household context. As well as creating households from EHR data, variables within the EHR record can be used to characterise the household composition and typology for further household context. The household context is important because the composition of a household plays a role in the social, economic and health experience of the occupants.

The biases in household assignment are small but important to identify so that their impact can be considered for different research populations and purposes. The predominant reasons for lack of household assignment in this study are no valid household RALFs and either not alive or no regular GP registrations. This is likely to relate to underlying data quality in the GP patient record which is influenced by demographic, geographic and organisational factors, as described in Harper et. al. 2021 [4]. It is not clear why in both the child use case and the larger population cohort, assignment rates were higher for patients living in or attending schools in Tower Hamlets and lower for patients living in or attending schools in City & Hackney when the opposite may be expected due to Tower Hamlets' higher proportion of properties that are flats which can translate into lower UPRN match rates. There was no clear pattern between level of local deprivation. These results need further exploration. The issue of not having a UPRN assigned from prior address-matching can be lessened by recording patient addresses and UPRNs at registration in the BS7666 address standard format [27] in AddressBase, as is happening in NHS Scotland with the CHI2 patient system [28].

Researchers should be aware that the quality of data in the GP patient record generally has implications for the accuracy of the identified household. As well as the quality of the address determining if a correct UPRN can be assigned, if a patient does not actually live at the recorded address at a point in time, and is not recorded at their correct address, this will affect the accuracy of any household occupancy and composition measures. It is beyond the scope of this paper but it is well documented that GP list inflation and gaps are recognised issues [29] and is known to be non-random with young men, young adults and healthy people less likely to keep their registration details up-to-date or to not be registered at all. Harper and Mayhew 2012 [30] created a population and household method from linked GP patient and local authority data to deal with this.

There is a relatively small but growing body of work on household-level studies. Concordant poor physical and mental health between household members has been found [31–34]. Household composition and the health status of household members were found to be relevant to children's health. For example, children in smaller households have better health, educational and economic outcomes compared with children from larger families [35–37]; single children (no other child in the household) and children sharing a household with older children with obesity were found to be more likely to be living with obesity [38–40]. Household structure and living arrangements were found to influence self-rated health, mobility limitations and depressive symptoms in adults [41]. The household utility can support more household studies based on real-world EHR data in a faster standardised way.

Household level data provides granular evidence, rather than aggregated ecological inference, of the wider upstream determinants of health to drive effective household level interventions and policies. Knowing the actual demographic, health and property context for a household unit rather than the average or counts for a combined area provides greater statistical strength and stronger evidence.

The utility can be implemented quickly and in real-time supporting snapshot and longitudinal approaches to understanding household circumstances and outcomes.

Next steps

Future work is planned as part of the wider ADR UK [16] funded programme of work on Healthy Households. First is to extend the utility by calculating a defined set of household level variables for each household RALF identified, including total occupancy, breakdown of occupancy count by specified age and sex groups, and household composition type (e.g. three-generational or single-adult households). This will be based on the demographic characteristics of all NEL DDS GP patients identified as living at the same household RALF at an event date.

Then we will work with the SAIL Databank in Wales and the Scottish National Safe Haven in Scotland to scale and standardise the household RALF method and use on English, Welsh and Scottish data. Full population household spines will be created for each of the three countries using EHRs.

The programme will also explore a robust validation method of household counts, investigating the possibility of benchmarking and comparing against other household count and occupant datasets. One option is to compare against household counts from the 2021 Census, ideally at line-level rather than aggregated level. Our team will apply for permission to access line-level Census data under the Healthy Households ADR UK funded project. ONS develop population and household counts from linked anonymous administrative data [42], which offer another comparison dataset, although

any comparison exercise would need to acknowledge the differing methods and definitions used to create each source. We envisage our method not as a Census household count replacement, but as a way to create and identify household units of analysis at fixed or variable dates for research.

Finally, linkages are planned to housing data held by government and local authorities to develop a dynamic method of assessing over-crowding at the household level and to develop a robust validation method.

Conclusion

The household RALF utility has been developed for use with NHS primary care EHRs to identify household units of analysis in a standardised way. Transparency in methods using electronic health records and other administrative data for research is important for reproducibility and robustness of analyses. The utility is innovative and fit-for-purpose and it will support important population health research based on the household context.

Acknowledgments

This work was supported by funding from Endeavour Health Charity and ADR UK (Administrative Data Research UK) an Economic and Social Research Council investment (part of UK Research and Innovation) Grant number: ES/X00046X/1 and funded by a grant from Barts Charity (ref: MGU0419). This work also uses data provided by patients and collected by the NHS as part of their care and support, specifically data provided by patients in east London and recorded by the NHS general practitioners who shared de- identified data for research purposes via the Discovery Data Service which was curated with the support of the Queen Mary University Clinical Effectiveness Group and the north east London Discovery Programme.

Statement on conflicts of interest

None to be declared.

Ethics statement

Ethics approval was not required or obtained. Approval for access to the person identifiable data (patient addresses) used in this study was provided by the north east London Discovery Data Service data controllers to the Clinical Effectiveness Group as appointed data sub-processors for the sole purpose of developing and evaluating the household algorithm for direct patient care. This access was limited to approved individuals with appropriate information governance training working in a secure trusted data environment. Of the authors, Gill Harper, Carol Dezateux, and Paul Simon had access to identifiable data.

Only aggregated patient data are reported in this study.

Data availability statement

The data controllers for the data used in this study is the north east London Discovery Data Service (DDS). This access was limited to approved individuals with appropriate information governance training working in a secure trusted data environment. This data is not publicly available and the DDS do not allow the authors to onwardly share this data. Any applications for this data can be made to DDS directly who will advise on the correct procedures.

References

1. Geoplace LLP. Available at https://www.geoplace.co.uk/.

2. Ordnance Survey 'Adopting the UPRN'. Available at https://www.ordnancesurvey.co.uk/newsroom/blog/adopting-the-uprn.

3. Ordnance Survey. Public Sector Geospatial Agreement (PSGA). Available from https://www.ordnancesurvey.co.uk/business-government/public-sector-geospatial-agreement.

4. Harper G, Stables D, Simon P, Ahmed Z, Smith K, Robson J, Dezateux C. 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. 2021;6(1). https://doi.org/10.23889/ijpds.v6i1.1674

5. NIHR. Change in alcohol outlet density and alcohol-related harm to population health. Research Award ID 09/3007/02. Available at https://fundingawards.nihr.ac.uk/award/09/3007/02

6. Harper G, Mayhew L. Using administrative data to count and classify households with local applications. Applied Spatial Analysis and Policy. 2016 Dec;9:433-62. https://link.springer.com/article/10.1007/s12061-015-9162-2

7. Office of National Statistics. Annual Assessment of ONS's Progress on the administrative Census. July 2018. Available at https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments/annualassessmentofonssprogressontheadministrativedatacensusjuly2018.

8. Office of National Statistics. Research Outputs: An update on developing household statistics for an Administrative Data Census. Available at https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/householdsandfamilies/researchoutputsanupdateondevelopinghouseholdstatisticsforanadministrativedatacensus.

9. Lloyd T, Crellin E, Brine RJ, Shen JY, Wolters AT. Association between household context and emergency

hospital use in older people: a retrospective cohort study on indicators for people living alone or living with somebody with frailty, developed from routine healthcare data in England. BMJ open. 2022 May 1;12(5):e059371. https://doi.org/10.1136/bmjopen-2021-059371

10. Johnson RD, Griffiths LJ, Hollinghurst JP, Akbari A, Lee A, Thompson DA, Lyons RA, Fry R. Deriving household composition using population-scale electronic health record data—A reproducible methodology. PLoS One. 2021 Mar 29;16(3):e0248195. https://doi.org/10.1371/journal.pone.0248195

11. Stafford M, Deeny SR, Dreyer K, Shand J. Multiple long-term conditions within households and use of health and social care: a retrospective cohort study. BJGP open. 2021 Apr 1;5(2). https://bjgpopen.org/content/5/2/BJGPO.2020.0134

12. Thelwall S, Zaidi A, Nsonwu O, Rice W, Chudasama D, Lamagni T, Dabrera G. The role of multi-generational household clusters in COVID-19 in England. medRxiv. 2021 Nov 24:2021-11. https://www.medrxiv.org/content/10.1101/2021.11.22.21266540v2

13. Lynda F, Ciara G, David C, Sam C, Jen B, Martin R, Jane W, Campbell M, Hutchinson S, Robertson C, Helen MC. Risk of hospitalisation with covid-19 among teachers compared to healthcare workers and other working-age adults. A nationwide case-control study. medRxiv. 2021 Feb 8:2021-02. https://www.medrxiv.org/content/10.1101/2021.02.05.21251189v1

14. Shaw, R. How the UPRN saved lives: a story of vaccinating for COVID. GeoPlace Annual Conference 10 June 2023. Available at https://vimeo.com/826914331?share=copy

15. Thompson DA, Abbasizanjani H, Fry R, Marchant E, Griffiths L, Akbari A, Hollinghurst J, North L, Lyons J, Torabi F, Davies G. Staff–pupil SARS-CoV-2 infection pathways in schools in Wales: a population-level linked data approach. BMJ Paediatrics Open. 2021;5(1). https://doi.org/10.1136/bmjpo-2021-001049

16. Harper G, Mayhew L. Using administrative data to count local populations. Applied Spatial Analysis and Policy. 2012 Jun;5(2):97-122. https://link.springer.com/article/10.1007/s12061-011-9063-y

17. Office for National Statistics. Patient Register: quality assurance of administrative data used in population statistics, Dec 2016. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016

18. Healthy Households Project. 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. Press release available at https://www.adruk.org/news-publications/news-blogs/data-linkage-funding-will-drive-new-insights-into-the-health-of-households-821/

19. NHS Digital. FHIT (Fast Healthcare interoperability Standards) Available at https://digital.nhs.uk/services/fhir-apis

20. Ordnance Survey. AddressBase Premium. Available at https://www.ordnancesurvey.co.uk/business-government/products/addressbase-premium

21. Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, John G, Verplancke JP. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. Journal of Public Health. 2009 Dec 1;31(4):582-8. https://doi.org/10.1093/pubmed/fdp041

22. Firman N, Wilk M, Marszalek M, Griffiths I, Harper G, Dezateux C. Is obesity more likely among children sharing a household with an older child with obesity? Conference Proceedings for ADR UK Conference 2023 Vol. 8 No. 2. Available at https://ijpds.org/article/view/2203

23. NHS Digital. National Child measurement Programme. Available at https://digital.nhs.uk/services/national-child-measurement-programme/

24. National Statistics UPRN Lookup. Available at https://geoportal.statistics.gov.uk/datasets/e75da94b5c7642ea920bba5e6a84baa7/about

25. Understanding Society UK Household Longitudinal Survey. Available at https://www.understandingsociety.ac.uk/.

26. NHS Digital Health Survey for England. Available at https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england.

27. British Standard 7666. Available at https://www.geoplace.co.uk/press/2015/british-standard-7666-2006-its-impact-use-within-local-government

28. CHI2 – Community Health Index, as per personal correspondence with Professor Chris Dibben, Director of the Scottish Centre for Administrative Data Research.

29. Office for National Statistics. Patient Register: quality assurance of administrative data used in population statistics, Dec 2016. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulationstatisticsdec2016

30. Harper G, Mayhew L. Using administrative data to count local populations. Applied Spatial Analysis and Policy. 2012 Jun;5:97-122. https://link.springer.com/article/10.1007/s12061-011-9063-y

31. Qin Y, Guo Y, Tang Y, Wu C, Zhang X, He Q, He J. Concordance of chronic conditions among the household members in Shanghai: a cross-sectional study. BMJ open. 2019 Dec 1;9(12):e031240. https://doi.org/10.1136/bmjopen-2019-031240

32. Meyler D, Stimpson JP, Peek MK. Health concordance within couples: a systematic review. Social science & medicine. 2007 Jun 1;64(11):2297-310. https://doi.org/10.1016/j.socscimed.2007.02.007

33. Patel SA, Dhillon PK, Kondal D, Jeemon P, Kahol K, Manimunda SP, Purty AJ, Deshpande A, Negi PC, Ladhani S, Toteja GS. Chronic disease concordance within Indian households: A cross-sectional study. PLoS medicine. 2017 Sep 29;14(9):e1002395. https://doi.org/10.1371/journal.pmed.1002395

34. Nielsen J, Bahendeka SK, Whyte SR, Meyrowitsch DW, Bygbjerg IC, Witte DR. Household and familial resemblance in risk factors for type 2 diabetes and related cardiometabolic diseases in rural Uganda: a cross-sectional community sample. BMJ open. 2017 Sep 1;7(9):e015214. https://doi.org/10.1136/bmjopen-2016-015214

35. Tang-Péronard JL, Heitmann BL. Stigmatization of obese children and adolescents, the importance of gender. Obesity Reviews. 2008 Nov;9(6):522-34. https://doi.org/10.1111/j.1467-789X.2008.00509.x

36. Parsons TJ, Power C, Logan S, Summerbelt CD. Childhood predictors of adult obesity: a systematic review. International journal of obesity. 1999 Nov 1;23.

37. Lersch PM. Fewer siblings, more wealth? Sibship size and wealth attainment. European Journal of Population. 2019 Dec;35(5):959-86. https://doi.org/10.1007/s10680-018-09512-x

38. Whitaker RC, Wright JA, Pepe MS, Seidel KD, Dietz WH. Predicting obesity in young adulthood from childhood and parental obesity. New England journal of medicine. 1997 Sep 25;337(13):869-73. https://doi.org/10.1056/NEJM199709253371301

39. Reilly JJ, Kelly J. Long-term impact of overweight and obesity in childhood and adolescence on morbidity and premature mortality in adulthood: systematic review. International journal of obesity. 2011 Jul;35(7):891-8. https://doi.org/10.1038/ijo.2010.222

40. Chanfreau J, Barclay K, Keenan K, Goisis A. Sibling group size and BMI over the life course: Evidence from four British cohort studies. Advances in Life Course Research. 2022 Sep 1;53:100493. https://doi.org/10.1016/j.alcr.2022.100493

41. Hughes ME, Waite LJ. Health in household context: Living arrangements and health in late middle age. Journal of health and social behavior. 2002 Mar;43(1):1.

42. Office for National Statistics. Progress updates – transforming our population and migration statistics. October 2023. Available at https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/censusanddatacollectiontransformationprogramme/futureofpopulationandsocialstatistics/administrativedatacensusannualassessments#:...:text=Our%20research%20to%20transform%20population,to%20be%20a%2.

Abbreviations

ABP:	AddressBase Premium
ASSIGN:	AddreSS matchInG to unique property reference Numbers
CCG:	Clinical Commissioning Group
CEG:	Clinical Effectiveness Group
DS:	Data Service
HER:	Electronic Health Records
EMIS:	Egton Medical Information Systems
GP:	General Practice/Practitioner
ICB:	Integrated Care Board
MSSQL:	Microsoft SQL
NCMP:	National Child Measurement Program
NEL:	North East London
NHS:	National Health Service
QMUL:	Queen Mary University of London
RALF:	Residential Anonymised Linkage Field
SAIL:	Secure Anonymised Information Linkage
UPRN:	Unique Property Reference Number

Supplementary Appendices

Supplementary Appendix 1

Household method logic coded in Python

```
#!/usr/bin/env python3

# new_pae.py
# Script takes 1 parameter: event_date (date to define place of residence )
# eg python3 new_pae.py '2012-01-01'

# Import libraries/modules
import pymssql  # interacting with Microsoft SQL Server
import pandas as pd  # data manipulation and analysis
import configparser  # reading configuration files
from datetime import datetime  # datetime class from the datetime module
import sys

##############################################################################
#############################
#           FUNCTIONS
##############################################################################
#############################

#### function DH() converts a date string ("YYYY-MM-DD") into a numeric (ret)
# runs faster than the pandas.to_datetime()
def DH(date):
    if (date == "None"): return 0  # Return 0 if the date is "None"

    # Split the date string into year, month, day and convert to integer
    z = date.split("-")
    year, month, day = z[0], z[1], z[2]
    y, m, d = int(year), int(month), int(day)

    # Calculate the number of leap years (r) and days (ret) since 1840 and days in February (leap)
    r = round((y - 1) // 4) - (round((y - 1) // 100)) + (round((y - 1) // 400)) - 446
    ret = 366 * r + ((y - 1841 - r) * 365) + d
    leap = 29 if (y % 4 >0) else 28 if (y % 100 >0) else 29 if (y % 400 >0) else 28

    # Add the number of days for each month (var) prior to the current month
    var = [31, leap, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31]
    for i in range(len(var)):
        m = m - 1
        if (m == 0): break
        ret = ret + var[i]

    return ret  # Return the numerical representation of the date (ret)

#### function read_ini_extra() reads configuration from an INI file or a dictionary array.
# Extracts database connection parameters, NAME and DEBUG settings.
def read_ini_extra(file_path, dict_obj=None):
    global ZDB, ZUSER, ZPASS, ZHOST, ZPORT  # Declare global variables to store database connection parameters

    # Use ConfigParser() to read ini file or dictionary
    config = configparser.ConfigParser()
    if dict_obj:
        config.read_dict(dict_obj)
    else:
        config.read(file_path)
```

```
    # Get the value of DEBUG setting and convert it to boolean
    debug = config["APP"].getboolean("DEBUG")

    # Get the value of NAME setting
    name = config.get('APP', 'NAME', fallback='NAME is not defined')

    # Get the database connection parameters
    ZDB = config["DATABASE"].get("DB")  # Database name
    ZUSER = config["DATABASE"].get("USERNAME")  # Database username
    ZPASS = config["DATABASE"].get("PASSWORD")  # Database password
    ZHOST = config["DATABASE"].get("HOST")  # Database host
    ZPORT = config.get('DATABASE', 'PORT', fallback='PORT is not defined')  # Database port with a fallback value if not defined

    return debug  # Return the DEBUG setting (debug)

#### function GMSV3 determines the GMS registration status of a patient based on the event date.
# Checks if the patient exists and not deceased. Iterates through the patient's registration record (episodes of care) to determine their status.
def GMSV3(nor, event_date):
    b = 2
    zevent_date = DH(event_date) # Use DH() to convert event_date to numeric

    # Check if patient exists + if patient is deceased
    if (nor not in patient):
        return 4  # Patient not found

    dod = patient[nor][0] # Get date of death

    if (not str(dod) == "None"):
        dod_h = DH(str(dod))
        if (zevent_date >= dod_h):
            return b  # Patient is deceased

    # Iterate through the patient's episodes of care record
    # Check the episode is for a Regular/GMS registration (not dummy, emergency, temporary etc)
    # Check if the event date falls within or on the episode start and end dates + episode was active
    for i in x[nor]:
        z = i.split("~")
        id = z[0]
        date_start = z[1]
        date_end = z[2]
        type = z[3]

        d1 = DH(date_start)
        d2 = DH(date_end)

        if (type != "1335267"):
            continue # If registration type is not 1335267 (Regular/GMS patient), skip

        if (not d1==0 and d1 >zevent_date):
            continue # If the episode start date is after the event date, skip

        if (d1 < = zevent_date and d2 == 0):
            b = 1
            break # episode was active on the event date

        if (d2 >= zevent_date and d2 >= d1):
            b = 1
            break # event date falls within the episode date range
```

```
        if (d1 < zevent_date and d2 < d1):
            b = 3
            break # event date is after the end date of the episode

    return b  # Return a status code (b)

##################################################################################
#############################
#           FETCH PATIENT / ADDRESS DATA
##################################################################################
#############################
patient = {}; x = {}; adr = {}; adridx = {}; matchbig = {}  # Initialize dictionaries for storing data

# Connect to the database + set start time + clear the patient dictionary
ret = read_ini_extra("/tmp/bob.ini") # Use read_ini_extra() to get debug mode from INI file
conn = pymssql.connect(server=ZHOST, user=ZUSER, password=ZPASS, database=ZDB)
zstart = datetime.now()

########## patient
##################################################################################
##########
patient.clear()
# Execute SQL query to fetch patient data from database in batches of 1000000
cursor = conn.cursor()
cursor.execute('SELECT id, date_of_death FROM [compass_gp].[dbo].[patient] ORDER BY id OFFSET 3000001 ROWS FETCH
NEXT 1000000 ROWS ONLY;')

# Loop through each patient data record and add the patient ID and date of death into the patient dictionary
row = cursor.fetchone()
c = 1
while row:
    if (c % 10000 == 0):
        print(c);  # Print progress every 10000 rows
    id = row[0] # Patient ID
    date_of_death = row[1]
    patient[id] = [date_of_death]
    c = c + 1
    row = cursor.fetchone()
print(c)  # Print the total number of fetched rows

########## patient registration record (episode_of_care)
#################################
x.clear()
# Execute SQL query to fetch episode_of_care data from database into dictionary (x)
cursor = conn.cursor()
cursor.execute('SELECT id, patient_id, registration_type_concept_id, date_registered, date_registered_end FROM [compass_gp].[dbo
[episode_of_care];')

# Loop through each episode_of_care data record and add the episode_of_care ID, patient ID, registration type ID and
registration start/end dates for each patient
row = cursor.fetchone()
c = 1
while row:
    if (c % 10000 == 0):
        print(c)  # Print progress every 10000 rows
    id = row[0]  # Episode of care ID
    patient_id = row[1]
    type = row[2]
    date_start = row[3]
    date_end = row[4]
```

```
    if (patient_id in patient): # Check the episode of care patient ID matches with patient ID in the patient dictionary
        x.setdefault(patient_id, []).append(str(id) + "~" + str(date_start) + "~" + str(date_end) + "~" + str(type))  # Add
episode data to the dictionary
    c = c + 1
    row = cursor.fetchone()
print(c)  # Print the total number of fetched rows

########## patient address
############################################################################
adr.clear()
adridx.clear()
# Execute SQL query to fetch patient address data from database into dictionaries (adr) and (adridx)
cursor = conn.cursor()
cursor.execute('SELECT id, patient_id, start_date, end_date, use_concept_id, lsoa_2011_code, msoa_2011_code FROM
[compass_gp].[dbo].[patient_address];')

# Loop through each patient address record
# Add (for each patient) address ID to dictionary (adridx) and patient ID, residency start/end date, use ID, LSOA and MSOA, as
a list, to dictionary (adr)
row = cursor.fetchone()
c = 1
while row:
    if (c % 10000 == 0):
        print(c)  # Print progress every 10000 rows
    id = row[0]  # Address ID
    patient_id = row[1]
    start_date = row[2]
    end_date = row[3]
    use = row[4]
    lsoa = row[5]
    msoa = row[6]
    if (patient_id in patient): # Check the patient_address patient ID matches with patient ID in the patient dictionary
      adr.setdefault(patient_id, []).append([id, str(start_date), str(end_date), str(use), str(lsoa), str(msoa), DH(str(start_date))])
# Add to dictionary adr(list)
        adridx[id] = [] # Add to dictionary (adridx)
    c = c + 1
    row = cursor.fetchone()
print(c)  # Print the total number of fetched rows

########## patient address match
############################################################################
matchbig.clear()
# Execute SQL query to fetch patient address match data from database into dictionaries
cursor = conn.cursor()
cursor.execute('SELECT id, patient_address_id, uprn, qualifier, uprn_property_classification FROM [compass_gp].[dbo].
[patient_address_match];')

# Loop through each patient address match record
# Add (for each address) match ID, address ID, uprn, qualifier and classification to dictionary (matchbig)
row = cursor.fetchone()
c = 1
while row:
    if (c % 10000 == 0):
        print(c)  # Print progress every 10000 rows
    match_id = row[0]
    adr_id = row[1]
    if (adr_id not in adridx): # Check if the address ID exists in the adridx dictionary
        row = cursor.fetchone()
        continue
    uprn = row[2]
    qualifier = row[3]
```

```
    classification = row[4]
    if (adr_id in matchbig): # Update dictionary (matchbig) with the latest match information for each address ID
        m_id = matchbig[adr_id][0]
        if (match_id >m_id):
            matchbig[adr_id] = [match_id, uprn, qualifier, classification]
    else:
        matchbig[adr_id] = [match_id, uprn, qualifier, classification]
    c = c + 1
    row = cursor.fetchone()
print(c)  # Print the total number of fetched rows

##########################################################################
############################
#          MATCH PATIENT ADDRESS
##########################################################################
############################
event_date = str(sys.argv[1]) # takes input parameter
zevent_date = DH(event_date) # Use DH() to convert event_date into a numeric
uprn = ""

# Open a file for output and add the required headers
outputFile = open("/tmp/hh_output.txt", "w")
outputFile.write("patient_id\taddress_id\tuprn\taddress_start_date\taddress_end_date\taddress_use\tproperty_classification
\tlsoa\tmsoa\n")

# Define dictionary (VPROP) with property codes
VPROP = {"R": "", "RD": "", "RD01": "", "RD02": "", "RD03": "", "RD04": "", "RD06": "", "RD07": "", "RD10": "",
"RH02": "", "U": "", "UC": "",  "UP": "", "X": ""}
# R = Residential
# RD = Dwelling
# RD01 = Caravan
# RD02 = Detached
# RD03 = Semi-Detached
# RD04 = Terraced
# RD06 = Self Contained Flat (Includes Maisonette / Apartment)
# RD07 = House Boat
# 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

# Loop through each patient in the patient dictionary
c = 1  # Initialize a counter for progress tracking
for nor in patient:
    if (c % 10000 == 0):
        print(c)  # Print progress every 10000 patients
    c = c + 1
    ret = GMSV3(nor, event_date)# Use GMSV3() to determine the registration status of the patient
    if (ret == 2): # If the patient is dead (status code 2), write the patient ID and status to the output file
        outputFile.write(str(nor) + "\t2\n")
        continue
    if (nor in adr):# If the patient is in adr dictionary, sort addresses based on start date and address ID
        blist = sorted(adr[nor], key=lambda x: (x[6], x[0]), reverse=True)

        # SubLoop through each address associated with the patient
        # Extract patient_id, address ID, start date, end date, use concept ID, LSOA and MSOA into list (blist)
        for i in range(len(blist)):
            id = int(blist[i][0])
            start_date = blist[i][1]
```

```
        end_date = blist[i][2]
        use = blist[i][3]
        lsoa = blist[i][4]
        msoa = blist[i][5]

        #1335358 = Home
        #1335360 = Temporary
        #1335361 = Old / Incorrect
        if (use == "1335360"):# If address concept ID is 1335360 (temporary address), skip
            continue

        # If the address ID exists in matchbig dictionary, extract UPRN, qualifier, and classification
        if (id in matchbig):
            uprn = matchbig[id][1]
            qualifier = matchbig[id][2]
            classification = matchbig[id][3]

        if (uprn == ""): # If UPRN is empty, skip
            continue

        if (classification not in VPROP.keys()): # If the classification is not in VPROP dictionary keys, skip
            continue

        if (qualifier != "Best (residential) match"): # If the qualifier is not "Best (residential) match", skip
            continue

        d1 = DH(start_date)  # Convert start date to numeric
        d2 = DH(end_date)  # Convert end date to numeric

        # If the event date is within the date range of the address, write address details to the output file
        if (d1 <= zevent_date or d1 == 0) and (d2 >= zevent_date or d2 == 0):
            outputFile.write(str(nor) + "\t" + str(id) + "\t" + str(uprn) + "\t" + str(start_date) + "\t" + end_date + "\t"
            + use + "\t" + classification + "\t" + lsoa + "\t" + msoa + "\n")
            break  # Exit the SubLoop back into the main Loop

# Close the output file + Database Connection. Print run times
outputFile.close()
conn.close()
print(zstart)
print(datetime.now())
```

Supplementary Appendix 2: Proportions of the combinations of the four reasons a household RALF was not assigned to 10,025 children in the use case cohort

	Multiple different valid household UPRNs	No valid household UPRNs	No address records at event date	Not alive or no regular registrations at event dates	Total	%
0000	0	0	0	0	6	0.1
0001	0	0	0	1	3,091	30.8
0100	0	1	0	0	2,027	20.2
0101	0	1	0	1	3,506	35.0
0110	0	1	1	0	593	5.9
0111	0	1	1	1	379	3.8
1000	1	0	0	0	188	1.9
1001	1	0	0	1	229	2.3
1100	1	1	0	0	1	0.0
1101	1	1	0	1	5	0.0
Total					10,025	100

Note: the combinations are in binary format where 1 = any of the person's multiple address records met the criteria.