International Journal of Population Data Science

Journal Website: www.ijpds.org

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Evaluation of the ASSIGN open-source deterministic address-matching algorithm for allocating unique property reference numbers to general practitioner-recorded patient addresses

Gill Harper[1,*], David Stables[2], Paul Simon[2], Zaheer Ahmed[1], Kelvin Smith[1], John Robson[1], and Carol Dezateux[1]

Submission History

Submitted:	11/04/2021
Accepted:	19/10/2021
Published:	08/12/2021

[1]Clinical Effectiveness Group, Centre for Primary Care, Wolfson Institute of Population Health, Barts and the London School of Medicine and Dentistry, Queen Mary, University of London
[2]Endeavour Health Charity, Low Farm, Barthorpe, Malton, YO17 9RW

Abstract

Introduction

Linking places to people is a core element of the UK government's geospatial strategy. Matching patient addresses in electronic health records to their Unique Property Reference Numbers (UPRNs) enables spatial linkage for research, innovation and public benefit. Available algorithms are not transparent or evaluated for use with addresses recorded by health care providers.

Objectives

To describe and quality assure the open-source deterministic ASSIGN address-matching algorithm applied to general practitioner-recorded patient addresses.

Methods

Best practice standards were used to report the ASSIGN algorithm match rate, sensitivity and positive predictive value using gold-standard datasets from London and Wales. We applied the ASSIGN algorithm to the recorded addresses of a sample of 1,757,018 patients registered with all general practices in north east London. We examined bias in match results for the study population using multivariable analyses to estimate the likelihood of an address-matched UPRN by demographic, registration, and organisational variables.

Results

We found a 99.5% and 99.6% match rate with high sensitivity (0.999,0.998) and positive predictive value (0.996,0.998) for the Welsh and London gold standard datasets respectively, and a 98.6% match rate for the study population.

The 1.4% of the study population without a UPRN match were more likely to have changed registered address in the last 12 months (match rate: 95.4%), be from a Chinese ethnic background (95.5%), or registered with a general practice using the SystmOne clinical record system (94.4%). Conversely, people registered for more than 6.5 years with their general practitioner were more likely to have a match (99.4%) than those with shorter registration durations.

Conclusions

ASSIGN is a highly accurate open-source address-matching algorithm with a high match rate and minimal biases when evaluated against a large sample of general practice-recorded patient addresses. ASSIGN has potential to be used in other address-based datasets including those with information relevant to the wider determinants of health.

Keywords

data linkage; electronic health record; addresses; address-matching; quality assurance; population health; place-based health

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

https://doi.org/10.23889/ijpds.v6i1.1674

Introduction

Data linkage is being increasingly used in health data science, with growing examples of spatial linkage of electronic health records (EHRs) to environmental information for population health research [1–4]. Address-matching is data linkage that enables spatial linkage by specifically matching non-standardised addresses recorded in an administrative dataset to a reference address gazetteer that provides standardised address formats, property reference numbers, and geographic co-ordinates.

Linking places to people is a core element of the UK government's geospatial strategy [5]. In 2019, the Public Sector Geospatial Agreement [6] gave more than 5,000 public sector organisations unlimited access to Ordnance Survey data, including Unique Property Reference Numbers (UPRNs) - the unique identifier for every addressable location in Great Britain. UPRNs are described as the 'golden thread' which links datasets together, with the potential 'to underpin huge advances in our digital society, improving our lives and equipping the economy to recover from the effects of Coronavirus' [5]. UPRNs are now a mandated standard across the public sector, however challenges remain to implement this fully within the National Health Service (NHS) enabling geospatial linkage for research, innovation, and public benefit.

The UPRN acts as an address standardiser, a household identifier, and a high-resolution geocoder, and ultimately as the granular spatial link to environmental information to be used for direct patient care as well as for health research. Address-based geography using UPRNs moves away from the acknowledged limitations of area-based geography ecological approaches and enables more accurate patient-level analysis of the effect of geographical and household exposures and covariates on health outcomes.

Robust methods are important for linking addresses in health data to UPRNs. Schinasi et al. (2018) [4] concluded that such linkage is a major research opportunity, and that future research should include more detailed descriptions of methods used to geocode addresses and for dealing with missing or poor quality geographic information. They recommended assessment of the extent and impact of biases including the adoption or design of formal methods to assess the extent to which patterns of missing geographic data will lead to biased results.

While other address-matching algorithms are available in the UK [7–9] few, if any, have been developed specifically for patient recorded addresses available in EHRs, and their methods, accuracy and potential biases are often not transparent or evaluated, limiting the extent to which users of address-matching results can be aware of and assess implications for analyses.

From our experience we propose that there are five general factors that will affect the match rates, quality, and bias of match success of any address-matching algorithm:

1. How the address is provided, i.e. the quality of the address provided by the patient when registering with respect to its completeness, spelling mistakes or omissions.

2. How the address is recorded, i.e. manually by a data entry person using free-text or auto-fill prompts, and the level of attention to accuracy when doing this.

3. The content and quality of the property gazetteer being matched to, i.e. whether the gazetteer is up-to-date and complete.

4. The geography of the address as some geographic areas are more prone to variation or errors in how the address is provided that may differ from the standardised address in the property gazetteer. For example, apartment numbering can be represented in multiple ways or properties in rural areas can be addresses with a house name or a number.

5. The matching algorithm, i.e. the quality and appropriateness of the method used to find a match.

We describe ASSIGN (**A**ddre**SS** Match**I**n**G** to Unique Property Reference **N**umbers), an address-matching algorithm specifically designed, developed and validated by Dr Gill Harper and Dr David Stables for the linkage of patient addresses as routinely recorded in EHRs to the UPRNs in the Ordnance Survey Great Britain property gazetteer database AddressBase Premium (ABP) [10]. ASSIGN as implemented in the north east London Discovery Data Service (DDS) enables the UPRNs from ABP to be assigned to each patient address in near real-time and subsequent changes to patient addresses and gazetteer databases to be automatically updated as required.

Overall, our objective was to transparently describe and quality assure the ASSIGN address-matching algorithm and examine potential biases in match results so that users of the algorithm and its outputs have this information available to them and have clarity on how their analyses may be affected by it.

If an address-matching algorithm is not accurate, an incorrect UPRN can result in the incorrect residential location being attributed to a patient. This can result in misclassification of environmental exposure estimates and consequently in epidemiologic affect estimates, potentially systematically, for example of air pollution exposure on asthma related emergency department visits [11]. It can also result in misassignment of occupants to a UPRN which when used as a proxy of a household, can introduce error in studies where the household occupancy or type is the risk factor, for example in COVID-19 studies [12–14].

Knowledge of address-matching algorithm accuracy and error supports confident use of UPRNs not only within EHRs but across the growing variety of sectors who are moving towards the implementation of UPRNs on their address data.

Methods

Study population

The study population comprises 1,757,018 patients aged $\geq$18 years, alive and currently registered as at census date 16$^{\text{th}}$ November 2020 with one of 277 general practices providing primary care services to the entire geography covered by seven north east London Clinical Commissioning Groups

(CCGs), all of which publish primary care EHR data on a daily basis into the north east London DDS and associated subscriber database. These patients were recorded as living at 945,196 unique addresses. This includes patients registered in all general practices in City and Hackney, Newham, Tower Hamlets, Waltham Forest, Barking and Dagenham, Havering, and Redbridge. At the time of sampling, 257 practices used the Egton Medical Information Service (EMIS) [15], 12 the SystmOne [16], and eight the Vision [17] clinical record supplier systems.

Address-matching algorithm

The ASSIGN algorithm was developed by exploiting the address-matching experience of the designers and with inner north east London GP recorded patient addresses as the test addresses. Repeated checks of false positives and false negatives were made to inform coding improvements and increase match rate and accuracy with each iteration. The input address to be matched is named the 'candidate' address, and the addresses in ABP to be matched to are named the 'standard' addresses. The method consists of three stages: reformat, match and return.

Reformat

The ABP files are loaded into a database and mapped directly to the combination of eleven standard address object fields that exist across both the Royal Mail Delivery Point Address (DPA) and local authority Local Property Identifier (LPI) versions of addresses in ABP: flat, building, number, dependent thoroughfare, street, dependent locality, locality, town, postcode, organisation, vertical, concatenating where required.

These are stored and heavily indexed using a set of single and compound indexes designed to improve search performance at run time. In addition, certain performance improving indexes are generated based on semantic equivalence or semantic importance. Examples include correcting spelling errors, de-pluralisation, replacing or removing punctuation and lower casing, and removing extraneous words that are unnecessary in the match process, for example, the range of words that are equivalent to the word 'flat' such as 'apartment' or 'maisonette'.

When a candidate address is submitted the address string is parsed using a combination of Regex [18] matching expressions and index checking to form the same eleven address object fields. For example, the postcode is identified by checking the format and position in the string (postcodes are usually submitted at the end of the string or in a separate comma delimited field). A further candidate address version is created by applying the same reformatting techniques as applied to the standard addresses, so that both the eleven address fields non-formatted and the eleven address fields formatted candidate addresses are available to the algorithm.

The final reformatting step is positional checking, for example, a candidate address abbreviation 'st' would be mapped to 'street' as a spelling correction, but not if it was presented as the first word in a field 'St David's' for example would be retained as 'St David's'.

Match

The objective of the matching algorithm is to reach a high level of confidence that the matched candidate address refers to the same location as the standard address and more so than any other available standard address. Blocking by matching postcode area, potential matching standard addresses are 'tried on for size' deterministically by applying matching judgement rules in rank order. The rules that are applied are determined on the content of the candidate address string and the text manipulation required. Higher ranking rules have required the least amount of address string manipulation, so that rank 1 is an exact word for word match for the entire address string. Rank 1 is the most frequent match rule.

These rules mirror human pattern recognition and manipulate and compare the address strings until the best available match is found. Human pattern recognition refers to knowing that similar or the same words in different orders, or transposed characters, or the correct spelling of a misspelled word usually means the same thing. The algorithm codes these using, for example, Levenshtein distance [19], pattern matching with Regex [18], field swapping and pluralisation.

The algorithm can be considered as a decision tree handling a combination of ANDs, ORs or NOTS with branching occurring on the OR conditions. The nodes of the trees relate to comparison of the different address fields and are pass/ fail tests and travelling down one of the next branches means a test has been passed. If a test fails the process goes back up the feeder branch to the next branching node, and tries the next untried branch, until all branches are exhausted. This has similarity to a tableaux tree [20] except the nodes branch on human judgement-based decision making rather than pure logic.

A match is made with one of four overall qualifiers that qualifies the relationship between the candidate address and the matched standard address in relation to approximate geography, or no match is made. The qualifiers are:

1. Best match: the closest match out of all available
2. Child: candidate address is a 'child' sub-property of the UPRN it has been matched to
3. Parent: candidate address is the 'parent' building shell of the UPRN it has been matched to
4. Sibling: candidate address is a near neighbour of the UPRN it has been matched to

Return

Where there is a match, the algorithm returns the UPRN, the overall qualifier, the standard address, the match pattern, and match rule identifier employed to get that match. The match rule is a label identifying which section of the code made the match, and the match pattern depicts how five address objects were manipulated to achieve the match. These five address objects are merged from the original eleven: flat, building, number, street, postcode. Twelve possible match terms (Table 1) exist and can be combined in up to 50 different ways on the five address fields. These are restricted to plausible terms, for example, postcodes are never swapped with streets.

Table 1: The twelve match terms applied to the five address fields to describe the match pattern

Match term	Character	Description
mapped also to	&	Indicates a match using more than one candidate field
moved to	>	Means that the candidate field was moved to another field to match e.g. number moved to flat
moved from	<	Means that the candidate field was moved from another field to match on this field
field merged	f	when moved from and to, the fields are then merged to match
ABP field ignored	i	ABP field was ignored in order to match i.e. the ABP address contained more precise detail than the candidate but was unnecessary in order to match. This usually means that the candidate field is null
Candidate field dropped	d	The candidate field was dropped in order to match i.e. the candidate address has more precise detail than the authority address. The ABP address would probably be null
Matched as parent	a	The candidate field matched as being at a higher level than the ABP field, for example flat 6 matching to flat 6a
Matched as child	c	The candidate field matched as being at a lower level than the ABP field, for example candidate flat 6a, ABP flat 6
Partial match	p	The candidate field was partially matched to the ABP field (or vice versa) typically 2 out of 3 words
Possible spelling error	l	The candidate field and ABP field were matched using the Levenshtein distance algorithm taking account of misspellings
Level based match	v	The level of a flat in a building (vertical from the street) was used to create the match e.g. 2b for second floor b
Equivalent	e	The fields are equivalent, albeit not necessarily spelled the same, using various equivalence lists, word swaps, word drops etc

An example of a match pattern is 'Pe,Se,Ne,Bp,Fe'. This means that the postcode, street, number, and flat fields were equivalent matches between the candidate and standard address, and the building field was a partial match between the candidate and standard address.

The ASSIGN algorithm code is available as fully open-source [21] for free use and information on the algorithm method [22] is freely provided for users. Supplementary Appendix 1 describes ASSIGN in the GUILD [23] format.

ASSIGN seeks to match the input addresses to addresses in Ordnance Survey's AddressBase Premium (ABP) [10]. This is a comprehensive property gazetteer of all current, historic and future addresses, properties and land areas in Great Britain. Each property address is recorded in national standard BS7666 [24] format and is represented by a Unique Property Reference Number (UPRN). Property classification type and geographic co-ordinates are also provided for each UPRN. Updates to ABP are provided by the Ordnance Survey every six weeks as Epochs. When the ASSIGN algorithm matches the candidate address to a UPRN, metadata relating to the match and variables of interest from ABP is assigned. Match metadata is listed in Table 2.

ASSIGN is designed so that only records in ABP that are of relevant property types are made available for matching. These include all residential property types and a considered selection of commercial property types that we found can be given as a person's place of residence for example if they live above a public house. The choice of property types can be varied as required, for example, if matching patient addresses solely to commercial addresses such as care homes. We evaluated Version 4.2.1 of the ASSIGN algorithm and Epoch 75 of AddressBase Premium which were implemented in the north east London DDS at the time of data extraction.

Data sources

Gold standard reference datasets

The algorithm was run on two 'gold standard' external reference address datasets with previously assigned and verified UPRNs in order to calculate accuracy rates. The first of these datasets comprised 9,177 local authority sourced addresses in Wales, and the second 9,475 local authority sourced addresses from the London Borough of Tower Hamlets in north east London. The ASSIGN algorithm has been developed using north east London patient addresses. Therefore, addresses from the rural geography of Wales and from local authority sourced addresses that tend to be of poorer address quality than patient addresses were considered to be a challenging test of ASSIGN's performance.

North east London Discovery Data Service (DDS) subscriber database

For all analyses reported in this paper, we used identifiable data from general practitioner electronic health records data which are held in the DDS subscriber database and curated by the Queen Mary University of London based Clinical Effectiveness Group (CEG). The GP EHR data are provided daily from GP system suppliers to the CEG DDS database and contain demographic and clinical data and address history for each patient registration. Approval for access to the person identifiable data (patient addresses) used in this study was provided by the DDS data controllers to the CEG as appointed data sub-processors for the purpose of developing and evaluating the ASSIGN algorithm for direct patient care purposes only. This access was limited to approved individuals

Table 2: Unique property reference number (UPRN) match metadata

Metadata field	Description
From ASSIGN:	
Algorithm version	Version of algorithm used
Match date	Date match made
Qualifier	One of four match qualifiers: best match, child, parent, sibling
Match rule	Label identifying which section of code made the match
Match pattern	The combination of manipulation qualifiers used on each of the five address fields used to make the match
From ABP:	
UPRN	Unique Property Reference Number, from ABP
Epoch	ABP Epoch used
Property classification	The property classification type, from ABP
x-coordinate	UPRN geographical easting coordinate, from ABP
y-coordinate	UPRN geographical northing coordinate, from ABP
latitude	UPRN geographical latitude coordinate, from ABP
longitude	UPRN geographical longitude coordinate, from ABP
ABP address	UPRN associated address string, from ABP

ABP = AddressBase Premium.

with appropriate information governance training working in a secure trusted data environment.

Primary outcome

The primary outcome was a binary variable indicating whether a UPRN had been matched or not matched to the patient address using the ASSIGN algorithm.

Explanatory variables

We selected a range of patient level demographic and registration characteristics, and organisational features to evaluate match rates and biases. These are listed in Table 3.

Statistical methods

In the absence of a formal standard method to evaluate address-matching algorithms, we considered the GUILD [23] data linkage reporting principles to be a relevant framework for this purpose because address-matching is fundamentally a data linkage exercise linking address data between two sources to find a match. GUILD proposes which information may be required at each step of the linkage pathway to improve the transparency, reproducibility, and accuracy of linkage processes, and the validity of analyses and interpretation of results. We follow this framework as much as possible, in particular for the calculation of match and accuracy rates.

We applied ASSIGN to the two gold standard external reference datasets and calculated the data linkage accuracy metrics described in Table 4. We estimated the match rate obtained from applying ASSIGN to the 945,196 distinct patient addresses from our study population.

Descriptive summary statistics by three age bands (18–19, 20–64 and ≥65), five ethnic groups (White, South Asian, Black, Other (including Chinese and Mixed) and Not Stated), Sex (female, male, other) and IMD 2019 score quintile (1 = most deprived, 5 = least deprived) were calculated for the entire study population, separately for those with and without an ASSIGN-matched UPRN, in order to compare the characteristics of each group, including those with missing data. In total, 268,382 had missing values across these four variables, the majority from missing ethnic groups.

The absolute difference in the proportion matched, relative to the reference group for each explanatory variable was calculated. We considered an absolute difference in match rates of 1% or greater to be potentially an important difference.

We performed a Poisson multilevel mixed-effects generalized linear model in a complete case analysis to estimate UPRN match prevalence ratios and their 99% confidence intervals after mutual adjustment for all explanatory variables described previously, including GP practice as a random effect. We explored between general practice variation in match rates.

All analyses were conducted using Stata/MP 15 (StataCorp LP).

Results

Match quality

When assessed against the Welsh and Tower Hamlets gold-standard datasets, the match rates were, respectively, 99.5% and 99.6%; the sensitivity 0.999 and 0.998; the positive predictive value 0.996 and 0.998; and the F-measure 0.997 and 0.998. Overall, there were 35 (0.38%) incorrect matches and 12 (0.13%) missed true matches in the Welsh dataset and 16 (0.17%) incorrect matches and 20 (0.21%) missed true matches in the Tower Hamlets dataset.

The ASSIGN algorithm matched 924,094 (98%) of the 945,196 unique patient addresses in the study population to a UPRN. ASSIGN processed 38,000 records per minute.

Table 3: The demographic, GP registration and GP organisational explanatory variables

Demographic variables:	
Age on census date (16/11/2020)	Years
Self-reported ethnic group	NHS 16 + 1 classification [25]
Sex	Male, female, other
Deprivation	LSOA level IMD 2019 quintiles
Mobility	Number of different GP registrations in previous 12 months, number of address changes in previous 12 months
GP registration variables:	
Age at registration	Years (<1, 1–14, 15–29, 30–64, 65–84, 85 and over);
Duration of registration	Days (quartiles)
GP organisational variables:	
Commissioner	GP practice Clinical Commissioning Group (CCG)
EHR supplier system	EMIS, SystmOne, or Vision

LSOA = Lower Super Output Area, IMD = Index of Multiple deprivation, CCG = Clinical Commissioning Group, EHR = Electronic Health Record.

Table 4: Data linkage accuracy metrics (modified from GUILD [23])

Accuracy metric	**Description**
Positive Predictive Value (PPV)	The proportion of record pairs classified by the algorithm as links that are true matches. Also known as precision.
Sensitivity	The proportion of true matches that are correctly classified as links. Also known as recall.
F-measure	The harmonic mean between positive predictive value and sensitivity. Often used to compare the overall efficiency of a method. F-measure = 2*(PPV*sensitivity)/(PPV + sensitivity)

Those addresses without a match were more likely in specific postcode areas, and for invalid addresses or postcodes, or address strings beginning with an alphabetic character indicating a flat rather than a house. Full details on the GUILD reporting of match and accuracy rates are provided in the Supplementary Appendix 1.

Population characteristics

An ASSIGN matched UPRN was available for 1,731,920 (98.6%) of the 1,757,018 adults in the study population. Supplementary Appendix 2 shows the UPRN matched and unmatched rates for age at census date 16$^{\text{th}}$ November 2020, ethnic background, sex, and IMD 2019 quintile for the study population. Around half (49.2%) were female, 85.3% were aged 20 to 64 years at the census date, and 41.7% were from White, 24% South Asian, 11% Black, or 6.7% Other ethnic groups. The majority (67.5%) lived in the two most deprived IMD 2019 quintiles, with 24.4% living in the most deprived quintile, reflecting the high levels of social disadvantage in north east London. Higher proportions of an unsuccessful UPRN match were found for men, those aged 20 to 64 years at the census date, or from the Other ethnic group.

Match rates and absolute differences

Absolute match rate differences to the reference groups greater than 1% were found for people aged 15–29 or ≥85 years, from Chinese or Not Stated ethnic groups, with a missing IMD 2019 quintile or GP registration duration, with the longest GP registration duration quartiles, with ≥2 address changes in the previous 12 months, or who were registered with a GP practice using the SystmOne clinical record system or registered with a GP practice in Tower Hamlets.

The match rate was consistently high with a minimum of 94.4%, and match rates were similar for any missing and non-missing categories, with the exception of the 0.2% of the study population with missing IMD 2019 values which had a substantially lower match rate of 23.5%. As the IMD score is assigned via the postcode, if this is missing or of poor quality, it is also likely that a UPRN cannot be assigned. The match rate in those with missing ethnicity codes (n = 265,525) was similar to that reported for those from White ethnic groups.

Full details of the UPRN match rates and absolute difference in the proportion matched relative to the reference group, for all explanatory variables and the complete study population are given in Supplementary Appendix 3.

Bias in UPRN match success

The adjusted complete case analysis prevalence ratios and 99% confidence intervals are presented in Table 5, which excludes 278,875 patients with missing data, the majority excluded due to missing ethnicity codes.

Based on absolute differences greater than 1% from the reference category, people aged 15-29 or 85 years and over, those of Chinese ethnic background, with ≥3 address changes

Table 5: Absolute differences in percentage of population matched to a UPRN, adjusted complete case analysis prevalence ratios and 99% CIs with respect to reference category by demographic, GP registration and organisational characteristics

	Number *n*	**Absolute difference relative to reference group (%)**	**Prevalence ratio**	**99% CI lower**	**99% CI upper**
Patient age at registration (years)					
<1	**30,029**	**Ref**			
1–14	93,868	−0.13	0.999	0.997	1
15–29	485,945	**−1.46**	0.986	**0.982**	**0.989**
30–64	811,582	−0.8	0.992	0.990	0.994
65–84	52,385	−0.55	0.995	0.992	0.997
85 and over	4,334	**−1.94**	0.981	**0.966**	**0.995**
Ethnic background					
British	**375,405**	**Ref**			
African	100,142	0	1	0.998	1.002
Any other Asian background	60,999	−0.2	0.998	0.994	1.002
Any other Black background	43,764	0.28	1.003	1.000	1.005
Any other White background	336,182	−0.31	0.997	0.995	0.999
Any other ethnic group	52,610	−0.31	0.997	0.993	1.001
Any other Mixed background	14,944	−0.8	0.992	0.988	0.996
Bangladeshi	145,379	0.55	1.006	1.003	1.009
Caribbean	47,653	0.43	1.004	1.002	1.006
Chinese	21,819	**−3.26**	0.968	**0.946**	**0.989**
Indian	120,431	−0.11	0.999	0.994	1.003
Irish	12,945	−0.22	0.998	0.994	1.002
Not stated	25,826	−0.79	0.993	0.987	0.999
Pakistani	93,146	0.26	1.003	1	1.005
White and Asian	4,905	−0.51	0.995	0.990	1
White and Black African	9,918	−0.62	0.994	0.986	1.002
White and Black Caribbean	12,075	−0.42	0.996	0.991	1.001
Sex					
Female	**736,398**	**Ref**			
Male	741,745	−0.19	0.998	0.997	0.999
IMD 2019 quintile					
1 (most deprived)	**367,429**	**Ref**			
2	666,104	0.07	1.001	0.998	1.003
3	268,097	0.12	1.001	0.997	1.005
4	116,122	−0.25	0.997	0.989	1.005
5 (least deprived)	60,391	0.07	1.001	0.995	1.006
GP registration duration (quartiles)					
1 (shortest)	**386,610**	**Ref**			
2	388,014	0.66	1.007	1.004	1.009
3	381,225	**1.38**	1.014	**1.01**	**1.018**
4 (longest)	322,294	**1.77**	1.018	**1.014**	**1.022**
Number of GP registrations in preceding 12 months					
1	1,336,709	Ref			
2	126,645	0.1	1.001	0.999	1.003
3 or more	14,789	-0.29	0.997	0.992	1.002

Continued.

Table 5: Continued

	Number ***n***	**Absolute difference relative to reference group (%)**	**Prevalence ratio**	**99% CI lower**	**99% CI upper**
Number of address changes in preceding 12 months					
1	**1,083,883**	**Ref**			
2	305,838	−0.69	0.993	0.99	0.996
3 or more	88,422	**−3.12**	0.969	**0.957**	**0.981**
GP system					
EMIS	**1,370,370**	**Ref**			
SystmOne	77,354	**−2.63**	0.973	**0.967**	**0.98**
VISION	30,419	0.52	1.005	0.999	1.012
Clinical Commissioning Group					
Newham	**306,438**	**Ref**			
Barking & Dagenham	122,432	−0.12	0.999	0.993	1.004
City & Hackney	232,840	−0.88	0.991	0.986	0.996
Havering	135,262	0.22	1.002	0.998	1.006
Redbridge	212,088	−0.36	0.996	0.990	1.003
Tower Hamlets	244,643	**−1.35**	0.986	**0.977**	**0.995**
Waltham Forest	224,440	−0.65	0.993	0.987	1

Complete case analysis; $\mathrm{N} = 1{,}478{,}143$.
IMD = Index of Multiple Deprivation.
Quartile definitions for GP registration duration: Quartile 1 (shortest): 0–32 months; Quartile 2: 33–77 months; Quartile 3: 78–183 months; Quartile 4 (longest) $>$ 184 months.
EMIS: Egton Medical Information Systems.
Reference groups and values with an absolute match rate difference to the reference group of $>$1% are in bold.

in the preceding 12 months, registered at a GP practice using SystmOne, or at a GP practice in Tower Hamlets were less likely to have an address matched to a UPRN. Conversely, people registered with their GP practice for more than 6.5 years were more likely to have an address matched to a UPRN than the reference group (Figure 1).

At the practice level, GP practice UPRN match rates ranged from 84.9% to 99.96% with an average of 98.6% (data not shown). The three GP practices with UPRN match rates below 90% included one GP practice for homeless people and two using SystmOne supplier systems. There was no clear association between GP practice UPRN match rate and GP practice list size.

Discussion

Key findings

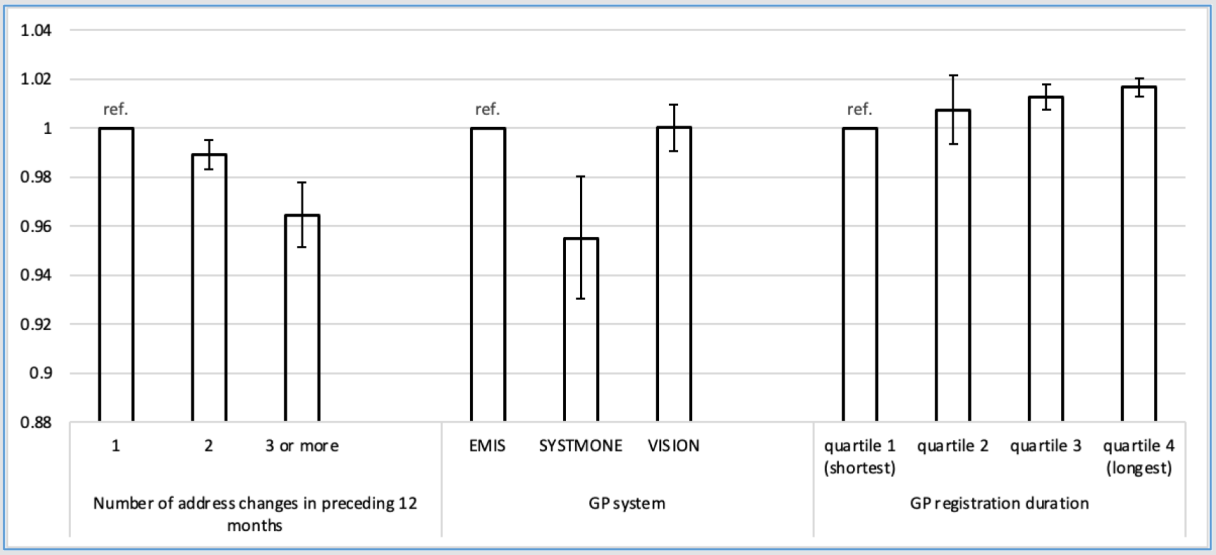
This is to our knowledge the first address-matching algorithm developed specifically to assign UPRNs to patient addresses recorded at registration for NHS general medical practitioner services. Using GUILD [23] specified criteria and methods we have shown that the ASSIGN algorithm achieved a greater than 99.5% match rate in two gold standard datasets drawn from diverse populations with high accuracy as indicated by the sensitivity, PPV and the F-measures. Incorrect matches were extremely low overall, with marginally higher percentages of incorrect matches for the Welsh addresses and of missed true matches in the Tower Hamlets addresses. The high value of the F-measures (0.99) for the ASSIGN algorithm exceeds the threshold of ≥0.8 specified by Ferrante and Boyd (2012) [26] for 'very good' linkage algorithms.

A similarly high match rate (98.6%) was also achieved by ASSIGN when applied to routinely entered GP registered patient addresses for an entire population of predominantly working age, ethnically diverse and socially disadvantaged adults in the complete geography of north east London. We found relatively small differences in some demographic and provider organisation characteristics among the 1.4% patients for whom a match to a UPRN was unsuccessful.

We found that UPRN matching success was less likely among patients aged 15–29 or ≥85 years, and those from Chinese ethnic backgrounds, with missing IMD, who were highly mobile (as assessed by three or more changes in address in the preceding 12 months) or were registered at a GP practice in Tower Hamlets CCG, or using the SystmOne clinical record system. Conversely, UPRN matching success was more likely for patients with missing GP registration dates, or with longer duration of GP registration.

In conclusion, we consider ASSIGN to be a transparent, robust and quality assured address-matching algorithm with a high and accurate match rate with minimal biases in those not matched when evaluated against a whole population

Figure 1: Adjusted prevalence ratios and 99% CIs for number of address changes in the preceding 12 months, GP EHR system, and GP registration duration



dataset of NHS addresses registered as part of routine NHS processes.

Strengths and limitations

Strengths of our study include the use of robust best-practice methods to calculate and evaluate the accuracy of the ASSIGN algorithm using two gold standard datasets from different populations in the UK reflecting very different demographics, geography, and property types. In doing so, we have addressed many of the methodological issues highlighted by Schinasi et al. (2018) [4], by providing a detailed and transparent account of methods we used to geocode addresses, and to evaluate missing or poor quality geographic information, and have undertaken a rigorous evaluation of bias in matching success. To our knowledge, similar accounts of accuracy checks and bias have not been provided by other address-matching algorithms currently in use in the UK.

We evaluated the ASSIGN algorithm in addresses routinely recorded for more than 1.75 million adults who include all those registered for general medical services in an extensive geographic area in north east London. This diverse urban geography provided challenging address quality for developing, optimising and evaluating the algorithm. As the ASSIGN algorithm was developed on NHS patient addresses, it has a high potential for health service specific applications and is readily scalable. In addition, ASSIGN is open-source and freely available for others to use. ASSIGN has potential to be used in other address-based datasets including those with information relevant to the wider determinants of health.

The Clinical Effectiveness Group in north east London has pioneered the recording of ethnic background in general practice which is higher than that reported in other geographies or in acute care EHRs [27]. Although an ethnicity code was missing for 15% of our population, we found that the match rate for those with missing ethnicity was similar to that observed in those from White ethnic backgrounds.

While a number of alternative metrics are available to summarise the linkage performance we selected three metrics to be harmonised with GUILD [23] and others such as Office for National Statistics (ONS) [28].

We reported the UPRN match rate based on all match qualifiers combined and not separately for the 2.2% of matches that were ‘child’, ‘parent’ or ‘sibling’ qualifiers which would not be exact matches to the actual patient address. The implications of this will depend on the use of the UPRN: for example, these qualifiers are fit for purpose when using UPRNs for geographical analyses but may be less appropriate for household analyses. We are currently undertaking further work to evaluate approaches to using UPRNs to represent households.

We did not evaluate address-matching success for patients who do not register with a GP practice at all or who are registered at non-residential addresses such as homeless or migrant people.

Interpretation

The ASSIGN algorithm has achieved a very high accurate match rate as evidenced by performance against the two gold standard external datasets with the slightly higher incorrect match rate for the Welsh addresses, reflecting the greater challenge of addresses which contain Welsh language words and spelling.

In the context of the very high match rates achieved, the biases in match success are small but important to identify. The impact of these biases can then be considered when using UPRNs in different populations and for a range of purposes.

Reasons specific to the study population that could influence the five known factors associated with a non-match were considered. Quality of the recorded patient address as well as the address type are important aspects as certain address types are more likely to vary from the address format given in AddressBase Premium, particularly addresses for flats which are more prevalent in urban areas. For example, Tower Hamlets has a higher rate of properties that are flats, which tend to be more poorly recorded addresses. There was also evidence of a slightly lower UPRN match success among those who are more mobile as evidenced by address and practice changes, and duration of registration at the practice. Address-matching was slightly less successful for younger people having taken account of mobility, and the reasons for this are unclear. Of interest were the differences noted by GP EHR supplier systems and further investigation of the address format in SystmOne may be warranted. In summary, those without a successful UPRN match - while small in absolute numbers – demonstrate some demographic, geographic and organisational characteristics indicative of underlying poorer address quality. Some of these factors may be amenable to improvement at the point of address recording in general practices and warrant further exploration.

Specifically, the GP practice is key to the accurate recording of the patient address and to improving address quality in the NHS. We are considering how results from this analysis could be fed back to GP practices to improve systems for patient address recording as well as to confirm accuracy of address with patients since many aspects of direct patient care depend on accurate patient addresses.

The momentum of address-matching and assigning UPRNs to address data created by the UK government's geospatial data strategy has not been matched by greater transparency and evaluation in methods used to assign UPRNs as highlighted by Schinasi et al. (2018) [4]. The Secure Anonymised Information Linkage (SAIL) databank [29] in Wales has a 14 year history of data linkage of national datasets including by address and UPRN with NHS Wales Informatics Service as the Trusted Third Party (TTP) organisation that carries out the linkage. To date address keys (e.g. UPRNs) have been assigned to addresses using Experian QAS [30] with the Postcode Address File [31] and ESRI LocatorHub [32] which, together with other internal methods in the Welsh Address Matching Service, does not have a transparent methodology. The methodology behind the ONS address-matching service [7] is open-source code and is documented and performance evaluated by match rate and by clerically checking the quality of matches compared to other commercial solutions, but there has been no evaluation of bias. The Scottish Improvement Service's Data Hub's address-matching methodology is not documented or evaluated in detail, stating that 'no thorough clerical review of automatic matches' had yet been carried out [8]. The Ordnance Survey's Match and Cleanse service [9] does not currently provide transparent documentation of the method or any quality assurance.

The ASSIGN method is innovative in its transparency of methodology, quality assurance and bias, is open-source, and is scalable. We have now implemented automatic UPRN matching for the patient addresses of 6.9 million London citizens registered with general practitioners who are included in the London Discovery Programme. We are currently exploring wider implementation of ASSIGN in different geographic areas in the UK, as well as across different organisations to support integration of data between health and local authorities including schools and social care settings and to other non-residential property types, particularly care homes.

Conclusion

The ASSIGN address-matching algorithm has been developed for use with NHS recorded patient addresses in an ethnically diverse urban population. It offers a transparent, accurate and quality-assured method for assigning UPRNs and advancing the use of geospatial linkage for effective health care and population health management, for supporting planning and policy for whole systems approaches, and for health data science research.

Acknowledgements

This work was supported by a UKRI Rutherford Postdoctoral fellowship (GH), and by funding from Endeavour Health Charity, Barts Charity (Grant/Award Number: MGU0419), and Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities.

This work used 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 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 ASSIGN algorithm for direct patient care. This access was limited to approved individuals with appropriate information governance training working in a secure trusted data environment.

Only aggregated patient data are reported in this study.

References

1. Schwartz BS, Stewart WF, Godby S, Pollak J, DeWalle J, Larson S, Mercer DG, Glass TA. Body mass index and the built and social environments in children and adolescents using electronic health records. American journal of preventive medicine. 2011 Oct 1;41(4):e17–28. https://doi.org/10.1016/j.amepre.2011.06.038

2. Bazemore AW, Cottrell EK, Gold R, Hughes LS, Phillips RL, Angier H, Burdick TE, Carrozza MA, DeVoe JE. "Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. Journal of the American

Medical Informatics Association. 2016 Mar 1;23(2):407–12. https://doi.org/10.1093/jamia/ocv088

3. Laranjo L, Rodrigues D, Pereira AM, Ribeiro RT, Boavida JM. Use of electronic health records and geographic information systems in public health surveillance of type 2 diabetes: a feasibility study. JMIR public health and surveillance. 2016;2(1):e12. https://doi.org/10.2196/publichealth.4319
4. Schinasi LH, Auchincloss AH, Forrest CB, Roux AV. Using electronic health record data for environmental and place based population health research: a systematic review. Annals of epidemiology. 2018 Jul 1;28(7):493–502. https://doi.org/10.1016/j.annepidem.2018.03.008
5. Geospatial Commission. Unlocking the power of location: The UK's Geospatial Strategy, 2020 to 2025. Geospatial Commission; June 2020. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894755/Geospatial_Strategy.pdf
6. Ordnance Survey. Public Sector Geospatial Agreement (PSGA). Available from https://www.ordnancesurvey.co.uk/business-government/public-sector-geospatial-agreement
7. Office for National Statistics. ONS working paper series no 17 - Using data science for the address matching service. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsworkingpaperseriesno17usingdatasciencefortheaddressmatchingservice#authors
8. Clark D, Dibben C. A guide to CHI-UPRN Residential Linkage (CURL) file. Scottish Centre for Administrative Data Research and Public Health Scotland; November 2020. Available from: https://www.isdscotland.org/Products-and-Services/EDRIS/_docs/CURL-Report-November-2020.pdf
9. Ordnance Survey. Match and Cleanse API. Available from https://developer.ordnancesurvey.co.uk/match-cleanse
10. Ordnance Survey. AddressBase Premium. Available at https://www.ordnancesurvey.co.uk/business-government/products/addressbase-premium
11. Kinnee EJ, Tripathy S, Schinasi L, Shmool JL, Sheffield PE, Holguin F, Clougherty JE. Geocoding error, spatial uncertainty, and implications for exposure assessment and environmental epidemiology. International journal of environmental research and public health. 2020 Jan;17(16):5845. https://doi.org/10.3390/ijerph17165845
12. Forbes H, Morton CE, Bacon S, McDonald HI, Minassian C, Brown JP, Rentsch CT, Mathur R, Schultze A, DeVito NJ, MacKenna B. Association between living with children and outcomes from covid-19: OpenSAFELY cohort study of 12 million adults in England. bmj. 2021 Mar 18;372. https://doi.org/10.1136/bmj.n628
13. Haroon S, Chandan JS, Middleton J, Cheng KK. Covid-19: breaking the chain of household transmission. bmj. 2020 Aug 14;370. https://doi.org/10.1136/bmj.m3181
14. Nash D, Qasmieh S, Robertson M, Rane M, Zimba R, Kulkarni S, Berry A, You W, Mirzayi C, Westmoreland D, Parcesepe A. Household factors and the risk of severe COVID-like illness early in the US pandemic. medRxiv. 2020 Jan 1. https://doi.org/10.1101/2020.12.03.20243683
15. EMIS. Available from https://www.emisgroupplc.com/who-we-are/our-history/
16. SystmOne. Available from https://tpp-uk.com/
17. Vision. Available from https://info.visionhealth.co.uk/gp-solution
18. Regex. Available from https://en.wikipedia.org/wiki/Regular_expression
19. Levenshtein distance. Available from https://en.wikipedia.org/wiki/Levenshtein_distance
20. Tableaux tables. Available from https://en.wikipedia.org/wiki/Method_of_analytic_tableaux
21. ASSIGN open-source code. Available from https://github.com/endeavourhealth-discovery/ASSIGN
22. ASSIGN description. Available from https://wiki.discoverydataservice.org/index.php?title=UPRN_address_matching_algorithm
23. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, Dibben C, Goldstein H. GUILD: GUidance for information about linking data sets. Journal of Public Health. 2018 Mar 1;40(1):191–8 https://doi.org/10.1093/pubmed/fdx037
24. Aligned Assets. British Standard BS7666. Available from https://www.aligned-assets.co.uk/british-standard-bs7666/
25. NHS. NHS data model and dictionary – ethnic category. Available from https://datadictionary.nhs.uk/data_elements/ethnic_category.html?hl=ethnic%2Ccategory
26. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. Journal of biomedical informatics. 2012 Feb 1;45(1):165–72. https://doi.org/10.1016/j.jbi.2011.10.006
27. Hull S, Mathur R, Boomla K, Chowdhury TA, Dreyer G, Alazawi W, Robson J. Research into practice: understanding ethnic differences in healthcare usage and outcomes in general practice. British Journal of General Practice. 2014 Dec 1;64(629):653–5. https://doi.org/10.3399/bjgp14x683053
28. ONS. Developing standard tools for data linkage. February 2021. Available from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/developingstandardtoolsfordatalinkagefebruary2021

29. SAIL databank. Available from https://saildatabank.com/

30. Experian QAS. Available from https://www.experian.co.uk/qas/index.html

31. Experian. What is the PAF? Available from https://www.experian.co.uk/business/glossary/postcode-address-file/

32. ESRI. LocatorHub. Available from https://www.esriuk.com/en-gb/support/esri-uk-products/locatorhub

Abbreviations

ABP:	AddressBase Premium
ASSIGN:	**A**ddre**SS** Match**InG** to Unique Property Reference **N**umbers
CEG:	Clinical Effectiveness Group
CCG:	Clinical Commissioning Group
DDS:	Discovery Data Service
DPA:	Delivery Point Address
EHR:	Electronic Health Record
GP:	General Practitioner
GUILD:	Guidance for Information about Linking Data Sets
IMD:	Index of Multiple Deprivation
LPI:	Local Property Identifier
LSOA:	Lower Super Output Area
NHS:	National Health Service
ONS:	Office for National Statistics
PPV:	Positive Predictive Value
RALF:	Residential Anonymous Linking Field
SAIL:	Secure Anonymised Information Linkage
TTP:	Trusted Third Party
UPRN:	Unique Property Reference number

Supplementary Appendix 1: GUILD report on ASSIGN: The checklist uses data linkage reporting principals from the GUILD Guidance for Information about Linking Data Sets

Data provision

Concept	**Discovery data service (DDS) patient addresses**	**AddressBase Premium**
Population included	Distinct current GP registered patient addresses as at 16th November 2020 from 7 CCG GP practices in north east London for persons aged 18 and over. n = 945,196 distinct addresses Reporting on distinct addresses so that the number of patients with the same address does not skew results.	Records for Greater London area plus 8km buffer Epoch 75. n = 10,595,513 (local authority Land and Property Identifier LPI and Royal Mail Delivery Point Address DPA records)
Linkability: how generated	Addresses provided by patients either online or on a paper form when registering with GPs	Master list of addresses sourced from Ordnance Survey, Royal Mail and local authorities
Linkability: how processed	Entered manually by GP practice administrators	Managed and maintained by GeoPlace[2]
Linkability: how quality controlled	Varies by practice: either no quality control, or check against a street list, or Google searches	GeoPlace stringent data quality processes. Run 359 checks on each record before being accepted into the database. BS7666[3] standard.
Linkability: updates	When informed by patient. Updated addresses are available to Discovery Data Service in real-time	Every 6 weeks
Linkability: cleaning and validation	Address data quality measures calculated. The addresses are reformatted: • into eleven standard address object fields: flat, building, number, dependent thoroughfare, street, dependent locality, locality, town, postcode, organisation, vertical • a second version of the eleven standard address object field is created by correcting spelling errors, de-pluralisation, replacing or removing punctuation and lower casing, and removing extraneous words that are unnecessary in the match process, for example, the range of words that are equivalent to the word 'flat' such as 'apartment' or 'maisonette' • positional checking is carried out e.g. the abbreviation 'st' would be mapped to "street" as a spelling correction, but not if it was presented as the first word in a field "St David's" for example would be retained as "St David". See https://github.com/endeavourhealth-discovery/uprn-match/tree/master/UPRN/yottadb for address preformatting routines.	The addresses are reformatted: • into eleven standard address object fields: flat, building, number, dependent thoroughfare, street, dependent locality, locality, town, postcode, organisation, vertical • the eleven standard address object fields are indexed with single and compound indexes to improve search performance time • the eleven standard address object fields are indexed with performance improving indexes based on semantic equivalence or semantic performance including correcting spelling errors, de-pluralisation, replacing or removing punctuation and lower casing, and removing extraneous words that are unnecessary in the match process, for example, the range of words that are equivalent to the word 'flat' such as 'apartment' or 'maisonette'
Linkability: replaced with artificial identifiers to reduce disclosure before linkage	N/A	N/A

Supplementary Appendix 1: Continued

Data linkage		
Concept	**DDS patient addresses**	**AddressBase premium**
Process: characteristics used for linkage	Address and postcode	Address and postcode
Process: patterns of missingness	There are 945,196 total *distinct* addresses of which 804 (0.09%) have a missing or invalid address or postcode[1]. [1]An incomplete address <8 characters in length; or contains no alphanumeric characters; or contains the words: unknown, no fixed abode, dummy, nfa, not found, not entitled, overseas, not known, not given, overseas, patient, visitor, unk, address, zz99, @, place of birth, none; or begins with: a special character, london, xx, or x; or does not follow full UK postcode format	N/A
Process: expected range of values after cleaning	N/A	N/A
Process: de-duplication	Duplicate address strings relating to different patient-address pairs removed in previous step. Duplicate addresses that are formatted differently were included because they could not easily be identified as relating to the same address until UPRNs are assigned.	N/A Duplicate versions of UPRN in ABP due to different versions of the same address reflecting aliases and the address life cycle
Process: description of algorithm	**Reformat** Candidate and standard addresses are reformatted as per ‘cleaning and validation’ section. **Match** Blocking by matching postcode area, potential matching standard addresses are assessed deterministically by applying matching judgement rules in rank order of extent of string manipulation (rank 1 = no manipulation), using a decision tree to determine which string comparison match tests are passed and which fail until all branches are exhausted and the best match is found. These rules mirror human pattern recognition and are coded using e.g. Levenshtein distance[4], pattern matching (Regex), field swapping and pluralisation. A match is made with one of four overall qualifiers that qualifies the relationship between the candidate address and the matched standard address in relation to approximate geography, or no match is made. The four qualifiers are: • Best match: the closest match out of all available • Child: candidate address is a ‘child’ sub-property of the UPRN it has been matched to • Parent: candidate address is the ‘parent’ building shell of the UPRN it has been matched to • Sibling: candidate address is a near neighbour of the UPRN it has been matched to **Return** Where there is a match, the algorithm returns the UPRN, the overall qualifier, the standard address, the match pattern and match rule identifier employed to get that match. The match rule is a label identifying which section of the code made the match, and the match pattern depicts how five address objects were manipulated to achieve the match. These five address objects are merged from the original eleven: flat, building, number, street, postcode. Twelve possible match terms (see Table 1) exist and can be combined in up to 50 different ways on the five address fields. These are restricted to plausible terms, for example, postcodes are never swapped with streets.	

Continued.

Supplementary Appendix 1: Continued

<table>
<tr><th colspan="3">Data linkage</th></tr>
<tr><th>Concept</th><th>DDS patient addresses</th><th>AddressBase Premium</th></tr>
<tr><td></td><td colspan="2">An example of a match pattern is ‘Pe,Se,Ne,Bp,Fe’. This means that the postcode, street, number, and flat fields were equivalent matches between the candidate and standard address, and the building field was a partial match between the candidate and standard address.
The algorithm is described here: https://wiki.discoverydataservice.org/index.php?title=UPRN_address_matching_algorithm
The algorithm is available for free open-source use here: https://github.com/endeavourhealth-discovery/ASSIGN</td></tr>
<tr><td>Process: new derived linkage variables</td><td colspan="2">N/A</td></tr>
<tr><td>Process: blocking methods</td><td colspan="2">By postcode area</td></tr>
<tr><td>Record-level indicators of the process</td><td colspan="2">UPRN, qualifier, match rule, match pattern</td></tr>
<tr><td>Aggregate linkage results: number of records linked and unlinked</td><td>Of 945,196 distinct address strings:
924,094 matched (<b>98%</b>)
21,102 unmatched (<b>2%</b>)
Of 924,094 matched, broken down by qualifier:
<table>
<tr><th>Qualifier</th><th>Count</th><th>%</th></tr>
<tr><td>Best match</td><td>904,259</td><td>97.85</td></tr>
<tr><td>Child</td><td>9,912</td><td>1.07</td></tr>
<tr><td>Parent</td><td>686</td><td>0.07</td></tr>
<tr><td>Sibling</td><td>9,237</td><td>1.00</td></tr>
<tr><td>Total matched</td><td>924,094</td><td></td></tr>
</table></td><td>N/A</td></tr>
<tr><td>Aggregate linkage results: comparison of characteristics of linked and unlinked records</td><td colspan="2">Address characteristics:
<table>
<tr><th>Characteristic</th><th>Total</th><th>Linked</th><th>Unlinked</th></tr>
<tr><td>Total</td><td>1,549,669</td><td>1,425,497</td><td>124,172</td></tr>
<tr><td>Of which:</td><td></td><td></td><td></td></tr>
<tr><td>E postcode %</td><td>61.2</td><td>61.2</td><td>62.0</td></tr>
<tr><td>N postcode %</td><td>7.3</td><td>7.3</td><td>9.0</td></tr>
<tr><td>R postcode %</td><td>18.7</td><td>19.0</td><td>6.0</td></tr>
<tr><td>I postcode %</td><td>12.3</td><td>12.3</td><td>11.8</td></tr>
<tr><td>Other postcode %</td><td>0.5</td><td>0.3</td><td>8.6</td></tr>
<tr><td>Address begins with numeric character %</td><td>75.9</td><td>76.5</td><td>52.7</td></tr>
<tr><td>Address begins with alphabetic character %</td><td>24.0</td><td>23.5</td><td>46.6</td></tr>
<tr><td>Address begins with special character %</td><td>0.0</td><td>0.0</td><td>0.7</td></tr>
<tr><td>Invalid address or postcode %</td><td>0.1</td><td>0.0</td><td>3.5</td></tr>
</table>
There are higher proportions of ‘Other’ postcodes, addresses beginning with an alphabetic character (i.e. a flat rather than a house) or a special character, and invalid addresses or postcodes in unmatched compared to matched.
Differences between matched and unmatched addresses across all characteristics were found to be significant using chi square tests, but this could be attributable to the large sample size.
Patient and registration characteristics are compared in section ‘Population characteristics’ of the paper.</td></tr>
<tr><td>Aggregate linkage results: representativeness of the linked data set</td><td colspan="2">See paper section ‘Bias in UPRN match success’</td></tr>
<tr><td>Aggregate linkage results: flow diagram of linkage steps</td><td colspan="2">N/A – the linkage steps pathway is different for different addresses depending on the content and required manipulation of the address string</td></tr>
</table>

Continued.

Supplementary Appendix 1: Continued

<table>
<tr><th colspan="3">Data linkage</th></tr>
<tr><th>Concept</th><th>DDS patient addresses</th><th>AddressBase premium</th></tr>
<tr><td>Linkage accuracy: how error rates were estimated</td><td colspan="2">Algorithm applied to two ‘gold-standard’ external reference data sets.
1) 9,177 Welsh local authority addresses.
2) 9,475 Tower Hamlets local authority addresses

True false positive matches, false matches, missed matches, and true negative matches are quantified to calculate:
• Positive Predictive Value (PPV) or Precision - the proportion of record pairs classified by the algorithm as links that are true matches
• Sensitivity or Recall– the proportion of true matches that are correctly classified as links.
• The F-measure – The harmonic mean between positive predictive value and sensitivity. Often used to compare the overall efficiency of a method</td></tr>
<tr><td>Linkage accuracy: estimates of error rates</td><td colspan="2"><table>
<tr><th>Measure
Measure</th><th>DDS address linkage results on
Welsh gold-standard addresses</th><th>DDS address linkage results on
Tower Hamlets gold-standard addresses</th></tr>
<tr><td>Sensitivity</td><td>0.999</td><td>0.999</td></tr>
<tr><td>PPV</td><td>0.996</td><td>0.998</td></tr>
<tr><td>F-measure</td><td>0.997</td><td>0.998</td></tr>
</table></td></tr>
<tr><td>Disclosure controls</td><td colspan="2">Addresses and UPRNs remain in the identifiable zone of Discovery Data Service only.
UPRNs are pseudonymised into Residential Anonymous Linking Fields for third party use</td></tr>
</table>

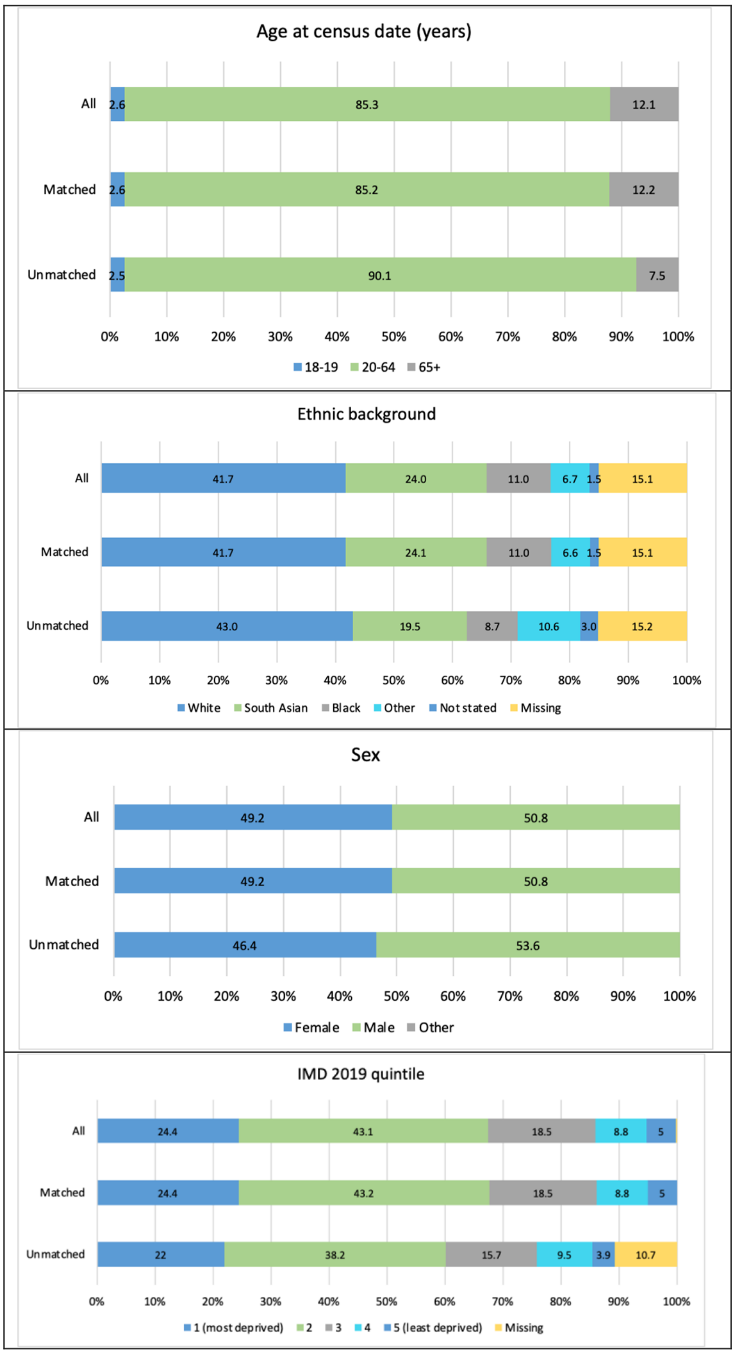
[1] Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.C., Smith, P., Dibben, C. and Goldstein, H., 2017. GUILD: GUidance for Information about Linking Data sets. *Journal of Public Health*, 2017 Mar 28:1–8.

[2] www.geoplace.co.uk

[3] https://www.aligned-assets.co.uk/british-standard-bs7666/

[4] https://en.wikipedia.org/wiki/Levenshtein_distance

Supplementary Appendix 2: Summary characteristics of the study population according to whether patient address was matched or not matched to a UPRN by the ASSIGN algorithm



UPRN = Unique Property Reference Number.
IMD = Index of Multiple Deprivation.

Supplementary Appendix 3: UPRN match rates and absolute differences in proportion matched with respect to reference category for all explanatory variables N = 1,757,018

	Number ***n***	**Address-matched to UPRN (%)**	**Absolute difference relative to reference group (%)**
Age at census date 16/11/2020 (years)			
Missing	8,116	99.62	0.06
>1	**50,740**	**99.56**	**Ref**
1–14	133,371	99.33	−0.22
15–29	570,251	98.06	−1.49
30–64	929,452	98.71	−0.85
65–84	59,973	98.77	−0.85
85 and over	**5,115**	**96.72**	**−2.84**
Ethnic background			
Missing	265,524	98.56	-0.08
British	**382,170**	**98.64**	**Ref**
African	100,743	98.68	0.03
Any other Asian background	61,521	98.38	−0.27
Any other Black background	44,131	99.01	0.37
Any other White background	337,905	98.4	−0.24
Any other ethnic group	52,823	98.42	−0.22
Any other mixed background	15,018	97.88	−0.77
Bangladeshi	145,920	99.28	0.64
Caribbean	48,203	99.16	0.51
Chinese	**21,961**	**95.51**	**−3.14**
Indian	121,134	98.51	−0.13
Irish	13,113	98.41	−0.24
Not stated	**26,196**	**97.09**	**−1.56**
Pakistani	93,538	98.9	0.25
White and Asian	4,947	98.08	−0.56
White and Black African	9,971	97.9	−0.74
White and Black Caribbean	12,200	98.21	−0.43
Sex			
Female	**864,337**	**98.65**	**Ref**
Male	892,638	98.49	−0.16
Other	43	95.35	−3.3
IMD 2019 quintile			
Missing	**3,502**	**23.5**	**−75.21**
1 (most deprived)	**428,373**	**98.71**	**Ref**
2	757,212	98.74	0.02
3	325,075	98.79	0.08
4	154,523	98.45	-0.26
5 (least deprived)	88,333	98.88	0.17
GP registration duration (quartiles)			
Missing	**8,116**	**99.58**	**1.94**
1 (shortest)	**437,228**	**97.64**	**Ref**
2	437,422	98.36	0.72
3	**437,603**	**98.92**	**1.28**
4 (longest)	**436,649**	**99.36**	**1.72**

Continued.

Supplementary Appendix 3: Continued

	Number ***n***	**Address-matched to UPRN (%)**	**Absolute difference relative to reference group (%)**
Number of GP registrations in preceding 12 months			
1	**1,595,729**	**98.58**	**Ref**
2	144,755	98.61	0.03
3 or more	16,534	97.67	−0.91
Number of address changes in preceding 12 months			
1	**1,316,956**	**98.98**	**Ref**
2	**343,808**	**97.89**	−**1.09**
3 or more	**96,254**	**95.41**	−**3.57**
GP system			
Missing	4,960	99.62	0.83
EMIS	**1,629,199**	**98.79**	**Ref**
SystmOne	**87,783**	**94.39**	−**4.4**
Vision	35,076	98.86	0.08
Clinical Commissioning Group			
Newham	**326,386**	**99.16**	**Ref**
Barking & Dagenham	168,008	98.59	−0.57
City & Hackney	259,973	98.25	−0.91
Havering	221,328	99.38	0.22
Redbridge	251,128	98.61	−0.55
Tower Hamlets	**278,520**	**97.7**	−**1.46**
Waltham Forest	251,675	98.35	−0.81

Quartile definitions for GP registration duration: Quartile 1 (shortest): 0–32 months; Quartile 2: 33–77 months; Quartile 3: 78–183 months; Quartile 4 (longest) > 184 months.

EMIS: Egton Medical Information Systems.

Reference groups and values with an absolute match rate difference to the reference group of >1% are in bold.