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

Data linkage

Concept	DDS patient addresses	AddressBase Premium
	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	
Process: new derived linkage variables	N/A	
Process: blocking methods	By postcode area	
Record-level indicators of the process	UPRN, qualifier, match rule, match pattern	
Aggregate linkage results: number of records linked and unlinked	Of 945,196 distinct address strings: 924,094 matched (**98%**) 21,102 unmatched (**2%**) Of 924,094 matched, broken down by qualifier:	N/A

Qualifier	Count	%
Best match	904,259	97.85
Child	9,912	1.07
Parent	686	0.07
Sibling	9,237	1.00
Total matched	924,094	

Aggregate linkage results: comparison of characteristics of linked and unlinked records

Address characteristics:

Characteristic	Total	Linked	Unlinked
Total	1,549,669	1,425,497	124,172
Of which:			
E postcode %	61.2	61.2	62.0
N postcode %	7.3	7.3	9.0
R postcode %	18.7	19.0	6.0
I postcode %	12.3	12.3	11.8
Other postcode %	0.5	0.3	8.6
Address begins with numeric character %	75.9	76.5	52.7
Address begins with alphabetic character %	24.0	23.5	46.6
Address begins with special character %	0.0	0.0	0.7
Invalid address or postcode %	0.1	0.0	3.5

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.

Aggregate linkage results: representativeness of the linked data set	See paper section 'Bias in UPRN match success'	
Aggregate linkage results: flow diagram of linkage steps	N/A – the linkage steps pathway is different for different addresses depending on the content and required manipulation of the address string	

Continued.

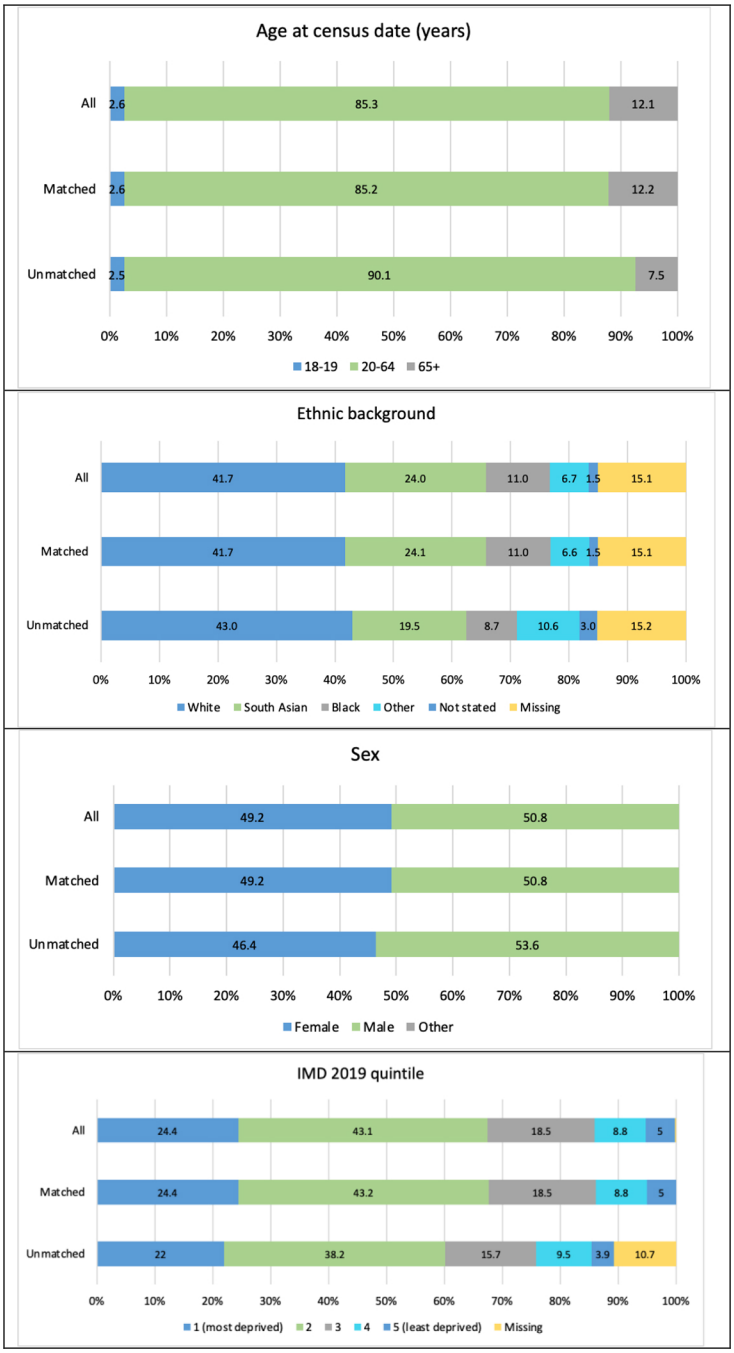
Supplementary Appendix 1: Continued

Data linkage

Concept	**DDS patient addresses**	**AddressBase premium**
Linkage accuracy: how error rates were estimated	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	
Linkage accuracy: estimates of error rates	Measure Measure: Sensitivity PPV F-measure	DDS address linkage results on Welsh gold-standard addresses 0.999 0.996 0.997
		DDS address linkage results on Tower Hamlets gold-standard addresses 0.999 0.998 0.998
Disclosure controls	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	

[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.