International Journal of Population Data Science

Journal Website: www.ijpds.org

Strategies for ascertaining timing of address changes in administrative data

Joseph Lam[1*], Mario Cortina-Borja[1], Peter Christen[2,3], Richard Thomas[4], Robert Aldridge[5], Ruth Blackburn[1], Ruth Gilbert[1], Andy Boyd[4], and Katie Harron[1]

Submission History

Submitted:	10/04/2025
Accepted:	30/09/2025
Published:	10/12/2025

[1] Great Ormond Street Institute of Child Health, University College London, UK
[2] University of Edinburgh, UK
[3] Australian National University, AU
[4] United Kingdom Longitudinal Linkage Collaboration, Bristol Medical School, University of Bristol, UK
[5] Institute for Health Metrics and Evaluation, University of Washington, USA

Abstract

Addresses captured in administrative health data can facilitate the analysis of place-based exposures and their influence on health. If patients change address, hospital records are updated when individuals next use health services: precise timing of addresses changes must therefore be inferred. There is therefore a need to determine the temporal accuracy of address recording in these data.

Using deidentified data, we compared Lower Super Output Areas (LSOAs) derived from English hospital records (NHS_E, $n = 40{,}102$), with linked data on addresses recorded in the UK Longitudinal Linkage Collaboration cohorts (Cohorts, $n = 40{,}963$) from January 1989 to April 2023. We compared the accuracy of three methods for estimating the timing of changes in LSOA: 1) inferring the end date of the current address as the start date of the next reported address minus 1 day (N-1 method); 2) using median date between current and next start date as the end date for current address, and to update the start date for the next address (Median method); 3) generating the address end date as a function of beta distribution defined in terms of the current and next start dates, assuming that most people update their address soon after they move (random method).

In total, 39,216 (95.7%) Cohort members had at least 1 matching LSOA reported in both Cohorts and NHS_E data. 47% of these matching LSOAs were recorded in the hospital data within the two years prior to or following the Cohorts recorded dates. All three methods demonstrated LSOA agreement of ∼78%, with negligible differences across methods.

Methods for estimating timing of changes in hospital-recorded LSOA, based on address, are reasonably similar when compared to LSOA of residence in Cohorts data despite different updating methods. Researchers should consider the assumptions and implications of each method and justify their approach. More accurate and standardised methods are needed for recording address across systems.

Keywords
UK LLC; geospatial; addresses; bias

*Corresponding Author:
Email Address: joseph.lam.18@ucl.ac.uk (Joseph Lam)

https://doi.org/10.23889/ijpds.v10i1.2995
December 2025 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Introduction

In epidemiologic and health services research, residential address data are increasingly being used to assess place-based exposures – including household, environmental, social, and structural factors – and their influence on health outcomes [1]. For instance, use of solid fuels for cooking has been associated with premature mortality [2]; residential proximity to major roads has been associated with increased asthma prevalence in children [3]; living in socioeconomically deprived neighbourhoods has been linked to increased risk for depression, anxiety, and chronic illnesses like diabetes and hypertension [4, 5].

Establishing a spatial-temporally accurate residential history from administrative data sources like hospital records is challenging. Studies frequently rely on a single address recorded at a specific time point (for example, the recorded address at hospital admission or at birth) as a proxy for an individual's residence over an extended exposure period [6, 7]. Such an approach can be problematic because residential mobility creates spatial-temporal variation that undermines the precision of exposure measurements and may yield biased estimates.

Increasingly, researchers recognise the importance of temporal and spatial accuracy in place-based exposures captured in administrative health data. However, precise information on the timing of address changes is often unavailable and must therefore be inferred. Temporally precise representation of address changes is important for accurately representing (cumulative) exposure to environmental risk factors, residential mobility and the use of area-level deprivation measures such as the Index of Multiple Deprivation in the UK [8, 9]. Accurately representing residential mobility is also important for evaluating public health interventions. This is particularly important for populations with a higher residential mobility than the general public, such as those with unstable housing, those who are pregnant, and those with young children [10].

Within the UK there is no law compelling people to notify the state when they move address (except to register the change with the authority in charge of vehicle drivers' licences). This means people often notify service providers at their next interaction. When there are gaps or overlaps between recorded dates of addresses across or within different datasets, or when address end dates are not collected, researchers have to make assumptions on which dataset to prioritise, and when the address change happened. This is relevant for health administrative data, as addresses can potentially be recorded at each contact with health services (such as general practitioners, hospitalisation etc.), rather than being updated when people move to a new address.

This study describes and compares the accuracy of three methods for estimating the timing of changes in Lower Super Output Areas (LSOA; geographical output areas with between 400 and 1,200 households) in deidentified hospital records against self-reported deidentified address data from cohort studies.

Methods

We compared three methods based on different assumptions about the temporal accuracy of healthcare-recorded address data (Figure 1).

Method 1: N–1

The N-1 method assumes that the recorded start date corresponds to the actual date when an individual moved. Similar to Fecht et al.'s algorithm [8], which favours start dates over end dates, this method assumes that end dates are usually imputed in relation to start dates in database design. In our study, for N-1 method, when there are gaps or overlaps in the address period, the end date of the earlier address period is marked as one day before the next start date.

Method 2: Median

The Median method assumes that the reported start date of the new address is likely to be later than the actual date, as patients would only update their address at their next clinical or research encounter. This assumption could be valid in NHS_E. A study of six general practices in London found that nearly 40% of patients registered address change more than six months after they move, 13% delayed for over one year, and 7% delayed for over three years to update their address [11]. The size of this gap is likely to vary by the individual's health status, their family characteristics (e.g., if they have care responsibilities), the source of data (primary or secondary care), length of stay in each address, distance they moved, and ease of update of residential information (e.g. from 2022, National Healthcare Service allow individuals to update their emails and contact numbers online, but not their postal addresses). The Median method uses the midpoint between current and next address start date as the end date for the current address and subsequently, it updates the start date for the next address.

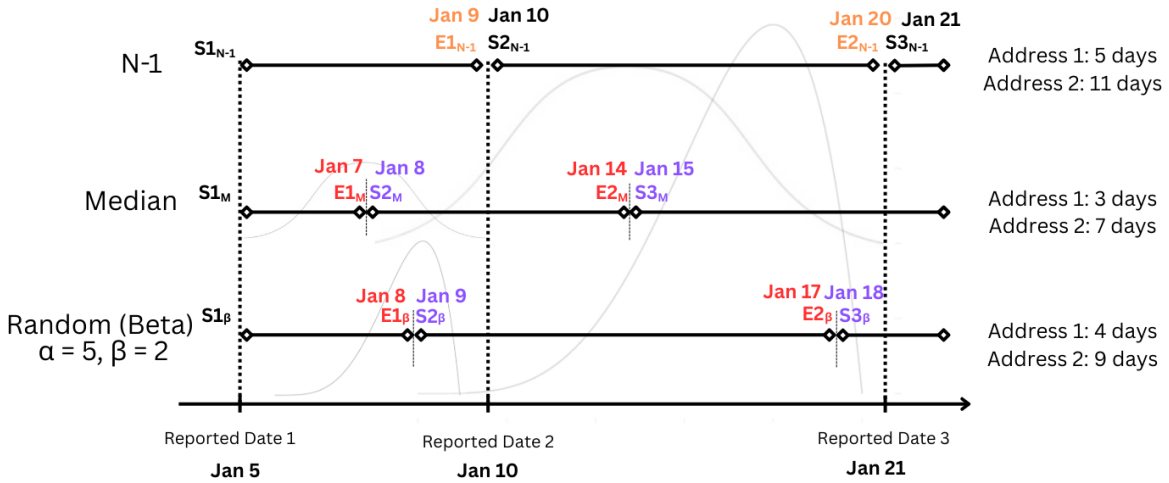
Method 3: Random (Beta)

The Random method also assumes that the recorded date is later than the actual date. However, instead of using the midpoint point as the end date for the current address, this method supposes that the percentage differences between the recorded and actual end dates are proportional to realisations from a Beta distribution with parameters $\alpha = 5$ and $\beta = 2$ [12]. This distribution was chosen for its left-skewed shape (long-tail of the distribution is on the left), which reflects the assumption that most of the actual end dates will be closer to the recorded end date than they were if their differences were assumed to be uniformly distributed. This assumes that most people update their addresses in NHS_E soon after they move.

Dataset

This study included 40,963 participants of the longitudinal cohorts which are partners in the UK Longitudinal Linkage

Figure 1: Three methods of ascertaining periods of time for each address. The same date ascertainment method is applied to both Cohorts and NHS_E address data. These dates are included only to illustrate how each method works



Highlighted in orange are end dates determined using N-1 method. Highlighted in red are the end dates, and in purple the subsequently adjusted start dates using median and random method. Median and Beta distribution is illustrated in the background to visualise how each method is applied.

Collaboration (UK LLC) [13]. The pooled studies within UK LLC form a dynamic sample which changes over time, we used denominator file version "Freeze 1" (taken 2022-10-19). A summary of each cohort included in this paper is described in Supplementary Table 1 with study sample characteristics and linkage rates provided online [14]. We included participants whose data had been linked to UK health administrative data, who had provided permission for geocoding, and who had provided at least 1 address with a valid start date in both cohort and health administrative data (as defined by inclusion in the UK LLC "geo indicators" dataset; version 2022-09-13). Within UK LLC, there is capability to look at the encrypted LSOA derived from participant address information collected directly by seven longitudinal population cohorts (referred to from here as Cohorts) as well as encrypted LSOA information from participants linked NHS England health administrative data. The NHS LSOA is sourced from both primary and secondary health services during patient interactions (referred to as NHS_E. © (2025), NHS England. Data re-used with the permission of NHS England. All rights reserved.). This dataset includes participants from Avon Longitudinal Study of Parents and Children (ALSPAC), Born in Bradford (BiB), Disorder Genetics Initiative and COVID-19 Psychiatry and Neurological Genetics Study (COPING), Extended Cohort for E-Health, Environment and DNA (EXCEED), Generation Scotland, Medical Research Council National Survey of Health and Development (NSHD), and TwinsUK [15–20]. The UK LLC system ensures the encryption of LSOA from linked and study data is comparable and also preserves the privacy of participants as neither the researcher nor UK LLC staff have access to the encryption key (which is managed by a trusted third party). This capability is rare and allows us to compare the impact of methods of date ascertainment in linked data.

Each of the cohorts collect information about changes in addresses differently, and where approaches differ across time within cohorts. For example, NSHD checked and traced the current address of their sample members with an annual birthday card [21]; ALSPAC has a curated address history database for each individual from multiple waves of study, asking them to provide month and year of residential moves [22]. Address changes between waves could be self-reported on the study online portal; COPING is conducted entirely online, and only recorded baseline addresses [18].

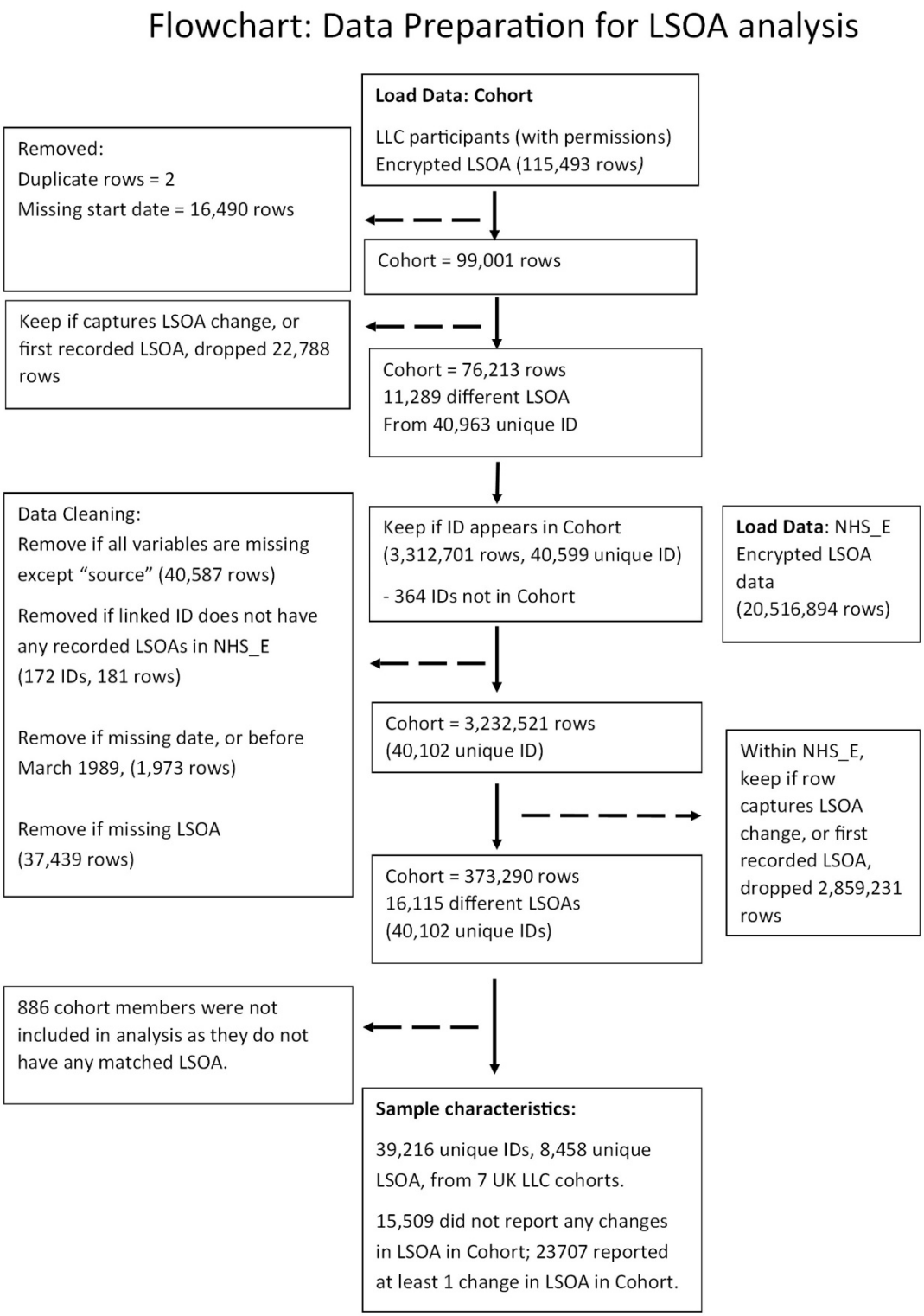
We included Cohorts-reported addresses from 1990-10-11 to 2022-10-13, and NHS_E-reported addresses from 1989-01-27 to 2023-04-06 (censor date for linked data). We sorted the LSOAs by their start date and only retained the rows where LSOA changes were captured (apart from the first recorded LSOA; Figure 2). For the 39,216 participants with at least one matching LSOA across data sources, we derived end dates for each address using each proposed method.

Statistical analysis

First, we explored whether there were differences in age category (≤18, 19-30, 31-59, 60-74, ≥75 years), sex or gender (male, female), or ethnicity (White, South-east Asian, Other Asian, Black, Mixed, Other, Missing or NA) between those who reported more than one LSOA, those who had a single LSOA (in either NHS_E or Cohorts), and those who did not appear in NHS_E or provided no LSOAs in NHS_E. Since sociodemographics were collected using different categories (such as sex and gender, or ethnicity groups) across different settings (self-reported, parent-reported, recorded by healthcare professionals), we first harmonised characteristics reported across Cohorts using the most disaggregated ethnic categories. We prioritised self-reported characteristics from Cohorts and filled missing information with NHS_E reported characteristics.

This description was restricted to the 39,216 individuals with at least one reported LSOA in both Cohorts and NHS_E. We also described the frequency of LSOA changes in both Cohorts and NHS_E.

Figure 2: Flow chart of data processing pipeline for this study



LLC = UK Longitudinal Linkage Collaboration, NHS_E = National Health Services Dataset, LSOA = Lower Super Output Areas.

We then assessed the agreement between reported timing of address changes in Cohorts and NHS_E, using a cross-sectional comparison. We examined how many LSOAs reported in Cohorts were also reported in NHS_E; and for those where the same LSOAs were reported, the minimum start date difference for each individual across all of their matching LSOA (Figure 3).

Longitudinal Comparison: Proportion of time in Cohorts represented by NHS_E

Finally, we estimated the proportion of time that the LSOA recorded in Cohorts was represented by the LSOA in NHS_E, over the total period for each individual (longitudinal comparison, Figure 4). We refer to this metric as "coverage".

Figure 3: Illustration of cross-sectional comparison

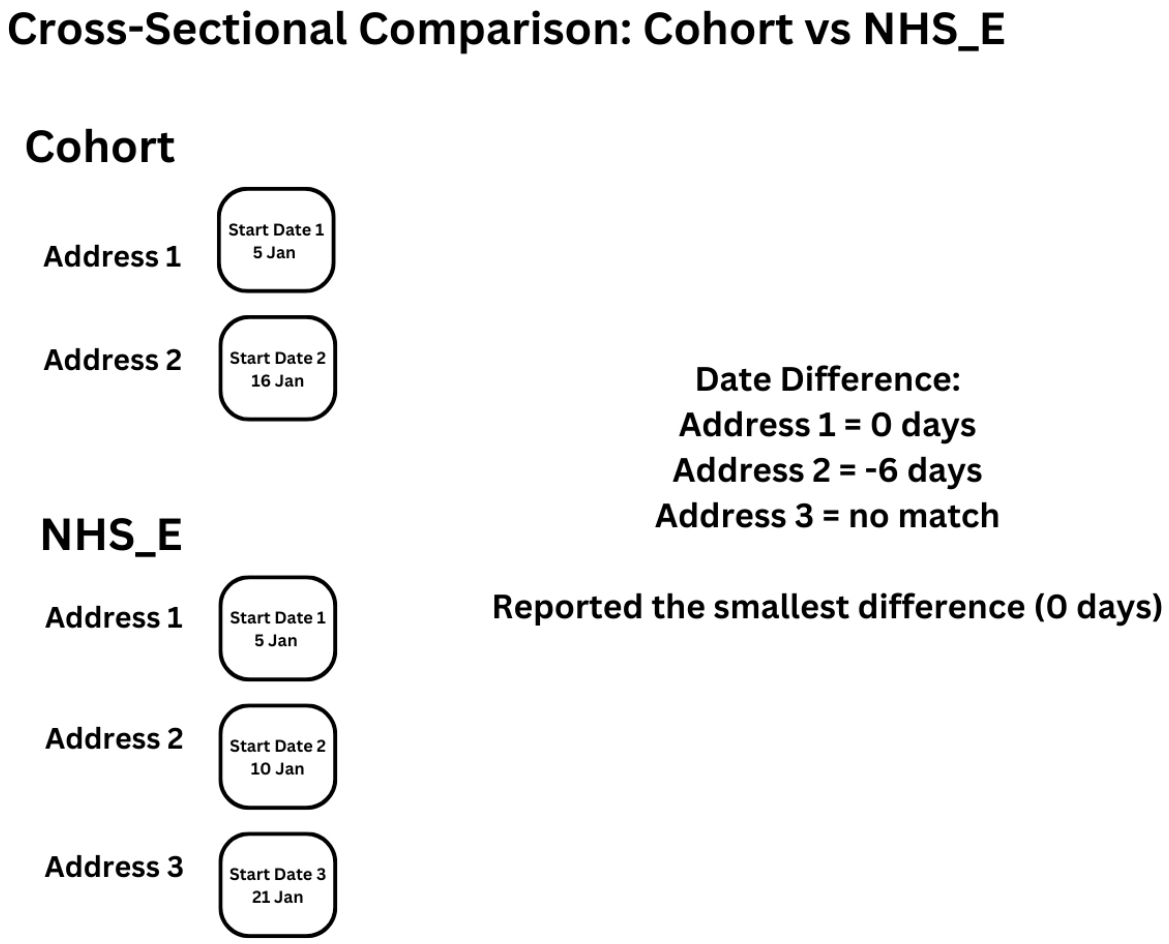
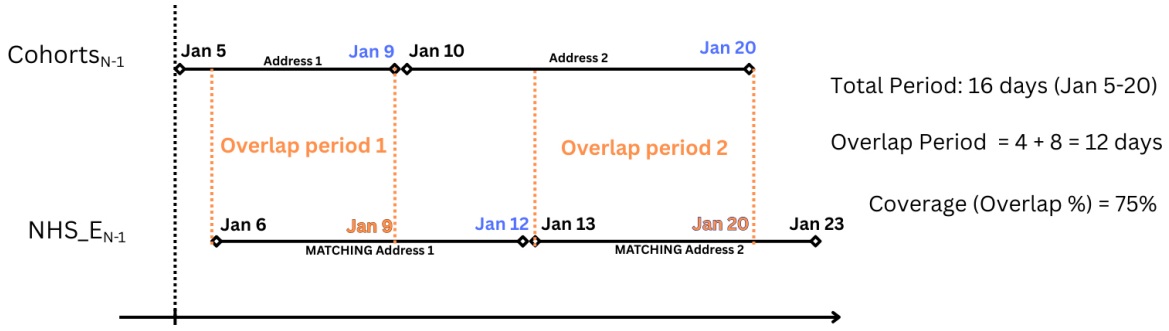


Figure 4: Illustration of longitudinal Comparison, proportion of time of addresses in Cohorts in agreement with NHS_E, using N-1 method



Let:

$$T^{Coh} = [t^{Coh}_{min}, t^{Coh}_{max}]$$

Be the observation period in Cohorts for a given individual.

$$T^{NHS}_j = [a_j,\ b_j],\ j = 1, \ldots,\ J$$

denote the intervals in NHS_E where the recorded LSOA matches the Cohorts LSOA.

The total Cohorts period (measured in days) is

$$L^{Coh} = t^{Coh}_{max} - t^{Coh}_{min} + 1$$

the overlap length between the Cohorts interval and NHS interval T^{NHS}_j is

$$L^{overlap}_j = \max\left(0, \min\left\{t^{Coh}_{max},\ b_j\right\} - \max\left\{t^{Coh}_{min},\ a_j\right\} + 1\right).$$

Then the total overlap length is

$$L^{overlap} = \sum_{j=1}^{J} L^{overlap}_j$$

and coverage is defined as:

$$Coverage = \frac{L^{overlap}}{L^{Coh}}$$

We repeated this analysis using each date ascertainment method (N-1, Median and Beta). Higher coverage implies higher agreement in both reported address and the timing between two data sources, with a maximum of 100%.

Sensitivity analysis compared Cohorts members with only one reported LSOA in Cohorts from those with more than one reported LSOA. We did not consider the between-study variations in address recording mechanisms in Cohorts.

In addition to the three main rules for assigning address change dates, we developed a supplementary approach in which the timing of the change is modelled as a latent variable with an empirically derived distribution. Using the subset of records where Cohorts and NHSD address start dates could be directly compared, we calculated the relative position of the true change date within the interval between consecutive address records. These relative positions were modelled using a Beta distribution whose shape parameters α and β were estimated directly from the data. This fitted Beta distribution represents the latent change-time model, describing the probability that a change occurs at any given fraction of the observed interval.

We applied a multiple imputation strategy to propagate this model into the full dataset. For each individual and address spell (where the end date was not directly observed), we drew 20 independent imputed end dates from the fitted Beta(α, β) distribution, scaled to the length of the interval until the next recorded address change, and constrained to precede the next start date. Each draw produced a complete set of address histories, resulting in M imputed datasets. The overlap analysis was repeated on each imputed dataset, and results were combined using Rubin's rules to produce pooled estimates and associated 95% confidence intervals. This approach integrates a data-driven latent timing model with multiple imputation to allow both improved realism in the simulated change dates and formal quantification of the variation it introduces into the evaluation metrics.

Results

Of the 40,963 Cohorts members, 39,216 (95.7%) had at least 1 matching LSOA in NHS_E, 886 (2.2%) had at least 1 valid LSOA in NHS_E, but did not match with any Cohorts LSOA, and 536 (2.1%) did not link to NHS_E, or did not provide any valid LSOA in NHS_E (Table 1).

For the 39,216 individuals with at least one reported LSOA in both Cohorts and NHS_E, there were no major differences by age category, sex, or ethnicity between those who reported more than one LSOA, those who had a single LSOA (in either data source), and those who were not included (Table 1).

For the 39,216 Cohorts members with a matching LSOA in NHS_E, almost half of start dates fell within two years before or after the Cohorts-reported start date (Figure 5). Another 18.1% of start dates were reported at least five years before Cohorts-reported start dates, and $<$3% of start dates were $>$2 years after the Cohorts-reported start date (Supplementary Table 3). Observed patterns suggest that variations in reporting were not specific to particular age groups; and were not driven by a particular cohort.

Longitudinal Comparison: Proportion of time in Cohorts represented by NHS_E

For the 39,216 participants in the longitudinal analysis, the overall mean coverage was 78.6% (95% Confidence Interval, CI 78.3% -78.9%) using the N-1 method, meaning that the LSOA recorded in Cohorts on average agreed and represented the LSOA in NHS_E for 78.6% of the time period. The Median and Random method worked similarly, with a coverage of 79.3% (95% CI 79.0%-79.6%) for Median, and 78.8% (95% CI 78.5%–79.1%) for Random method (Figure 6).

Coverage was lower ($\sim$73%) for individuals with more than one recorded LSOA change in Cohorts across all three methods (n = 23,707; Supplementary Figure 1), and coverage was higher ($\sim$87%) for individuals with only one recorded LSOA in Cohorts across all three methods (n = 15,509, Supplementary Figure 2, Supplementary Table 4).

Supplementary Analysis: Latent change-time model with multiple imputation

Applying the empirically estimated latent change-time model to the full dataset via multiple imputation produced results closely aligned with the main analysis. The fitted Beta distribution (α, β = 4.26, 2.96) to the Cohort-NHS_E date discrepancies indicated a tendency for address to occur later within the observed interval, as expected in our random (beta) model. Using this distribution to generate 20 imputed datasets, the imputation-pooled mean coverage was 81.9% (95% Confidence Interval = 81.7%–82.2%, Supplementary Figure 3). The distribution of person-level coverage proportions across imputations was stable, and the pooled means were within 3% of the fixed Beta model results. This consistency suggests that the empirically derived latent timing model does not materially alter the overall coverage estimates, but provides a more realistic and data-driven representation of change-time uncertainty.

Discussion

In this study, despite the difference in context and mechanism of data collection between data sources, we found acceptable spatial-temporal coverage of recorded LSOAs. Over 95% of Cohorts members had at least 1 matching LSOA reported in NHS_E. Overall, addresses in NHS_E covered 78% of individuals' Cohorts-recorded addresses both spatially and temporarily using all three methods. There is limited evidence of sizeable differences in coverage between the three methods.

Within Cohorts data, individuals report residential addresses at varying time intervals for each active study; within NHS_E, address changes were recorded prospectively at each contact with health services. There is a higher burden on individuals in Cohorts to report all previous addresses, which may introduce recall errors. Study participants may only report their current address at the time of data collection and are not always allowed to report multiple current residencies. On the other hand, late reporting of LSOA in NHS_E likely corresponds to the delayed reporting others have described [11]. Addresses recorded in NHS_E could contain more noise as individuals could be reporting different addresses when they access different health service providers; where corrections to address data appear as changes in address; or where patients are admitted in a condition where they cannot accurately or willingly provide their address. Since we only used LSOA instead of exact addresses, LSOAs would be considered a

Table 1: Demographics based on harmonised Cohorts and NHS_E data, for the analytical sample, those with no LSOA change, records without any LSOA match in Cohorts or NHS_E, or not provided any LSOA in NHS_E

		Analytical Sample	**No LSOA Change (Cohorts)**	**No LSOA Change (NHS_E)**	**No LSOA match**	**Not Included#**
	Total (*n*)	39,216	15,509	15,396	886	536
Variable	Category					
Age group	≤18 years	5030 (12.8)	1941 (12.5)	1941 (12.6)	<5%	<5%
	19-30 years	1417 (3.6)	546 (3.5)	557 (3.6)	>5%	>5%
	31-59 years	16950 (43.2)	6743 (43.5)	6592 (42.8)	>45%	>45%
	60-74 years	9434 (24.1)	3767 (24.3)	3692 (24.0)	>20%	>20%
	≥75 years	5144 (13.1)	2012 (13.0)	2105 (13.7)	>10%	>10%
	Missing	1241 (3.2)	500 (3.2)	509 (3.3)	<5%	<5%
Gender or sex∧	Female	24471 (62.4)	9670 (62.4)	9596 (62.3)	>60%	>60%
	Male	13785 (35.2)	5457 (35.2)	5408 (35.1)	>35%	>30%
	Missing or NA	960 (2.4)	382 (2.5)	392 (2.5)	<5%	<5%
Ethnicity	White	21801 (55.6)	8761 (56.5)	8682 (56.4)	>70%	>75%
	South-east Asian	5554 (14.2)	2148 (13.9)	2121 (13.8)	>5%	>15%
	Other Asian	383 (1.0)	166 (1.1)	155 (1.0)	<5%	<5%
	Black	372 (0.9)	149 (1.0)	137 (0.9)	<5%	<5%
	Mixed	458 (1.2)	202 (1.3)	175 (1.1)	<5%	<5%
	Other	276 (0.7)	104 (0.7)	122 (0.8)	<5%	<5%
	Missing or NA	10372 (26.4)	3979 (25.7)	4004 (26.0)	>20%	>25%
Number of Unique LSOAs		8,458	4,001	4,315	NA	NA

* <5% may include 0. Precise numbers for No LSOA match and ID not in NHS_E are suppressed due to risk of disclosure, therefore only percentages are given. ∧sex and gender terms were harmonised to options in "sex" variables. # Not Included when ID did not appear NHS_E or did not provide any LSOA in NHS_E.Around 40% of individuals in both data sources reported a single LSOA during the study period (Table 2). However, only 15.5% reported a single LSOA in both data sources (Supplementary Table 2). A greater proportion of individuals reported higher numbers of LSOA changes (4+) in NHS_E versus Cohorts data over a comparable time-period.

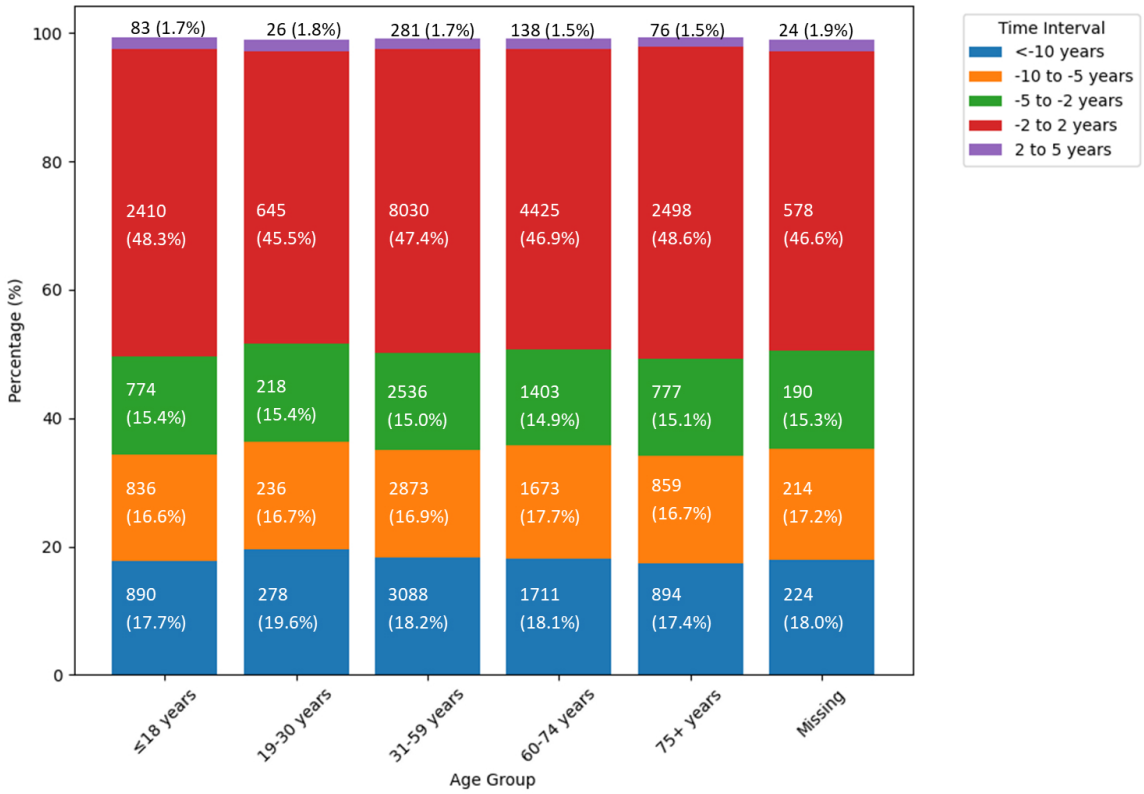
Table 2: Number of LSOA changes in Cohorts and NHS_E. Note that this table includes records from individuals that did not have any matching LSOA in NHS_E, who are excluded from subsequent analysis

	Number of LSOA Changes			
	Cohorts		**NHS_E**	
	n	**%**	***n***	**%**
0 (No changes)	16878	41.2%	15737	39.2%
1	9987	24.4%	2833	7.1%
2	5984	14.6%	3388	8.4%
3	3400	8.3%	1507	3.8%
4	1915	4.7%	2332	5.8%
5 to 9	2582	6.3%	5467	13.6%
10 to 14	183	0.4%	3032	7.6%
above 15	34	0.1%	5806	14.5%
Total	40,963		40,102	

match even if the exact addresses are different. Addresses in the same area recorded in NHS_E but not in Cohorts could explain why some NHS_E start dates fell before Cohorts in our cross-sectional comparison.

Choice of date ascertainment methods should take into account the characteristics of the underlying data. For example, N-1 method works best when records have minimal errors. However, if the recorded dates are wrong, missing or out-of-order, N-1 methods can introduce significant inaccuracies; the Median method works well for capturing short stays and frequent address changes, such as in populations with high mobility, but can underestimate long-duration stays if moves are infrequent. The Random (Beta) method is a flexible way to adjust systematic difference

Figure 5: Minimum difference (in years) between LSOA start dates recorded in Cohorts and NHS_E, for Cohorts members with a matching LSOA in NHS_E, by age group



NHS_E-reported start dates 5-10 years or >10 years after Cohorts-reported start dates were suppressed due to small numbers. Negative values mean that the NHS_E-reported start date fell before the Cohorts-reported start date; positive values mean that NHS_E-reported start dates fell after the Cohorts-reported start date.

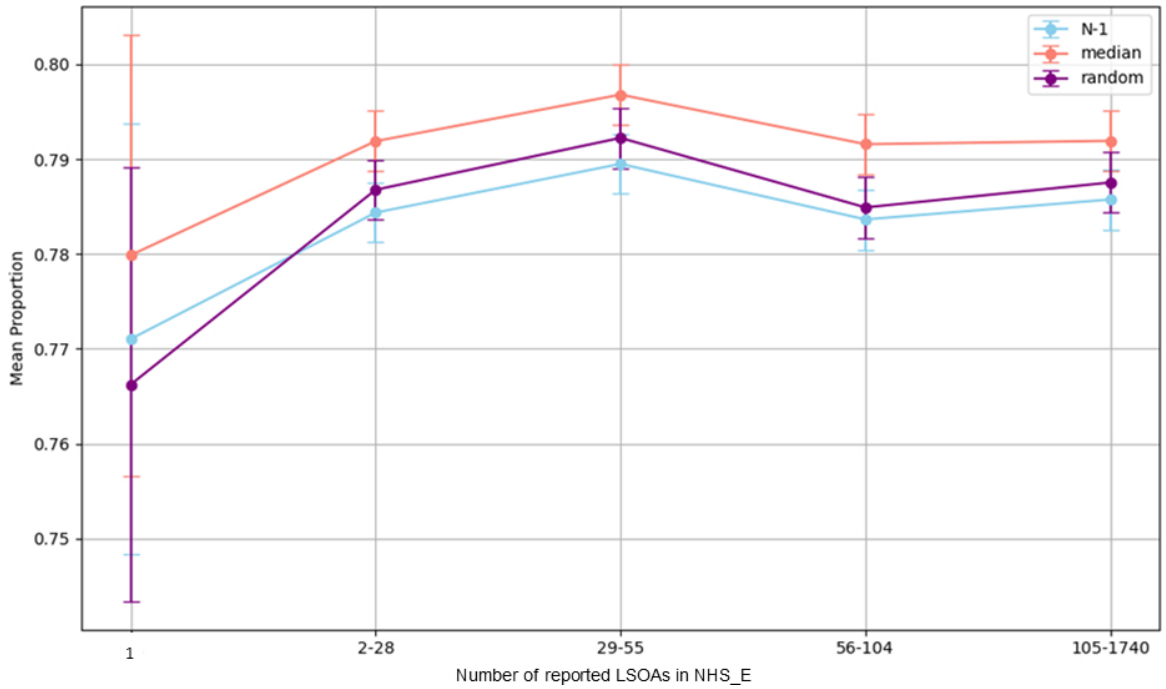
between recorded dates from actual dates (such as in health records), but introduces some model-based uncertainties. Researchers should consider how appropriate these different assumptions are for their study.

Researchers should also consider key contextual factors such as cohort characteristics and environmental exposure of interest when choosing date ascertainment methods. In terms of cohort characteristics, for example, households expecting newborns or with children are more likely to move than the general population [9, 23]. People experiencing poverty may have shorter residency periods at each address, and be less likely to move to less deprived neighbourhoods [24–26]. In our study, participants who moved tended to remain in equally or more deprived areas compared to less deprived areas, especially for those starting in more deprived areas (Supplementary Figure 4, 5). By nature, errors and biases in date ascertainment affect those who with higher residential mobility disproportionately. However, the size of bias varies by environmental exposure of interest. For example, Hoek et al. [27] suggested that for long-term exposures such as outdoor air pollution, biases arising from not accounting for neighbourhood-level exposures (e.g. journeys to school or work) are typically quite small where the population catchment area is also small. This was also described by Chen et al. [28] for a small cohort of pregnant mothers, where most moves in their cohort were within the same neighbourhood, or between neighbourhoods of similar characteristics. In these cases, improvements from using a more temporally and spatially accurate address may be small.

There are a few limitations to our current analysis. Firstly, our data only included individuals who were successfully linked to NHS_E and had at least one matching LSOA. The linkage used the latest available address provided in the NHS patient demographics service as one of the linkage keys. Individuals were more likely to be linked and included in the study if they had provided an up-to-date address. Wider evaluation of NHS_E data in the general population will likely find a lower proportion of matched LSOA than in our study. By demonstrating high temporal agreement between our data sources, our study showed an opportunity to utilise temporal information of addresses, instead of just using the most recently reported address for data linkage. Our findings provided a reference distribution of dis/agreement decays for addresses to help attribute appropriate weights for linkage [28].

Secondly, we evaluated our cohort using encrypted LSOA instead of individual addresses. This means that address changes within the same LSOA were not captured or described in our study. The current evaluation of address matches is therefore likely an under-estimation of true residential mobility; and an over-estimation of the extent of coverage of NHS_E LSOA. Using markers of higher geospatial resolution, such as encrypted Unique Property Reference Numbers, would allow us to investigate patterns of residential moves and exposure assignment with higher precision.

Figure 6: Coverage of NHS_E-reported LSOA relative to Cohorts-reported LSOA, using three methods, according to categorised number of reported LSOAs in NHS_E (1, quartile)



This corresponds to total number of LSOA reported, which includes repeated LSOAs that correspond to different time-periods (non-consecutive). The x-axis is chosen to visualise potential differences by total contact with health services, that coverage might be higher for individuals with more frequent NHS_E contacts.

Thirdly, our approach assumed that addresses reported in Cohorts are more valid or accurate than NHS_E: this may not be true. We cannot claim that high agreement between the two data sources suggest high validity of recorded addresses dates. We could not properly evaluate coverage for people who are living in multiple addresses, as it might not be captured in Cohorts, nor properly labelled in NHS_E. We assumed the mechanisms of recording and updating addresses to be the same between cohort studies, and between Cohorts and NHS_E, by applying the same date ascertainment method for both data sources. Future replications could apply different date ascertainment methods sensitive to data generation mechanisms of each data source.

For our supplementary analysis, the Beta distribution fitted in this analysis is derived from the smallest observed administrative interval that encloses a known change date in the Cohort–NHS_E data. As such, it captures patterns of relative timing within these minimal enclosing intervals, but does not reflect the full range of possible change-date variability when address observations are further apart in time. In practice, this means that the latent model characterises fine-scale within-interval noise, conditional on relatively short detection windows, and may underestimate the total uncertainty present in change-time ascertainment for records with longer unobserved periods. However, our main comparisons focus on relative performance between ascertainment strategies, and the supplementary multiple-imputation analysis using the latent Beta model produced results that were closely aligned with the deterministic and stochastic rules. This suggests that, while the model may not capture all sources of variability, its scope is sufficient for the comparative aims of this study. Approaches that explicitly model larger enclosing intervals, such as hierarchical timing models, could extend this work to address both within-interval and between-interval variability.

Conclusion

When possible, the choice of date ascertainment methods should be advised by external validation, where an alternative data source is seen as more valid than another. With presence of such external geodata (such as the Cohorts), formal evaluation of date ascertainment methods allows direct comparison of their impact on the main analysis. This study shows that choice of method for estimating timing of address changes may not substantially change the findings. However, any statistical approach to address poor quality temporal data for address movement will remain poor estimation of precise dates of movement. For longitudinal cohorts, when possible, we recommend implementing routine checks and internal correction of addresses with participants at each contact. Much more accurate information for some movements do exist within the UK's routine records (e.g., Driver and Vehicle Licensing Agency drivers licence changes, property transaction records) and these could be used to generate more accurate reference lists for research purposes, if public rights were respected and a protocol could be developed which established a social licence for this purpose. However, such a system would likely have substantial bias (e.g., towards car and property owners) which may impact on research and policy inequities.

Ethics statement

The UK Longitudinal Linkage Collaboration (UK LLC) is the national Trusted Research Environment (TRE) for the UK's longitudinal research community. Led by the University of Bristol and University of Edinburgh, UK LLC is designed to support the UK's unparalleled collection of Longitudinal Population Studies (LPS) by providing record linkage and secure analysis and data curation services. This project has been approved by UK LLC and its contributing data owners as part of a programme of work evaluating record linkage quality underpinning COVID-19 research. Information about this project and its outputs can be accessed via UK LLC's Data Use Register (https://ukllc.ac.uk/data-use-register/). UK LLC has ethical approval from the Health Research Authority Research Ethics Committee (Haydock Committee; ref: 20/NW/0446). Geographical data are maintained by England: '©Local Government Information House Limited copyright and database rights 2021, RDCA-395.' and Wales: 'Hawlfraint a hawliau cronfa ddata cyfyngedig Ty Gwybodaeth ar Lywodraeth Leol 2021, RDCA-395.'

Contributions

JL conceptualized the project, designed, analysed and wrote up the first draft. All authors contributed to critical reviewing and revising the manuscript. All authors read and approved the final manuscript before submission and agreed with the decision to submit the manuscript.

Acknowledgements

This work uses data accessed within UK LLC's Trusted Research Environment (TRE). We thank the Secure eResearch Platform (SeRP UK) team at Swansea University and NHS Digital Health and Care Wales for providing the TRE's infrastructure and support. This work uses data provided by participants of the contributing LPS, which have been collected through their LPS or as part of their care and support and/or interactions with UK government services. We wish to recognise and thank the participants and each contributing LPS team. We thank the following LPS for contributing data that made this research possible

- Avon Longitudinal Study of Parents and Children (ALSPAC)
- Born in Bradford (BIB)
- *Disorders Genetics Initiative (EDGI) and COVID-19 Psychiatry and Neurological Genetics (COPING) Study*
- Extended Cohort for E-Health, Environment and DNA (EXCEED)
- Medical Research Council (MRC) National Survey of Health and Development (NSHD)
- TwinsUK

This work uses data provided by patients and collected by the NHS as part of their care and support. NHS England collates patient data and gives permission for publicly beneficial uses via its Data Access Service. We thank the NHS and particularly NHS England for their work in curating participants' health records and for making these available for public good research designed to improve health services. Copyright © (2025), NHS England. Data re-used with the permission of NHS England. All rights reserved.

We thank SAIL Databank for their work in curating participants' health records and for making these available for public good research designed to improve health services. We thank the University of Leicester and City St George's, University of London for providing geospatial data and Ordnance Survey for providing AddressBase® Plus.

This research benefited from infrastructure provided by the National Institute for Health Research (NIHR) Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the NIHR or the UK Department of Health.

Funding

This work was supported by the Wellcome Trust [212953/Z/18/Z]. This research was supported by Health Data Research UK through the Social and Environmental Determinants of Health Drive Programme (for AB, RB and RG), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities (HDR-23003). UK LLC (and AB, RT) is funded by UK Research & Innovation (Medical Research Council MR/X021556/1, Economic and Social Research Council ES/X000567/1) and was initially established by the Health and Wellbeing National Core Study for COVID-19 research which was funded by the UK Government (MC_PC_20030, MC_PC_20059).

Conflict of interests

All authors declare no competing interests.

Availability of data and materials

Data used in this research are made available via UK Longitudinal Linkage Collaboration (UK LLC), which is a Trusted Research Environment (TRE) developed and operated by the Universities of Bristol and Edinburgh using an underlying 'Secure eResearch Platform' provided by Swansea University. These data cannot be used or shared outside this environment. Researchers can apply to access the UK LLC TRE using the procedure outlined in the UK LLC Data Access and Acceptable Use Policy. UK LLC uses a system of managed open access for researchers who demonstrate their project is intended to improve the public good.

References

1. Hughes AE, Pruitt SL. The Utility of EMR Address Histories for Assessing Neighborhood Exposures. Ann

Epidemiol. 2017 Jan;27(1):20–6. https://doi.org/10.1016/j.annepidem.2016.07.016

2. Northcross AL, Hwang N, Balakrishnan K, Mehta S. Assessing Exposures to Household Air Pollution in Public Health Research and Program Evaluation. Ecohealth. 2015;12(1):57–67. https://doi.org/10.1007/s10393-014-0990-3

3. McConnell R, Islam T, Shankardass K, Jerrett M, Lurmann F, Gilliland F, et al. Childhood incident asthma and traffic-related air pollution at home and school. Environ Health Perspect. 2010 Jul;118(7):1021–6. https://doi.org/10.1289/ehp.0901232

4. Ludwig J, Sanbonmatsu L, Gennetian L, Adam E, Duncan GJ, Katz LF, et al. Neighborhoods, obesity, and diabetes–a randomized social experiment. N Engl J Med. 2011 Oct 20;365(16):1509–19. https://doi.org/10.1056/NEJMsa1103216

5. Diaz V, Mainous A, Baker R, Carnemolla M, Majeed A. How does ethnicity affect the association between obesity and diabetes? DIABETIC MEDICINE. 2007 Nov;24(11):1199–204. https://doi.org/10.1111/j.1464-5491.2007.02244.x

6. Harari-Kremer R, Calderon-Margalit R, Yuval, Broday D, Kloog I, Raz R. Exposure errors due to inaccurate residential addresses and their impact on epidemiological associations: Evidence from a national neonate dataset. International Journal of Hygiene and Environmental Health. 2022 Sep 1;246:114032. https://doi.org/10.1016/j.ijheh.2022.114032

7. Youens D, Preen DB, Harris MN, Moorin RE. The importance of historical residential address information in longitudinal studies using administrative health data. Int J Epidemiol. 2018 Feb 1;47(1):69–80. https://doi.org/10.1093/ije/dyx156

8. Fecht D, Garwood K, Butters O, Henderson J, Elliott P, Hansell AL, et al. Automation of cleaning and reconstructing residential address histories to assign environmental exposures in longitudinal studies. Int J Epidemiol. 2020 Apr;49(Suppl 1):i49–56. https://doi.org/10.1093/ije/dyz180

9. Davies J, Bailey R, Mizen A, Pouliou T, Fry R, Pedrick-Case R, Stratton G, Johnson R, Christian H, Lyons R, Griffiths L. Residential mobility amongst children and young people in Wales: A longitudinal study using linked administrative records. International Journal of Population Data Science. 2024 Sep 17;6(1):2398. https://doi.org/10.23889/ijpds.v9i1.2398

10. Bell ML, Belanger K. Review of research on residential mobility during pregnancy: consequences for assessment of prenatal environmental exposures. J Expo Sci Environ Epidemiol. 2012 Sep;22(5):429–38. https://doi.org/10.1038/jes.2012.42

11. Millett C, Zelenyanszki C, Binysh K, Lancaster J, Majeed A. Population mobility: characteristics of people registering with general practices. Public Health. 2005 Jul 1;119(7):632–8. https://doi.org/10.1016/j.puhe.2004.09.004

12. Gupta AK, Nadarajah S, editors. Handbook of Beta Distribution and Its Applications. Boca Raton: CRC Press; 2004. 600 p.

13. Boyd A, Evans KM, Turner EL, Flaig R, Oakley J, Campbell KC, Thomas R, McLachlan S, Crane M, Whitehorn R, Calkin R. UK Longitudinal Linkage Collaboration (UK LLC): The National Trusted Research Environment for Longitudinal Research. International Journal of Population Data Science. 2025 Feb 17;10(1):2468. https://doi.org/10.23889/ijpds.v10i1.2468

14. Freeze number 1 dated August 2022 — UK LLC Dataset Documentation [Internet]. [cited 2025 Aug 15]. Available from: https://ukllc-book.netlify.app/docs/ukllc_key_facts/sample/freezes/freeze1

15. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013;42(1):111–27. https://doi.org/10.1093/ije/dys064

16. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort Profile: the Born in Bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42(4):978–91. https://doi.org/10.1093/ije/dys112

17. John C, Reeve NF, Free RC, Williams AT, Ntalla I, Farmaki AE, et al. Cohort Profile: Extended Cohort for E-health, Environment and DNA (EXCEED). Int J Epidemiol. 2019;48(3):678–679j. https://doi.org/10.1093/ije/dyz073

18. Bright SJ, Hübel C, Young KS, Bristow S, Peel AJ, Rayner C, et al. Sociodemographic, mental health, and physical health factors associated with participation within re-contactable mental health cohorts: an investigation of the GLAD Study. BMC Psychiatry. 2023 Jul 26;23(1):542. https://doi.org/10.1186/s12888-023-04890-x

19. Verdi S, Abbasian G, Bowyer RCE, Lachance G, Yarand D, Christofidou P, et al. TwinsUK: The UK Adult Twin Registry Update. Twin Res Hum Genet. 2019 Dec;22(6):523–9. https://doi.org/10.1017/thg.2019.65

20. Kuh D, Pierce M, Adams J, Deanfield J, Ekelund U, Friberg P, et al. Cohort profile: updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research. Wadsworth M CS Franklyn J, Trowell J, Halsall I, editor. Int J Epidemiol. 2011;40(1):e1-9. https://doi.org/10.1093/ije/dyq231

21. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort Profile: The 1946 National Birth Cohort (MRC National

Survey of Health and Development). International Journal of Epidemiology. 2006 Feb 1;35(1):49–54. https://doi.org/10.1093/ije/dyi201

22. Morris T, Manley D, Northstone K, Sabel CE. On the move: Exploring the impact of residential mobility on cannabis use. Social Science & Medicine. 2016 Nov 1;168:239–48. https://doi.org/10.1016/j.socscimed.2016.04.036

23. Hodgson S, Lurz PWW, Shirley MDF, Bythell M, Rankin J. Exposure misclassification due to residential mobility during pregnancy. International Journal of Hygiene and Environmental Health. 2015 Jun 1;218(4):414–21. https://doi.org/10.1016/j.ijheh.2015.03.007

24. Knighton AJ. Is a Patient's Current Address of Record a Reasonable Measure of Neighborhood Deprivation Exposure? A Case for the Use of Point in Time Measures of Residence in Clinical Care. Health Equity. 2018 May 1;2(1):62–9. https://doi.org/10.1089/heq.2017.0005

25. Gambaro L, Joshi H, Lupton R, Fenton A, Lennon MC. Developing Better Measures of Neighbourhood Characteristics and Change for Use in Studies of Residential Mobility: A Case Study of Britain in the Early 2000s. Appl Spatial Analysis. 2016 Dec 1;9(4):569–90. https://doi.org/10.1007/s12061-015-9164-0

26. Barr PJ, Shuttleworth I. Reporting address changes by migrants: The accuracy and timeliness of reports via health card registers. Health & Place. 2012 May 1;18(3):595–604

27. Hoek G, Vienneau D, de Hoogh K. Does residential address-based exposure assessment for outdoor air pollution lead to bias in epidemiological studies? Environmental Health. 2024 Sep 17;23(1):75. https://doi.org/10.1016/j.healthplace.2012.01.005

28. Chen L, Bell EM, Caton AR, Druschel CM, Lin S. Residential mobility during pregnancy and the potential for ambient air pollution exposure misclassification. Environ Res. 2010 Feb;110(2):162–8. https://doi.org/10.1016/j.envres.2009.11.001

Abbreviations

UK LLC:	United Kingdom Longitudinal Linkage Collaboration
LSOA:	Lower Super Output Areas