

Improving the Spatial Identification of Early Years Providers in England: A Comparison of Practical UPRN Linkage Approaches

Joana Cruz

joana.cruz@ucl.ac.uk

University College London

Bin Chi

Ed Lowther

University College London

Peat Allan

Pia Hardelid

University College London

Research Article

Keywords: childcare provision, geocoding, UPRN, spatial data infrastructure, service planning, spatial inequality

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9695052/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Accurate geolocation of Early Years Providers in England is important for understanding the spatial distribution of childcare provision and for supporting service planning, equity monitoring, and linkage to other administrative and spatial datasets. However, provider records in the Ofsted Childcare Providers and Inspections dataset do not routinely include Unique Property Reference Numbers, requiring retrospective address matching. We evaluated five address-matching approaches for assigning Unique Property Reference Numbers to 6,442 Early Years Provider records from the 2017 Ofsted register extract. Preprocessing was standardised across methods. A stratified validation subset enriched for agreement, disagreement, and candidate ambiguity was manually reviewed to assign AddressBase® Premium reference identifiers for 585 records, of which 565 had a definitive AddressBase® Premium identifier. Linkage was defined as whether a method returned at least one Unique Property Reference Number. For methods returning a single identifier, matching performance was assessed using precision, recall, and the F1 score against the AddressBase® Premium reference. For the Ordnance Survey Match & Cleanse API, evaluation used a candidate-set match definition rather than deterministic selection of a single final identifier. Because the approaches differed in output structure and decision rules, results should be interpreted as a comparison of practical UPRN linkage workflows rather than a like-for-like comparison of algorithms. Within the validation subset, linkage coverage ranged from 51% to 99%. Among single-output methods, ASSIGN 1 and the bespoke SQL workflow provided the most balanced precision–recall profile. High linkage coverage did not guarantee correct matching. Routine inclusion of Unique Property Reference Numbers within childcare provider registers would strengthen the spatial visibility of early years provision and support more reliable planning and cross-sector linkage.

Introduction

Early years providers (EYPs) form part of the local service infrastructure that shapes children's daily environments and families' access to care. Understanding where these settings are located is important for monitoring the spatial distribution of provision, identifying areas of undersupply or poor accessibility, and supporting local planning decisions concerning childcare capacity, neighbourhood need, and service equity (Academy of Medical Sciences, 2024; Magnuson et al., 2007; Melhuish & Gardiner, 2021). In England, formal childcare was used by 64% of children aged 0 to 4 years in 2024, equivalent to around 1.9 million children (DfE, 2025). This childcare is delivered through a mixed and decentralised system of private, voluntary, school-based and home-based providers (Stewart et al., 2024); the ability to locate providers accurately is therefore important for both policy and research.

Accurate location data on EYPs are needed to assess whether provision is equitably distributed and accessible, particularly in areas with high deprivation or unmet childcare need (Hurley et al., 2024). They also underpin spatial analyses of service catchments, neighbourhood context, and accessibility, and enable linkage to environmental and administrative datasets for wider place-based analysis. Without reliable spatial identifiers, such analyses may be inconsistent, incomplete, or biased, limiting their value for evidence-based planning.

The main administrative source of information on EYPs in England is the Ofsted Childcare Providers and Inspections dataset, which includes registered childcare providers on the Early Years Register and Childcare Register. Ofsted is the national inspectorate for childcare, education, and children's social care. Although the dataset contains address fields, it does not routinely include Unique Property Reference Numbers (UPRNs), limiting consistent linkage to geospatial and other administrative datasets.

UPRNs are unique, persistent identifiers assigned to addressable locations in Great Britain through Ordnance Survey AddressBase products (OS, 2023). They provide a standardised way to link records across administrative and spatial datasets. Despite growing policy emphasis on UPRN adoption across the public sector (Geospatial Commission, 2023), these identifiers are not yet consistently embedded in many administrative data systems, including Ofsted records. Reported barriers include resource constraints, inconsistent address-entry practices, and limited technical capacity (Knight & Pollock, 2025).

Where UPRNs are not already recorded, they must be assigned retrospectively using address-matching algorithms. However, address linkage is particularly challenging in institutional datasets such as EYP records, where addresses may be incomplete, informal, or associated with shared premises rather than clearly defined single-purpose delivery points. In some cases, such as providers operating from church halls, community centres, school sites, or other shared venues, the source record may support a robust spatial linkage but not an unambiguous claim to a single property-level “true” match without additional business rules. In this context, methods may differ not only in whether they return a UPRN, but also in whether the returned UPRN represents the most defensible spatial identifier for the provider record.

This study evaluates five address-matching methods for assigning UPRNs to EYP records from the Ofsted dataset. Specifically, we compare practical UPRN linkage approaches in terms of linkage coverage and correct matching against a manually assigned AddressBase® Premium (ABP) reference UPRN, while recognising that the approaches differ in output structure, decision rules, and intended operational use. By distinguishing between whether a method returns a UPRN and whether that UPRN is the correct match, we aim to clarify the balance between completeness and accuracy, and to identify the implications of different approaches for research and operational use in early years data. In

doing so, the study addresses a practical spatial-data problem: how to generate reliable, linkage-ready geographic identifiers for an important but poorly georeferenced component of children’s service infrastructure.

Methods

In this study, a **link** was defined as any instance in which a method returned at least one UPRN for a record. A **match** was defined as a returned output that correctly identified the manually assigned, or most defensible, ABP UPRN for that record. This distinction is important because some methods are designed to minimise non-linkage, which can increase apparent linkage coverage without improving correct matching. Address matching is often hindered by poor data quality, informal address structures (e.g. EYPs located in church halls or community centres), and inconsistent formatting (Kostadinovska et al., 2019). In this context, address-matching methods can vary substantially both in whether they return a UPRN and in whether the returned UPRN is the correct match.

Data Sources and Characteristics

We used two primary datasets: the Ofsted Childcare Providers and Inspections dataset (extract dated 31 December 2017) (Ofsted, 2021) and Ordnance Survey AddressBase® Premium (ABP; October 2023 release) (OS, 2023). We used the 31 December 2017 Ofsted extract to reflect a realistic linkage scenario in which administrative address data may be temporally misaligned with the reference address file and may contain outdated or inconsistent provider names due to changes over time. This allowed us to assess method performance under conditions more representative of longitudinal administrative data linkage than a fully contemporaneous, optimised linkage scenario. Residual discrepancies may arise from changes in the built environment (e.g. demolition, redevelopment, or subdivision), updates to address records within ABP, or changes in provider characteristics such as renaming, re-registration, or relocation.

The Ofsted dataset provides a national snapshot of registered childcare providers in England. Fields used for linkage were Provider Address 1–3 and Provider Postcode (Table S1). These address fields contain semi-structured address components (e.g. organisation or building name, street name, locality, and town), typically distributed across multiple lines rather than standardised into discrete elements. Provider Name, Provider Type (childminder, childcare on non-domestic premises, childcare on domestic premises, or home childcarer), Provider Unique Registered Number, and Register Combinations (Early Years Register [EYR], Compulsory Childcare Register [CCR], and Voluntary Childcare Register [VCR]) were retained for contextual validation and disambiguation. Records containing partially or fully redacted address components (e.g. childminders operating from private residences) were excluded from linkage.

ABP served as the reference dataset. It assigns each addressable location a UPRN and includes structured address components (sub-building, building number or name, thoroughfare, locality, post town, and postcode), classification codes, and organisational information. A full list of variables used from ABP is provided in Supplementary Table S1.

Study population and analytic sample

From 80,471 Ofsted records, we restricted the dataset to providers registered as “EYR Only” (Early Years Register only), defined as those caring for children from birth to 31 August following their fifth birthday and required to comply with the statutory Early Years Foundation Stage framework. Ofsted providers can be registered on one or more registers: EYR, CCR, and VCR. We restricted the analysis to EYR-only providers to define a more homogeneous population of early years settings and to avoid duplication or heterogeneity introduced by providers operating across multiple registers with differing regulatory requirements and address characteristics. After excluding records with redacted or unusable address information, 6,442 EYP records remained for linkage.

Because the aim was to assess the suitability of these methods for producing spatially usable identifiers for EYP records, evaluation focused not only on whether a UPRN was returned, but on whether the returned output correctly identified the ABP UPRN.

Address preprocessing

To support comparability across methods, address inputs were standardised before linkage. Provider Address 1–3 were concatenated in order to form a single-line address string, preserving the original sequence of address components. Address strings were cleaned consistently across methods by trimming leading and trailing whitespace, collapsing empty strings to missing, and standardising postcode formatting. Postcodes were converted to uppercase and internal spaces removed. This preprocessing ensured that all methods operated from the same underlying Ofsted address content, although some method-specific inputs differed by design. In particular, ASSIGN 2 additionally prepended provider name to the concatenated address string, and the bespoke method additionally used postcode-linked geography (Output Area [OA] and Lower Layer Super Output Area [LSOA]) via the Office for National Statistics Postcode Directory (ONSPD).

Address-matching methods

We evaluated five address-matching approaches: addressMatchR, OS Match & Cleanse API, ASSIGN 1, ASSIGN 2, and a bespoke SQL-based method. For each method, we evaluated whether a UPRN was returned (**link status**) and whether the returned output met the study definition of a correct match against ABP (**match status**). The five approaches were not assumed to be directly equivalent algorithms. Rather, they represent practical linkage workflows that could plausibly be used to assign UPRNs to EYP records. Some approaches return a single UPRN decision, whereas OS Match & Cleanse returns one or more candidate UPRNs. The bespoke SQL workflow also differs from packaged tools because it combines contextual geographic restriction, textual matching, numeric validation, and ABP classification filters. The comparison therefore evaluates how these practical approaches behave when applied to the same underlying Ofsted address records, rather than testing identical algorithms under identical decision rules.

addressMatchR

addressMatchR (R v4.3.1) implements a two-stage matching procedure (Krenz et al., 2023; NHS BSADS, 2023). First, exact matches are identified by joining on cleaned postcode and the full cleaned single-line address string. Records not matched exactly are then restricted to candidate lookup records sharing the same postcode. Within postcode, addresses are tokenised, with numeric tokens up-weighted relative to non-numeric tokens. Candidate pairs are scored using a combination of exact token matches and Jaro–Winkler similarity applied to non-numeric tokens, retaining token-level Jaro–Winkler scores greater than 0.8. Token-level scores are aggregated to an address-level score and normalised by the maximum attainable token weight for the primary address. For each primary address, the highest-scoring candidate or candidates are retained. For evaluation, a single UPRN per record was selected by ordering candidates by descending score (and match type where applicable) and retaining the first record. No additional score threshold was applied.

OS Match & Cleanse API

OS Match & Cleanse API is a proprietary cloud-based address-matching service operated by Ordnance Survey (Ordnance Survey (GB), 2026). It returns one or more candidate UPRNs with associated similarity scores in the range 0 to 1. The internal scoring rules and thresholds are not publicly disclosed.

For this study, cleaned single-line addresses were submitted via Python (v3.12; requests v2.31), API rate limits were respected, and all returned candidate UPRNs were retained. A record was classified as **linked** if at least one candidate UPRN was returned. A **match** was defined under a candidate-set definition: the ABP UPRN appeared anywhere within the returned candidate set. Because no reproducible external rule was applied to resolve multiple returned candidates to a single final UPRN, OS Match & Cleanse was evaluated as a set-valued method and its match status is not directly comparable to the single-output decision process used for the other methods.

ASSIGN

ASSIGN (October 2024 release) is an open-source deterministic algorithm designed to link structured address components to UPRNs (Harper et al., 2021). It parses address fields into components (e.g. building number, street, and postcode) and applies rule-based comparisons against ABP records. ASSIGN was originally developed for linkage of residential addresses within NHS administrative data and restricts candidate matches to predominantly residential property types and a subset of commercial properties that may serve as places of residence (e.g. flats above shops) (Harper et al., 2021). This design feature is relevant to EYP records, which are often located in non-residential or mixed-use buildings, and may reduce linkage coverage relative to methods that do not impose such constraints.

Two input configurations were evaluated:

- **ASSIGN 1**: concatenated address lines 1–3 plus postcode.
- **ASSIGN 2**: provider name prepended to concatenated address lines 1–3 plus postcode.

ASSIGN returns a categorical match classification. For this study, any returned UPRN classified by ASSIGN as a match was treated as linked. No probabilistic scoring or additional thresholding was applied.

Bespoke rule-based method

A custom rule-based workflow was implemented in PostgreSQL. This approach was designed as a practical linkage workflow rather than a direct equivalent of packaged matching tools. It combined geographic restriction, textual matching, numeric validation, and classification-

based filtering. Candidate ABP records were restricted using postcode and, where relevant, postcode-linked Output Area (OA) and Lower Layer Super Output Area (LSOA) identifiers derived from the ONS Postcode Directory. PostgreSQL full-text search functions were then applied to tokenised address fields. Building numbers and related numeric identifiers were required to match exactly where present, to reduce cross-number mislinkage. ABP classification codes were used to restrict candidates to property types considered plausible for childcare provision or shared institutional use. A record was classified as linked when a single UPRN satisfied the combined textual, numeric, geographic, and classification constraints. That returned UPRN was then compared with the manually assigned ABP reference UPRN to determine match status.

An overview of the methods is provided in Table 1.

Table 1

Summary of the address-matching methods evaluated. Transparency refers to the extent to which method rules and scoring procedures are openly described and reproducible. Geographic constraints refer to the use of postcode or area-based identifiers to restrict candidate matches. Output definition refers to the form of the returned result used in evaluation. Match definition refers to how a correct match was defined for each method. Ease of use reflects implementation complexity and user effort, and scalability reflects the ability to process large datasets efficiently.

Feature	addressMatchR	OS Match & Cleanse API	ASSIGN 1 / ASSIGN 2	Bespoke
Type	Open-source R package	Proprietary cloud-based API	Open-source deterministic tool	Custom rule-based SQL workflow
Matching logic	Hybrid deterministic plus token-weighted similarity	Proprietary similarity scoring	Deterministic rule-based matching	Deterministic textual, numeric, and geographic matching
Transparency	High	Low	High	High
Input used in this study	Concatenated address lines 1–3 plus postcode	Concatenated address lines 1–3 plus postcode	ASSIGN 1: concatenated address lines 1–3 plus postcode; ASSIGN 2: provider name plus concatenated address lines 1–3 plus postcode	Concatenated address lines 1–3 plus postcode, with OA/LSOA linkage via ONSPD
Geographic constraints	Postcode-based filtering	Internal to OS service	Postcode-based filtering	Postcode, OA, or LSOA restriction
Fuzzy matching	Yes (Jaro–Winkler)	Yes (proprietary)	No	Yes (text matching with numeric constraints)
Method output used in evaluation	Single highest-ranked UPRN	Candidate set of one or more UPRNs	Returned UPRN classified by ASSIGN as a match	Single UPRN satisfying all constraints
Match definition used in evaluation	Returned UPRN equals ABP UPRN	ABP UPRN appears anywhere in candidate set	Returned UPRN equals ABP UPRN	Returned UPRN equals ABP UPRN
Ease of use	High	Very high	Moderate	Moderate
Scalability	High	Very high (rate-limited)	High	Moderate
Software environment	R (v4.3.1)	Python (v3.12), web API	ASSIGN release October 2024	PostgreSQL
Special features	Returns UPRN, match type, and confidence score	Returns candidate UPRNs and match metadata	Designed for structured residential address linkage	Uses ONSPD to support postcode-linked geographic restriction

Validation sample construction

To evaluate performance, we constructed a stratified validation subset from the 6,442 eligible EYP records after preprocessing and execution of all five methods. Manual assignment of an ABP UPRN to all records was not feasible. The validation subset was therefore deliberately enriched for records exhibiting disagreement and ambiguity across methods, rather than designed to provide population-representative estimates of absolute performance.

Here, linkage behaviour refers to patterns in how different methods handled the same record, specifically whether non-OS methods returned the same UPRN, returned different UPRNs, or whether OS Match & Cleanse returned multiple candidate UPRNs.

For each record, we derived three indicators:

- **All-method agreement:** whether addressMatchR, ASSIGN 1, ASSIGN 2, and the bespoke method returned the same UPRN
- **Method disagreement:** whether at least two of these non-OS methods returned different UPRNs
- **OS ambiguity:** whether OS Match & Cleanse returned more than one candidate UPRN

We then drew up to 200 records from each of the following strata:

1. records where all non-OS methods agreed
2. records where non-OS methods disagreed
3. records where OS Match & Cleanse returned multiple candidate UPRNs

Because strata were not mutually exclusive, sampling was conducted independently within each stratum and the combined sample was subsequently deduplicated using provider URN. This design was intended to support comparative evaluation across a range of linkage scenarios, with deliberate enrichment for disagreement and ambiguity, rather than to estimate population-level absolute performance. Consequently, precision, recall, and F1 scores from the validation subset should not be interpreted as straightforward population-level estimates for all EYP records. They are best understood as comparative performance measures within a validation subset deliberately constructed to include a higher proportion of difficult, ambiguous, and method-discordant cases than would be expected in the full analytic sample.

Each record in the validation sample was manually reviewed by JC to assign a reference ABP UPRN where possible. Manual review considered agreement between the Ofsted provider address fields and ABP components, including building name or number, sub-building information, street, postcode, organisation name, and venue context where available. For records located in shared or communal premises, such as schools, church halls, community centres, or multi-use buildings, the assigned UPRN was the ABP record judged to best represent the physical delivery location described in the Ofsted record. Where the Ofsted record identified a named organisation operating within a larger site, the reviewer considered whether ABP contained a distinct organisational or sub-building record, or only a parent-site record. In cases where more than one ABP UPRN could plausibly represent the provider location, and the available address/contextual information did not support a single defensible choice, the record was classified as indeterminate. These indeterminate records were excluded from match-based analyses but retained for link-status analyses.

Validation and performance metrics

For each record in the validation subset, method outputs were compared with the manually assigned ABP UPRN. Records for which no definitive ABP UPRN could be assigned were excluded from match-based analyses but retained for analyses of link status.

Performance was assessed using precision (positive predictive value), recall (sensitivity), and F1 score, which summarise the trade-off between completeness and correct matching (Harron et al., 2016). These metrics were prioritised because true negatives are uncommon in this setting and measures such as specificity or negative predictive value are therefore less informative.

Record-level classification

For records with a non-missing ABP UPRN, outputs were classified as follows.

For **single-output methods** (addressMatchR, ASSIGN 1, ASSIGN 2, and the bespoke method):

- **True positive (TP):** a returned UPRN matched the ABP UPRN
- **False positive (FP):** a returned UPRN did not match the ABP UPRN
- **False negative (FN):** an ABP UPRN existed but no correct match was returned

For **OS Match & Cleanse**, evaluated under the candidate-set match definition:

- **True positive (TP):** the ABP UPRN appeared anywhere in the returned candidate set
- **False positive (FP):** at least one candidate UPRN was returned, but the ABP UPRN did not appear in the candidate set
- **False negative (FN):** no candidate set containing the ABP UPRN was returned

Because almost all valid addresses corresponded to an ABP UPRN, true negatives were rare in this context. Metrics that rely on true-negative counts therefore provided limited additional insight.

Performance metrics

Performance comparison focused on:

- **Precision (PPV):** $TP / (TP + FP)$
- **Recall (sensitivity):** $TP / (TP + FN)$
- **F1 score:** harmonic mean of precision and recall

Precision quantifies the proportion of returned outputs that were correct matches under the study definition. Recall quantifies the proportion of records with an ABP UPRN that were successfully matched under the study definition. The F1 score summarises the trade-off between completeness and correct matching.

Agreement analysis

To compare methods beyond summary performance metrics, we derived binary indicators of link status and match status for each method. For single-output methods, link status was defined as whether a non-missing UPRN was returned, and match status as whether the returned UPRN equalled the ABP UPRN. For OS Match & Cleanse, link status was defined as whether at least one candidate UPRN was returned, and match status as whether the ABP UPRN appeared anywhere within the returned candidate set.

Pairwise summaries were then calculated on these binary outcomes. For the link-level comparison, based on all records in the validation subset, the reported value is the proportion of records for which **both methods returned a UPRN**. For the match-level comparison, restricted to records with a non-missing ABP UPRN, the reported value is the proportion of records for which **both methods correctly matched the ABP UPRN**. These summaries are based on binary link and match outcomes and do not reflect exact equality of returned UPRNs across methods.

Reproducibility and data availability

All preprocessing, linkage, and evaluation procedures were implemented in R (version 4.3.1) and Python (version 3.12). Software versions and package dependencies were recorded, and session information was captured at runtime to support reproducibility. Performance metrics, linkage summaries, and pairwise comparison outputs were generated programmatically and exported as machine-readable CSV files.

Due to licensing restrictions, ABP record-level data and derived linkage outputs cannot be publicly shared. The analytic code used for preprocessing, linkage evaluation, metric calculation, and pairwise comparison will be made available via the project repository upon acceptance.

Results

The Ofsted Childcare Providers and Inspections dataset contained 80,471 provider records across England (Figure 1), from which we restricted to providers registered as **EYR Only** ($n = 6,615$). After excluding records with partially or fully redacted address information ($n = 173$), the final analytic sample comprised 6,442 EYP records. These records formed the basis for address cleaning, linkage to ABP, and evaluation of the five address-matching methods.

Across the full analytic sample ($n = 6,442$), linkage coverage, defined as the proportion of records for which at least one UPRN was returned, varied by method, ranging from 45.7% for ASSIGN 2 to 99.9% for OS Match & Cleanse API. addressMatchR achieved 98.8% linkage coverage, the bespoke method 81.5%, and ASSIGN 1 78.7%. These full-sample linkage coverage estimates should not be directly compared with those observed in the validation subset, which was deliberately enriched for records showing disagreement or ambiguity across methods.

Manual assignment of a definitive ABP UPRN was not feasible for all records; therefore, a stratified validation subset was constructed to enable detailed assessment of matching performance across records with differing levels of ambiguity. The validation dataset comprised 585 EYP records (Figure 1), derived using stratified sampling based on linkage behaviour across methods, with duplicates removed following sampling. A definitive ABP UPRN could be assigned for 565 records (96.6%); 20 records (3.4%) remained indeterminate after manual review.

Linkage coverage in the validation subset

Detailed linkage coverage results for the validation subset are presented in Table 2. Although the ordering of methods was broadly similar to that observed in the full analytic sample, absolute linkage coverage in this validation subset should be interpreted in light of the stratified sampling design. Within this subset ($n = 585$), linkage coverage varied across methods, with addressMatchR linking 99% of records, followed

by the bespoke method (83%), ASSIGN 1 (81%), OS Match & Cleanse API (76%), and ASSIGN 2 (51%). A similar pattern was observed when restricted to records with non-missing ABP UPRNs (n = 565).

Table 2 Linkage coverage across five address-matching methods applied to the stratified validation subset (n = 585). Linkage coverage represents the proportion of records for which at least one UPRN was returned. Coverage is also shown for the subset with non-missing AddressBase® Premium UPRNs (n = 565)

Address-matching method	Records linked (validation subset, n = 585)	Records linked (AddressBase® Premium subset, n = 565)
addressMatchR	579 (99%)	563 (100%)
OS Match & Cleanse API	445 (76%)	445 (79%)
ASSIGN 1	476 (81%)	467 (83%)
ASSIGN 2	297 (51%)	294 (52%)
Bespoke	483 (83%)	475 (84%)

Performance metrics

Performance metrics were calculated within the validation subset restricted to records with non-missing ABP UPRNs (n = 565; Table 3). When evaluated under the candidate-set match definition, OS Match & Cleanse API included the ABP UPRN within its candidate set for 445 of 565 records, corresponding to precision 1.00, recall 0.79, and F1 0.88 under this candidate-set match definition. Because these metrics do not reflect resolution to a single final UPRN, they are not directly comparable with the single-output methods shown in Table 3. No false positives were observed under this definition because match status was based on whether the ABP UPRN appeared anywhere in the returned candidate set, rather than on selection of a single final UPRN.

Within the stratified validation subset, single-output methods showed different precision–recall profiles. addressMatchR returned a correct match for most records with an assigned ABP UPRN but also produced a substantial number of incorrect returned UPRNs, resulting in high recall (0.99) and lower precision (0.57). ASSIGN 2 showed the opposite pattern, with higher precision (0.83) but lower recall (0.47), leaving many records with an ABP UPRN unmatched under the study definition (FN = 271). ASSIGN 1 and the bespoke SQL workflow showed a more even precision–recall profile, with F1 scores of 0.79 and 0.78, respectively.

Table 3 Performance of five address-matching methods against the AddressBase® Premium UPRN (n = 565 records with a non-missing AddressBase® Premium UPRN). True positives (TP) represent records correctly matched to the AddressBase® Premium UPRN; false positives (FP) represent records for which a UPRN was returned but did not match the AddressBase® Premium UPRN; false negatives (FN) represent records with an AddressBase® Premium UPRN for which no correct match was returned. Precision (positive predictive value), recall (sensitivity), and F1 score summarise the trade-off between matching correctness and completeness. OS Match & Cleanse metrics reflect candidate-set match status and are shown for descriptive comparison only

Address-matching method	True positives (TP)	False positives (FP)	False negatives (FN)	Precision (PPV)	Recall (sensitivity)	F1 score
addressMatchR	321	242	2	0.57	0.99	0.72
OS Match & Cleanse API ¹	445	0	120	1.00	0.79	0.88
ASSIGN 1	371	96	98	0.79	0.79	0.79
ASSIGN 2	244	50	271	0.83	0.47	0.60
Bespoke	364	111	90	0.77	0.80	0.78

¹For OS Match & Cleanse, a record was classified as matched if the ABP UPRN appeared anywhere in the returned candidate set. Metrics therefore reflect **candidate-set match status** rather than a single deterministic linkage decision and are not directly comparable to methods returning one UPRN per record.

Characteristics of OS Match & Cleanse API candidate sets

Among records with at least one OS Match & Cleanse API candidate, the median number of candidate UPRNs returned per record was 1 (IQR: 1–1). Most returned candidate sets were singletons. However, the method was still evaluated under a candidate-set match definition because candidate ranking was not externally validated and no reproducible disambiguation rule was applied. The ABP UPRN appeared within the candidate set for 445 of 585 validation records (76.1%).

Agreement between linkage approaches

Across all 585 validation records, the proportion of records for which both methods returned a UPRN varied across method pairs (Table S2). The highest shared linkage coverage was observed between addressMatchR and OS Match & Cleanse API (98.9%), followed by the bespoke method and OS Match & Cleanse API (82.6%), and addressMatchR and the bespoke method (81.9%). The lowest proportion of records for which both methods returned a UPRN was observed for pairings involving ASSIGN 2, particularly with the bespoke method (47.2%) and ASSIGN 1 (49.4%).

Among the 565 records with non-missing ABP UPRNs, the proportion of records for which both methods correctly matched the ABP UPRN was more variable (Table S3). The highest shared correct matching was observed between ASSIGN 1 and OS Match & Cleanse API (59.3%), followed by the bespoke method and OS Match & Cleanse API (55.0%), and ASSIGN 1 and the bespoke method (49.7%). Lower values were observed for comparisons involving ASSIGN 2, particularly with addressMatchR (31.3%) and the bespoke method (35.8%). These results reflect the proportion of records correctly matched by both methods, rather than overall agreement in match status.

Discussion

This study examined how different practical UPRN linkage approaches affect the spatial identification of EYPs in England. Across the five approaches evaluated, methods differed substantially in whether they returned a UPRN and whether the returned output corresponded to the manually assigned ABP reference UPRN. These differences matter because inaccurate or incomplete UPRN assignment can distort the spatial visibility of childcare provision, affecting analyses of service distribution, accessibility, and neighbourhood context. However, the approaches evaluated here were not fully equivalent: most returned a single UPRN decision, whereas OS Match & Cleanse was evaluated as a candidate-set method, and the bespoke SQL approach used contextual geographic, numeric, textual, and classification-based rules. The findings should therefore be interpreted as a comparison of practical linkage workflows applied to EYP records, rather than as a definitive ranking of algorithms under identical conditions. Within this framing, high linkage coverage did not necessarily correspond to correct matching. Among single-output methods, ASSIGN 1 and the bespoke SQL workflow provided the most balanced precision–recall profile in the validation subset, while OS Match & Cleanse frequently retrieved the correct UPRN within returned candidate sets but should not be interpreted as demonstrating equivalent single-decision matching accuracy.

Comparison with existing evidence

Previous evaluations of deterministic linkage tools such as ASSIGN have reported very high accuracy when applied to structured residential address data, particularly in NHS patient records where address components are consistently formatted and aligned with AddressBase® structure (Harper, 2018; Harper et al., 2021). Our findings suggest that such performance does not readily generalise to institutional datasets such as EYP records, which frequently include informal venue descriptions, shared premises (e.g. community centres and church halls), and incomplete address components (North Somerset Council, n.d.). These characteristics depart from the structured, component-based residential formats for which deterministic rule-based systems are typically optimised, and are consistent with the lower matching performance observed here for ASSIGN 2. Importantly, these cases should not be treated simply as data errors. For some EYP records, the address may accurately describe a service operating within a larger or shared site, while ABP may contain several plausible UPRNs for different parts of that site, or only a broader parent-site record. In such cases, the practical objective may be to assign a robust and reproducible spatial linkage suitable for research or planning, rather than to claim that the source record contains enough information to identify a uniquely correct property-level match.

Automated and hybrid tools such as addressMatchR and OS Match & Cleanse API are increasingly used in operational and research pipelines; however, few studies have evaluated their performance against a manually assigned ABP UPRN in non-residential administrative data

(Krenz et al., 2023; NHS BSADS, 2023). By explicitly separating linkage coverage from correct matching, this study shows that high linkage coverage did not necessarily correspond to high-quality matches. In our evaluation, addressMatchR achieved near-complete linkage coverage but lower precision, consistent with over-matching in complex address contexts. Under the candidate-set match definition, OS Match & Cleanse showed a high proportion of candidate sets containing the correct UPRN, although its proprietary scoring and limited transparency restrict reproducibility and independent audit in research settings.

Taken together, these findings add to the address linkage literature by suggesting that algorithm performance is strongly dependent on data structure and context. In less structured institutional datasets, approaches that combine deterministic constraints with contextual or rule-based refinement may offer a more favourable trade-off between linkage coverage and correct matching.

Interpretation and implications for spatial analysis and data infrastructure

Our findings have practical relevance for researchers, planners, data custodians and policymakers relying on spatially referenced service data in early childhood datasets. In applied spatial analysis, georeferencing error is not merely a technical issue; it affects what can be measured, mapped, and compared across places. No single method performed best across all evaluated dimensions, and trade-offs between linkage coverage, correct matching, and scalability should be considered in relation to data structure and intended use. Deterministic methods such as ASSIGN may perform well on structured residential data (Harper et al., 2021), but their performance appears to decline in less formal settings such as shared or non-standard premises commonly observed in EYP records. In this study, the bespoke approach showed higher precision than some alternatives and greater adaptability to complex address structures, although it required more implementation effort.

Importantly, linkage coverage alone should not be used as a proxy for correct matching. Methods such as addressMatchR achieved high linkage coverage but also produced many incorrect single-record matches. By contrast, OS Match & Cleanse frequently returned candidate sets containing the correct UPRN, but this does not demonstrate equivalent single-record matching accuracy. These findings highlight the importance of evaluating both whether a method returns a UPRN and whether that returned output is the correct match, particularly in applications where linkage error may influence downstream analyses or policy decisions.

Strengths and Limitations

To our knowledge, this is the first study to systematically evaluate multiple UPRN assignment methods specifically for EYP records, providing comparative evidence relevant to education planning, administrative data linkage, and spatial policy research. A major strength of the study is the systematic comparison of five address-matching approaches spanning open-source, proprietary, and bespoke implementations, applied to a national administrative dataset. Performance was evaluated against a manually assigned ABP UPRN, allowing precision, recall, and F1 score to be calculated rather than relying solely on linkage coverage. The explicit distinction between link and match also provides a clearer framework for evaluating method performance than linkage coverage alone, particularly in contexts where some algorithms are designed to minimise non-linkage. A further strength is that the study addresses a common but under-discussed source of uncertainty in applied spatial analysis: the assignment of persistent geographic identifiers to semi-structured service-provider records.

Several limitations should be considered. First, the validation subset was deliberately enriched for records exhibiting method disagreement or candidate ambiguity. This design improved the ability to compare how approaches behaved in difficult linkage scenarios, but it limits the extent to which absolute precision, recall, and F1 estimates can be generalised to the full population of EYP records. The performance metrics should therefore be interpreted primarily as comparative measures within a deliberately challenging validation subset, rather than as population-representative estimates of method performance.

Second, although ABP is widely regarded as a comprehensive reference dataset for addressable locations in Great Britain, some communal or informal venues, such as shared halls or temporary facilities, may be imperfectly represented or classified within the dataset. As such venues are common among EYP records, this may introduce some misclassification in the reference standard, potentially affecting both matching performance and manual validation.

Third, OS Match & Cleanse API operates as a proprietary service. Key elements of its scoring logic, internal thresholds, and reference data versioning are not publicly disclosed, limiting methodological transparency and reproducibility in research settings. Although match status can be evaluated under a candidate-set definition, the internal ranking and decision process cannot be independently audited.

Fourth, the ABP reference UPRN was assigned through manual review, which necessarily involved judgement in complex or shared-premises cases. Although indeterminate records were excluded from match-based analyses, some residual uncertainty may remain where provider records referred to organisations operating within larger sites, communal venues, or informal premises. In such cases, the distinction between the organisation, the physical delivery point, and the ABP addressable object may not always be straightforward. This limitation reflects the nature of the linkage problem itself, rather than only error in the source data or reference dataset.

Finally, while the bespoke SQL approach showed a relatively even precision–recall profile, it required greater implementation effort and contextual configuration than packaged tools. Without further automation or standardisation, such methods may present scalability challenges in large-scale operational pipelines.

Policy and research implications

These findings highlight the value of embedding geospatial identifiers such as UPRNs directly within administrative service-provider datasets. Routine inclusion of UPRNs at the point of provider registration would remove the need for retrospective address matching, reduce avoidable uncertainty, and enable more consistent integration with spatial and administrative datasets.

More broadly, accurate provider geolocation can support routine monitoring of local childcare capacity and identification of underserved areas (“nursery deserts”) (Pollard et al., 2023). It can also support analyses of access relative to deprivation and spatial planning for children’s services within rapidly changing neighbourhoods (Almeida et al., 2024; Langford et al., 2018; Lapietra et al., 2024). It may also support emergency and contingency planning where understanding the location of child-centred settings is operationally important. Accurate provider geolocation can also enable linkage to environmental exposure data, including air pollution, heat, and greenspace, for subsequent research on child health and climate-related inequalities (An et al., 2021; Schapiro et al., 2024; Vella-Brodrick & Gilowska, 2022).

From a systems perspective, the absence of embedded UPRNs introduces inefficiencies and avoidable uncertainty into research and policy analysis. Retrospective matching requires additional resources, introduces methodological variability, and may result in differential inclusion of providers depending on address quality. Integrating UPRNs within Department for Education and Ofsted administrative systems could strengthen interoperability across the public sector, align with the objectives of the Public Sector Geospatial Agreement, and support more robust place-based early years analysis (Langford et al., 2018).

Improving geospatial infrastructure in early years datasets is therefore more than a technical enhancement; it can support service planning, environmental risk assessment, and evidence-informed policy development. Emerging approaches, including machine-learning-based building identification and volunteered geographic data, may eventually assist linkage in difficult cases, but their completeness, validation, and governance would require careful assessment before use in administrative research.

Conclusion

Reliable spatial identification of Early Years Providers is important for planning, monitoring, and research. This evaluation shows that high linkage coverage alone does not guarantee that a returned UPRN is the correct or most defensible spatial identifier, particularly in provider datasets containing shared, informal, or weakly structured premises. Distinguishing between whether a method returns a UPRN and whether that returned output corresponds to the reference ABP UPRN is therefore essential when creating linkage-ready datasets for analyses of service distribution, accessibility, and inequality. The findings should be interpreted as a comparison of practical linkage approaches within a deliberately challenging validation subset, rather than as a definitive ranking of algorithms. Routine inclusion of UPRNs within administrative provider systems would reduce the need for retrospective address matching, strengthen the spatial visibility of childcare provision, and improve interoperability across education, spatial, environmental, and wider administrative datasets.

Declarations

Author Contribution

JC: conceptualisation, methodology, data curation, address preprocessing, implementation of the addressMatchR linkage approach, manual validation, formal analysis, interpretation, writing – original draft, and writing – review and editing. BC: methodology, implementation of the ASSIGN 1 and ASSIGN 2 linkage approaches, formal analysis, interpretation, and writing – review and editing. EL: technical implementation, computing infrastructure, linkage workflow support, interpretation, and writing – review and editing. PA: development of the bespoke rule-based SQL linkage workflow, operational address-matching expertise, interpretation of linkage behaviour, and writing – review and editing. PH: conceptualisation, supervision, methodological interpretation, and writing – review and editing. All authors reviewed and approved the final manuscript.

References

1. Academy of Medical Sciences (2024). *Prioritising early childhood to promote the nation’s health, wellbeing and prosperity*. Academy of Medical Sciences. <https://acmedsci.ac.uk/file-download/16927511>
2. Almeida, V., Hoffmann, C., Königs, S., Moreno-Monroy, A. I., Salazar-Lozada, M., & Terrero-Dávila, J. (2024). *Geographic inequalities in accessibility of essential services* (OECD Social, Employment and Migration Working Papers, Issue 307). OECD Publishing.

<https://doi.org/10.1787/12bab9fb-en>

3. An, F., Liu, J., Lu, W., & Jareemit, D. (2021). A review of the effect of traffic-related air pollution around schools on student health and its mitigation. *Journal of Transport & Health*, 23. <https://doi.org/10.1016/j.jth.2021.101249>
4. DfE (2025). *Childcare and early years survey of parents: reporting year 2024*. Department for Education Retrieved from <https://explore-education-statistics.service.gov.uk/find-statistics/childcare-and-early-years-survey-of-parents/2024>
5. Geospatial Commission (2023). *UK Geospatial Strategy 2030*. Retrieved from <https://www.gov.uk/government/publications/uk-geospatial-strategy-2030>
6. Harper, G. (2018). Quality assurance of linkage of routine data for singleton and multiple births. *British Medical Journal Open*, 8(3), e017898. <https://doi.org/10.1136/bmjopen-2017-017898>. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014.
7. Harper, G., Stables, D., Simon, P., Ahmed, Z., Smith, K., Robson, J., & Dezateux, C. (2021). Evaluation of the ASSIGN open-source deterministic address-matching algorithm for allocating unique property reference numbers to general practitioner-recorded patient addresses. *Int J Popul Data Sci*, 6(1), 1674. <https://doi.org/10.23889/ijpds.v6i1.1674>
8. Harron, K., Goldstein, H., & Dibben, C. (Eds.). (2016). *Methodological developments in data linkage*. Wiley Online Library. <https://doi.org/10.1002/9781119072454>
9. Hurley, P., Tham, M., & Nguyen, H. (2024). *International childcare: Mapping the deserts*. Mitchell Institute, Victoria University.
10. Knight, A., & Pollock, A. (2025). Barriers to UPRN adoption remain despite benefits. *Property Journal*. Retrieved 29/07/2025, from <https://ww3.rics.org/uk/en/journals/property-journal/uprn-benefits-barriers-adoption.html>
11. Kostadinovska, A., Asim, M., Pletea, D., & Pauws, S. (2019). Overview of Data Linkage Methods for Integrating Separate Health Data Sources. In S. Consoli, D. Reforgiato Recupero, & M. Petković (Eds.), *Data Science for Healthcare: Methodologies and Applications* (pp. 217–238). Springer International Publishing. https://doi.org/10.1007/978-3-030-05249-2_8
12. Krenz, K., Dhanani, A., McEachan, R. R. C., Sohal, K., Wright, J., & Vaughan, L. (2023). Linking the Urban Environment and Health: An Innovative Methodology for Measuring Individual-Level Environmental Exposures. *International Journal Of Environmental Research And Public Health*, 20(3). <https://doi.org/10.3390/ijerph20031953>
13. Langford, M., Higgs, G., & Dallimore, D. J. (2018). Investigating spatial variations in access to childcare provision using network-based Geographic Information System models. *Social Policy & Administration*, 53(5), 661–677. <https://doi.org/10.1111/spol.12419>
14. Lapietra, I., Colacicco, R., Rizzo, A., & Capolongo, D. (2024). Mapping Social Vulnerability to Multi-Hazard Scenarios: A GIS-Based Approach at the Census Tract Level. *Applied Sciences*, 14(11). <https://doi.org/10.3390/app14114503>
15. Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, 26(1), 33–51. <https://doi.org/10.1016/j.econedurev.2005.09.008>
16. Melhuish, E., & Gardiner, J. (2021). *Study of early education and development (SEED): Impact study on early education use and child outcomes up to age four years*. Department for Education.
17. NHS BSADS (2023). *addressMatchR: This package matches address. R package version 0.0.0.9000*. NHS Business Services Authority Data Science. <https://github.com/nhsbsa-data-analytics/addressMatchR>
18. North Somerset Council (n.d.). *Opening a new childcare setting within North Somerset*. <https://www.n-somerset.gov.uk/sites/default/files/2020-09/Opening%20a%20new%20childcare%20setting%20in%20North%20Somerset.pdf>
19. Ofsted (2021). *Childcare providers and inspections as at 31 December 2017*. Retrieved from <https://www.gov.uk/government/statistics/childcare-providers-and-inspections-as-at-31-december-2017>
20. Ordnance Survey (GB) (2026). *OS Match and Cleanse API*. Ordnance Survey. Retrieved 24/02/2026 from <https://docs.os.uk/os-apis/accessing-os-apis/os-match-and-cleanse-api>
21. OS. (2023). *AddressBase Premium © Crown copyright and/or database rights [2023] OS (Research Licence). © Improvement and Development Agency for Local Government copyright and database rights [2023] (Research Licence)*. Ordnance Survey. <https://www.ordnancesurvey.co.uk/products/addressbase-premium>
22. Pollard, T., Coote, A., Ewart-Biggs, H., Stephens, T., & Sandher, J. (2023). *A Fair Start for All: A Universal Basic Services Approach to Early Education and Care*. In: New Economics Foundation.
23. Schapiro, L. H., McShane, M. A., Marwah, H. K., Callaghan, M. E., & Neudecker, M. L. (2024). Impact of extreme heat and heatwaves on children's health: A scoping review. *The Journal of Climate Change and Health*, 19. <https://doi.org/10.1016/j.joclim.2024.100335>
24. Stewart, K., Gambaro, L., & Reader, M. (2024). Levelling down? Understanding the decline of the maintained nursery sector in England. *British Educational Research Journal*, 51(2), 1009–1038. <https://doi.org/10.1002/berj.4104>

Figures

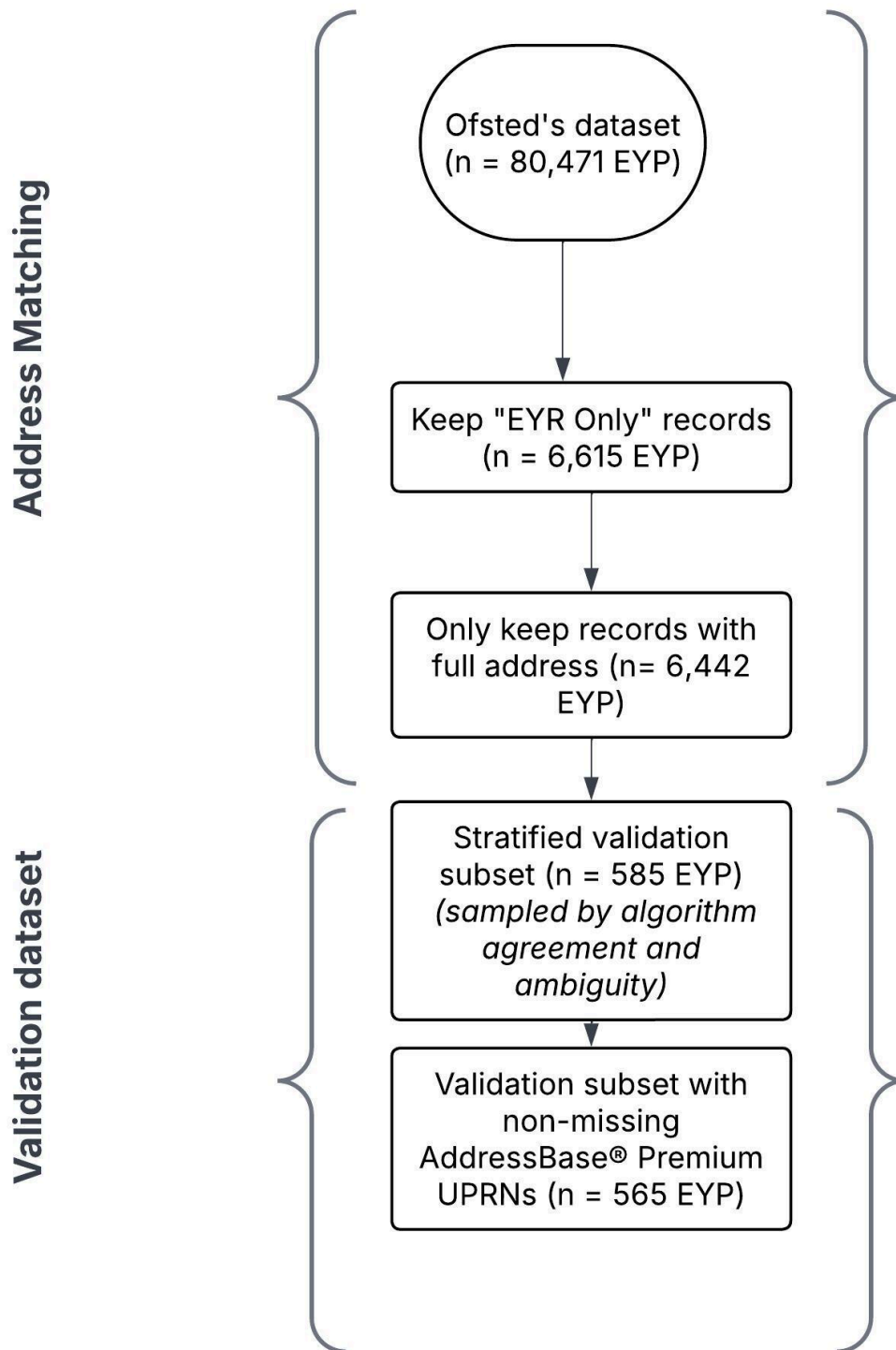


Figure 1

Flow diagram illustrating selection of Early Years Provider (EYP) records and derivation of the stratified validation subset (n = 585) used for manual verification

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables.docx](#)