International Journal of Population Data Science

Journal Website: www.ijpds.org

Improving opportunities for data linkage within Children Looked After administrative records in Wales

Grace A. Bailey[1,*], Alex Lee[1], Saira Ahmed[1], Ieuan Scanlon[1], Laura E. Cowley[1], Amy Stuart[1], Ian Farr[1], Caroline Brooks[1], Laura North[1], and Lucy J. Griffiths[1]

Submission History

Submitted:	16/01/2024
Accepted:	18/12/2024
Published:	19/02/2025

[1]Population Data Science, Swansea University Medical School, Swansea, SA2 8PP, UK

Abstract

Introduction
Linkage of population-based administrative data is a powerful tool for studying important public issues. To overcome confidentiality and disclosure issues, records are de-identified and allocated a unique identifier. Within the Secure Anonymised Information Linkage (SAIL) Databank, these are known as Anonymised Linking Fields (ALFs). Assignment of an ALF enables linkage of individuals across multiple routinely collected datasets. Within the Children Looked After (CLA) Wales dataset, only 37% of the children have an ALF, limiting linkage to other datasets and, as a result, potential research. There are also other known data issues, including discrepancies with the week of births, duplicate identifiers and year-on-year changes in identifiers.

Objectives
To improve accuracy and availability of the ALFs in the CLA dataset, and overall research quality.

Methods
Using several datasets within the SAIL Databank, we developed a six-step CLA matching algorithm to improve the ALF matching rate and correct for data errors. To assess the performance of our algorithm, we benchmarked against routine ALFs already identified via the algorithm currently used by SAIL.

Results
Our algorithm increased ALF matching by 25%, assigning 61% of individuals an ALF. Inconsistent weeks of birth, and incorrect and duplicate identifiers were resolved. When benchmarking against the current ALF-assigning algorithm used by SAIL, our algorithm had an overall sensitivity of 90%.

Conclusion
We have developed an algorithm which demonstrates comparable ALF matching performance to the current algorithm used within SAIL, and which greatly improves the ALF matching in the CLA dataset. This algorithm may help to overcome potential bias due to missing data, and increases the potential for linkage to other datasets. Further development and refinement could result in the algorithm being applied to other datasets in SAIL.

Keywords
administrative data linkage; children looked after; SAIL Databank

*Corresponding Author:
Email Address: g.a.bailey@swansea.ac.uk (Bailey GA)

https://doi.org/10.23889/ijpds.v6i1.2383

Introduction

In the United Kingdom, children looked after are defined as children under the age of 18 years who are looked after by the local authority and are in its care or provided with accommodation for a continuous period of more than 24 hours [1]. Children can become looked after under voluntary arrangements or under a compulsory care order granted by the courts. Throughout the last decade, there has been an upward trend in the number of children looked after in the UK, with the exception of Scotland where rates of children in care have been falling [2, 3]. These children are considered one of the most vulnerable groups in society, with previous work showing increased risk of poor outcomes across social, education and health domains [4–6].

In recent years, there has been growing interest in the use of administrative data to advance our understanding and insights into children in receipt of care and support, and children looked after by the state, as well as the implementation and effectiveness of social policies [7–9]. Current Government policy strongly supports the need to reduce the number of children in need of care and improve their outcomes, including the recent policy paper *Data Saves Lives* which promotes how it can improve children's social care services [10, 11]. Administrative data relates to individual or organisational level information, which has been derived from data routinely collected by public sector agencies for organisational purposes including, but not limited to, education, healthcare and social care agencies [12]. It contains abundant, detailed information that has significant potential for research.

As part of the UK's statutory guidance, local authorities are required to submit information about families and children who interact with children's statutory social services to the Government on an annual basis; this enables service provisions and outcomes to be monitored [13]. These returns, known as the Children Looked After (CLA) Census, contain individual-level data for all children and young people placed under the care of a local authority and are available for research purposes within Trusted Research Environments (a centralised, secure data platform) across all four UK nations.[1] Information about the variables within the datasets have been described elsewhere [14–18].

Within these TREs, person-level data can be linked between datasets using a unique identifier, allowing for a more comprehensive understanding about the needs of a population [19]. Within health datasets, the linkage variable is often based on unique identifiers such as the National Health Service (NHS) number (England and Wales), the Community Health Index (CHI) (Scotland) or the Health and Care number (HCN) (Northern Ireland). However, for the children's social care datasets (England, Scotland and Wales[2]), including the CLA dataset, the NHS/CHI number is not available. Instead, quasi-identifiers are used – these are attributes that are not themselves unique identifiers but in combination can be used to identify an individual (e.g. name, gender, date of birth, address and postcode) [20].

Demographic data collected within the CLA Census is limited; alongside the Unique Pupil Number (UPN) (England and Wales) or Scottish Candidate Number (SCN) (Scotland), only the date of birth, gender and ethnicity are collected. Other identifiable information such as the child's name, address and postcode are stored separately in the education dataset[3]. The UPN/SCN is used to link the two datasets to access the child's name and address. Children are automatically allocated a UPN/SCN on their entry to a maintained school. The majority of children in the UK will not attend school until after they turn four years old, although the compulsory school age in England and Wales is five years old. In the absence of the UPN/SCN, very few children under the age of four have sufficient quasi-identifiable information for probabilistic matching, meaning they do not have a unique identifier which allows linkage to other datasets. In some cases, children aged two or three may have an identifier if they attend a maintained nursery [21]. This is particularly problematic for researchers who wish to link the CLA dataset because it is well established that most children looked after are less than one years old [22].

Within the CLA dataset held in the Secure Anonymised Information Linkage (SAIL) Databank, only 37% of the CLA population have a unique identifier, the Anonymised Linkage Field (ALF), assigned via the routine algorithm (see Methods and Supplementary Appendix 1), drastically reducing the cohort sample that can be linked to other datasets [19, 23]. Previous researchers using the CLA dataset have also reported inconsistencies in the data, including discrepancies with the week of births, duplicate ALFs and year-on-year changes in quasi-identifiers amongst individuals in North Wales[4]. Lastly, the use of direct and indirect identifiers is particularly problematic for children looked after. These children are subject to recurring changes in quasi-identifiers e.g. frequent address changes which will contribute to the lack of a postcode.

To overcome the issue of limited routine ALF availability, we have developed a six-step CLA matching algorithm for the SAIL CLA dataset, with the main aim being to improve research quality by producing a standardised data table with additional matched ALFs. This tool utilises de-identified quasi-identifiers from several datasets already available to obtain the

[1]Wales: the Secure Anonymised Information Linkage (SAIL) Databank (Swansea University, Wales) (https://saildatabank.com/) England: the Office for National Statistics Secure Research Service. Note, several ADR UK flagship datasets, including Education and Child Health Insights from Linked Data (ECHILD), Longitudinal Education Outcomes (LEO) and Growing up in England (GUIE) datasets contain variables from the CLA Census via the National Pupil Database (NPD) (https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice) Scotland: The National Safe Haven is controlled by the Electronic Data Research and Innovation Service (eDRIS) who are part of Public Health Scotland (https://www.researchdata.scot/accessing-data/information-for-researchers/tres-and-data-access/) Northern Ireland: Children's social care data is available in the Social Services Client Administration and Retrieval Environment (SOSCARE) dataset (https://bso.hscni.net/directorates/digital/honest-broker-service/).

[2]SOSCARE (Northern Ireland) has an integrated health and social care system, whereby children's social care datasets contain the Health and Care Number (HCN), the equivalent of the NHS number. This allows linkage across health and social care datasets by matching of the HCN [45].

[3]Wales: Pupil Level Annual School Census (PLASC); England: National Pupil Database (NPD); Scotland: Scottish Exchange of Data (ScotXed).

[4]In the CLA collections, the four North Wales local authorities changed their child identifiers from one year to the next. This means that it is not possible to follow the same child longitudinally in these local authorities.

routine pre-assigned ALF held in SAIL. The improved ALF table has been made available as an additional resource to all researchers working with the CLA data to ensure projects can benefit from a benchmarked and reliable solution. Whilst this algorithm was developed specifically for the data held within SAIL Databank, it can be adapted to be applied to other anonymised population databases that use similar techniques to those in SAIL.

Methods

Data linkage

Within the SAIL Databank, the use of a unique Anonymised Linking Field (ALF) has been developed and implemented to enable an accurate matching process. The pseudonymisation and linkage processes have been described in detail elsewhere [19]. It is important to note that SAIL does not handle or contain any identifiable information, only pseudonymised data (e.g. week of birth instead of date of birth). Datasets are readily linkable via the ALF which is included within the datasets required for each research project. Data is pseudonymised through a split-file process by a trusted third party, as described in Supplementary Appendix 1.

In addition to the unique identifier (ALF), the social care datasets contain the system ID for each child, this is a unique child identifier assigned by the local authority and it is retained from year to year. However, in Wales, the child's identifiers do not continue with the child if the responsibility for their care is transferred to another local authority [24].

Data sources

The CLA census has previously been described elsewhere [14, 21]. Within SAIL, this table is referred to by the LACW (Looked After Children Wales) acronym however, in-keeping with the census, we will refer to it as the CLA. It is an annual census (1 April to 31 March) which collects information relating to children and young people who are 'looked after' by the local authority. In Wales, this refers to a child who is in care or provided with accommodation for a continuous period of more than 24 hours, as defined by Section 74 of the Social Services and Well-being (Wales) Act 2014 [24]. Within SAIL, CLA data is currently available between 1 April 2002 and 31 March 2021.

The following datasets are all available and pre-linked (i.e. where possible, records have an ALF assigned) within SAIL, and were used to identify additional ALFs. They include: Children and Family Court Advisory and Support Service (Cafcass Cymru/CAFW); Children in Need (CIN, referred to as CINW in SAIL); Children Looked After (CLA, referred to as LACW in SAIL); Children Receiving Care and Support (CRCS); Education Wales (EDUW), National Community Child Health Database (NCCHD); Welsh Demographic Service Dataset (WDSD); Welsh Longitudinal General Practice (WLGP) and Welsh Index of Multiple Deprivation (WIMD) 2014 (Table 1). Additional information about each of these datasets can also be found on the Health Data Research Innovation Gateway (https://www.healthdatagateway.org/).

The WDSD is a register containing demographic information about all individuals registered at a Welsh General Practice (GP). The demographic details included in this dataset are often considered as the gold standard compared with other datasets in SAIL, as they are confirmed as part of the GP registration process and used for patient contact and communication by NHS Wales. Within this dataset, residential postcodes are replaced by the Lower Layer Super Output Area (LSOA) code. Similarly, both home and placement residences are recorded and replaced by LSOAs in the social care datasets. LSOAs are geographic units designed for the reporting of small area statistics. They must have a minimum population size of 1,000, and a mean population size close to 1,600. There are 1,909 LSOAs in Wales [25]. LSOA codes can be used to ascertain an area-based measure of deprivation, known as the Welsh Index of Multiple Deprivation (WIMD) 2019. The WIMD ranks these LSOAs from most deprived to least deprived on a measure that considers a range of factors including income and employment, health, education, access to services, housing, community safety and the physical environment [25].

In addition, within SAIL, the Welsh Longitudinal General Practice (WLGP) dataset contains Read codes, a coded thesaurus of clinical terms used by general practitioners to report findings and procedures [26]. Children looked after can be identified using codes relating to 'adoption', 'care' and 'parent of looked-after-child' for the parent and child. A list of the codes is available in the concept library: https://conceptlibrary.saildatabank.com/phenotypes/PH1593/version/2945/detail/.

Matching algorithm pre-processing

Before proceeding with the primary linkage, we performed pre-linkage processing to improve match rates in the CLA using other datasets which will also be used in the primary process. Broadly, this includes:

- Linking CIN and CRCS to EDUW via the IRN to retrieve the ALF from EDUW to improve the CIN and CRCS match rates.
- Correcting cases in CLA where some local authorities changed all child system IDs between 2010 – 2012. Changes in system IDs disallowed the continuous link between a child's care records. Restoring this link allows for both improvements in the matching process and opportunities for further research. Individuals who do not have a recorded end date for their most recent episode in the CLA table and those who do not have a recorded start date for their earliest episode were identified and matched to find their pre- and post-change system IDs. These individuals were then given a new "unified" system ID, which was applied to all their pre- and post-change entries.
- Linking Cafcass Wales to NCCHD Births using the female parent ALF, child week of birth, and child gender code to improve the match rate for children in the Cafcass Wales dataset.

The matching algorithm is available here: https://github.com/SwanseaUniversityDataScience/LACW-ALF-Matching.

Table 1: Datasets and variables used in the algorithm

Dataset	Data provider	Description	Variables
Children and Family Court Advisory and Support Service (Cafcass Cymru/CAFW) (https://web.www.healthdatagateway.org/dataset/29a714e1-5289-4362-be24-2848c954344e)	Welsh Government	Cafcass is an organisation that is independent of the courts and social services but works under the rules of the Family Court and legislation to work with children and their families. They advise the courts on what is considered to be in the best interest of the child	Week of birth, gender, hearing date, hearing outcome type, maternal ALF
Children in Need (CIN) (referred to as CINW in SAIL) https://web.www.healthdatagateway.org/dataset/e4f6d9a8-88d0-4781-b192-cd165451b272)	Welsh Government	This discontinued census collected data on children and their families who were in need or provided with social services by the local authority (2009–2016)	Individual Record Number (IRN); is based on, and replaces the Unique Pupil Numbers (UPN) as a system ID across the education tables (EDUW) because UPNs would allow individuals to be identified as they are known in the real world
Children Looked After (CLA) (referred to as LACW in SAIL) (https://web.www.healthdatagateway.org/dataset/cfdafacb-48f0-4ad8-9f20-193a5eec2da4)	Welsh Government	An annual census that contains information relating to the child and their care episodes	Week of birth, gender, episode start date, legal status code, child's home Lower Layer Super Output Area (LSOA)
Children Receiving Care and Support (CRCS) (https://web.www.healthdatagateway.org/dataset/fb3fac03-a428-4b3f-8058-5622e1fd57d8)	Welsh Government	This superseded the Children in Need census, and is an annual census that includes information relating to children who have an eligible care and support plan	Individual Record Number (IRN)
Education Wales (EDUW) (https://web.www.healthdatagateway.org/dataset/204fa806-071d-4410-9c2c-143017d32d24)	Welsh Government	School and Pupil data covering all state-funded Welsh learning centres	Week of birth
National Community Child Health Database (NCCHD) (referred to as NCCH in SAIL) (https://web.www.healthdatagateway.org/dataset/20fe153c-a5e5-4991-900e-8fa9988e771a)	Digital Health and Care Wales	Includes information relating to birth registration, child health examinations and monitoring, and immunisations	Week of birth, gender, child's LSOA at birth, birth weights, birth times and maternal ALF to check whether children are twins
Welsh Demographic Service Dataset (WDS) (referred to as WDSD in SAIL) (https://web.www.healthdatagateway.org/dataset/8a8a5e90-b0c6-4839-bcd2-c69e6e8dca6d)	Digital Health and Care Wales	Contains information about individuals in Wales that use NHS services. Includes address and practice registration history	Week of birth, gender, child's home LSOA address, address start and end date
Welsh Longitudinal General Practice (WLGP) (https://web.www.healthdatagateway.org/dataset/33fc3ffd-aa4c-4a16-a32f-0c900aaea3d2)	Direct to SAIL	Covers 86% of General Practices in Wales, includes information relating to diagnoses, prescriptions and process of care codes	Week of birth, gender, event date, event LSOA

Continued

Table 1: Continued

Dataset	Data provider	Description	Variables
Welsh Index of Multiple Deprivation (WIMD) 2014		Welsh Government's measure of deprivation for small areas in Wales	WIMD is derived based on the LSOA. It ranks each LSOA from 1 (most deprived) to 1909 (least deprived), roughly grouping rankings into equal quintiles. It captures deprivation based on 8 domains

Abbreviations: ALF: Anonymised Linkage Field, LSOA: Lower Layer Super Output Area, WIMD: Welsh Index of Multiple Deprivation.

Matching algorithm procedure

Here, we provide an overview of the six-step CLA matching algorithm (summarised in Figure 1). Our linkage algorithm is an iterative process where each child proceeds through every step, providing in some cases that they meet the criteria to be matched in a given step, only stopping when they have been successfully assigned an ALF or have reached the end of the process.

1) The first step depended on the dates a child was recorded in CLA. If the child was in CLA pre-2016, we linked to CINW using the system ID and local authority code. If the child had an ALF in CINW, this was their ALF in CLA. If the child was in CLA post-2016, we linked to CRCS using the system ID and local authority code. If the child had an ALF in CRCS, this was their ALF in CLA.

2) Next, we required that a child had an IRN recorded in CLA. If they did, we used the IRN, week of birth, and gender code to link to EDUW. If the child had an ALF in EDUW, this became their ALF in CLA.

3) For the children unmatched by the first two steps, we continued the process by linking CLA to Cafcass Wales (CAFW) on the week of birth, gender, episode start date (CLA) and hearing date (CAFW), and a legal status code (CLA) and hearing outcome type (CAFW). If there was a fully matching CAFW record with an ALF, we considered it their ALF. This stage uses the "improved" CAFW match as described in one of the pre-processing tasks.

4) Next, we required that a child's first entry into CLA occurs under six months of age and that their first entry in CLA had a home LSOA recorded. We link CLA to NCCHD births using the week of birth, gender, child home LSOA (CLA) and LSOA of birth (NCCHD). If a single record in NCCHD matched all of these fields, then that ALF was taken to be the child ALF in CLA. This stage operates on the assumption that it is unlikely that the mother moved between giving birth and the child being taken into care if the child was under six months of age.

5) For the children who remain unmatched, we used the WLGP events table to filter for events with care event-related read codes and then linked to CLA, looking for a complete match on the week of birth, gender, child home or placement LSOA (CLA) and event LSOA (WLGP), and episode start/end date (CLA) and event date (WLGP). This step was carried out for every placement recorded for a child in CLA. If a single ALF matched the above fields for two or more placements of a child in CLA, this was taken to be their ALF.

6) Finally, using WDSD to examine GP registration dates, we looked for a match on the week of birth, gender, child home and placement LSOA (LACW) and address LSOA (WDSD), placement start/end date and address start/end date (WDSD). As in step 5, this was done for every placement recorded for a child in CLA. If a single ALF matched all of these fields for three or more placements of a child in LACW, this was taken to be their ALF.

7) Any remaining children were considered not to have an ALF.

Implementation and testing

The algorithm was implemented in Python 3.9.15 [27]. Where it was possible to link records between datasets, Python's Record Linkage package was used [28]. Following implementation, ibm-db and ibm-db-sa [29, 30] commands were then used to provide Python an interface to connect to IBM D2B database [31] and adapt data including preparing and issuing SQL statements.

Benchmarking

We conducted a benchmarking exercise to have a point of comparison for the performance of our algorithm (sensitivity). Since the algorithm utilised by SAIL had access to identifiable data, it is likely that our algorithm is indicative of correct identifiers. To benchmark our algorithm, we ran individuals with an ALF assigned via the routine SAIL processes (37% of children in CLA) through our algorithm to examine whether our algorithm was able to replicate ALFs. The routine algorithm used to assign ALFs within SAIL has produced specificity and sensitivity values of 99.8% and 94.6%, respectively in health datasets, and a sensitivity of 95.2% in social care datasets [19].

Results

CLA data is currently available from 2002/03 to 2020/21 and can be accessed by researchers through the SAIL Databank. During this period, we identified 38,260 individuals with a record in the CLA dataset, with 14,024 (37%) having been allocated an ALF via the standard SAIL algorithm (Supplementary Appendix 2).

Following the implementation of our algorithm, the overall match rate increased by 25%, with a total of 23,409 (61%) individuals having an ALF (via our algorithm and the routine SAIL algorithm). Figure 1 shows the number of 'new' ALFs assigned at each step of our developed algorithm. A total of 9,842 individuals were assigned an ALF via the algorithm. Note, an inidividual could be assigned an ALF for one or more care episodes where their system ID had changed if they had moved local authorities. This is why the unique number of ALFs assigned via the alogorithm, as per Supplementary Appendix 2 (n = 9,385), is lower.

The algorithm was also able to correct data quality issues. We also report the cohort demographics for the entire CLA cohort at the time of first entry into CLA before (routine ALF) and after (boosted ALF) applying the algorithm (Supplementary Appendix 2).

We were able to benchmark our algorithm by comparing routine ALFs assigned via the SAIL process to those assigned by our algorithm. Our algorithm demonstrated a high sensitivity rate of 90% when compared to the ALFs already assigned in SAIL (13,253/14,725)[5]. Only a small proportion were 'incorrect' or mis-matched (1.2%, 170/14,725). The algorithm was unable to assign the remaining 8.8% of children an ALF (1302/14,725). We were unable to calculate specificity because we do not have access to the de-anonymised data, and therefore we do not have a set of children that we can confidently say who should not be linked/recorded within the CLA dataset.

Individuals who were born before 1994 had the fewest pre- and post-match ALFs allocated, likely due to fewer datasets being available (e.g. CINW in 2009/10, EDUW in 2002/2003 (academic year)) and poorer data coverage. For example, although the WLGP data pre-dates 2000, the transition to electronic records was not widespread across practices until the late 1990s [32], and records are known to be less complete before 1 January 2000 [33].

As expected, there were also more children under the age of four who did not have an ALF when they entered care for the first time, having the most 'unlinked' ALFs (Figure 2). As previously alluded to, this is likely due to the UPN being assigned at entry into school but could also be explained by the lack of placement stability (poor recording of postcode) and continuity in the healthcare system that children looked after experience. The recognition of children looked after as a highly mobile population, together with limited experience of continued general practice care may contribute to the reduction in attributable ALFs [34]. However, the former explanation appears to be most probable, supported by comparable improvements in ALF percentages amongst those less than and greater than four years of age. There was also a high proportion of children aged 15+ years who remained 'unlinked'. One plausible explanation could be because these children have 'aged out' of care and/or left education, and therefore, they have no further contact with these sectors.

There appeared to be no difference in ALF improvement by sex; males and females both had around a 25% increase. Again, deprivation did not appear to impact ALF assignment, although there were fewer ALFs matched for individuals who were missing a WIMD score (Figure 3). This is to be expected and reflects the lack of a recorded postcode, an identifier used in the ALF linkage process. In line with previous studies, variation in data completeness was more prominent for ethnicity [35, 36]. Specifically, ethnic minority groups, with the exception of 'Mixed' ethnic group, were less likely to be successfully matched (Figure 4) [37, 38]. It has been speculated that this reflects inconsistencies in the quality and distinctiveness of registered names [39, 40]. Since those with asylum seeker status (a person who has left their own country and is seeking protection in another country) are likely to be born outside of the UK, this may also impact linkage success (e.g. lack of NHS number, postcode history).

[5]Note, this is higher than the unique number of ALFs (n=14,024) described in Supplementary Appendix 2 because a child may have several system IDs from where they have moved local authority

Discussion

Our CLA ALF improvement algorithm increased ALF matching from 37% to 61%, allowing an additional 9,385 (25%) individuals to be assigned an ALF. The algorithm also amended incorrect weeks of birth and gender, and incorrect/duplicate identifiers. Benchmarking against the SAIL ALF assignment algorithm demonstrated high levels of sensitivity (90%) for our CLA ALF algorithm.

Strengths

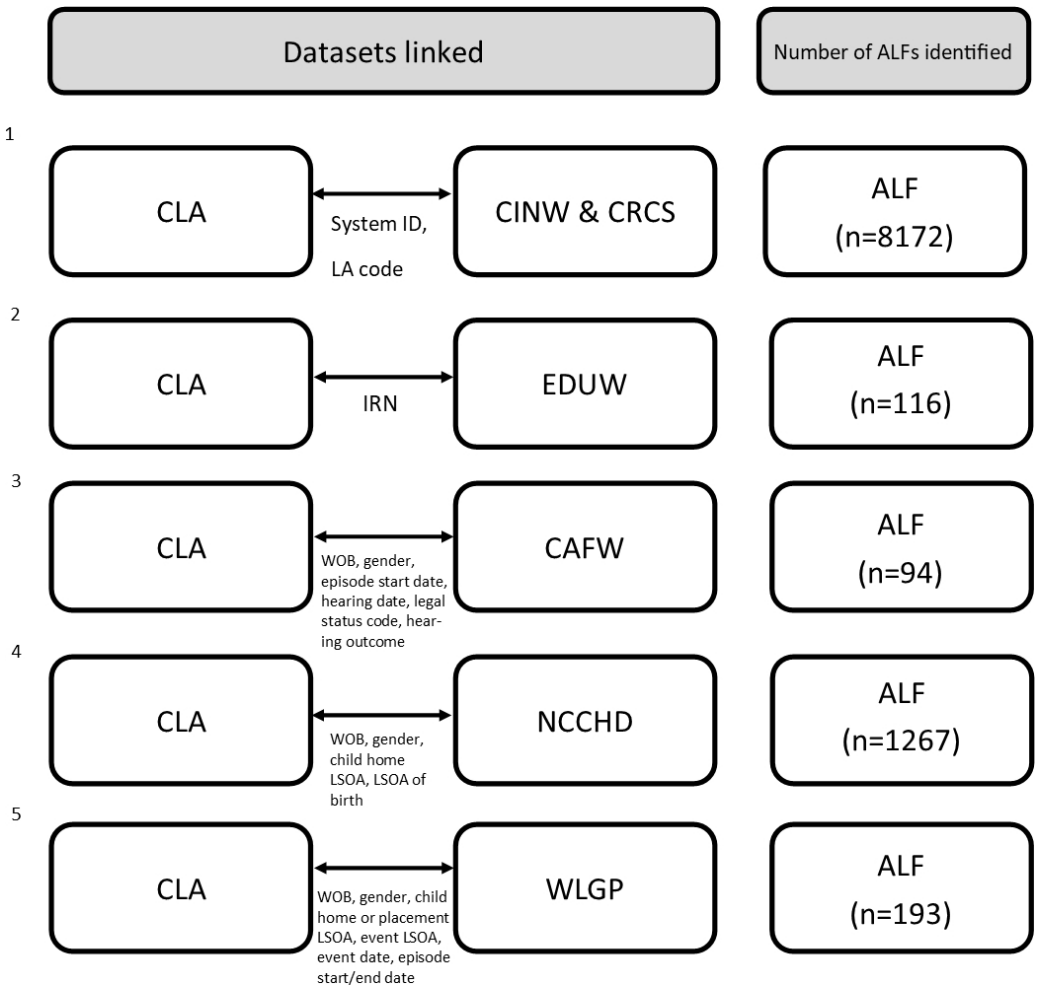
Our algorithm led to a large improvement in the number of individuals assigned an ALF within the CLA dataset. An additional 25% of identifiers were matched, and other known data errors such as incorrect weeks of birth, duplicate and inconsistent ALFs were corrected. Through improvement of ALF matching, we hope to reduce some bias that occurs with missing data, particularly ethnic and deprivation bias, demographics that may be associated with children in care. Our algorithm also demonstrated high levels of sensitivity (90%) when benchmarked against the routine algorithm used by SAIL to assign ALFs. This algorithm provides an underlying foundation for further development, particularly as SAIL continues to obtain new data sources. Other datasets may also benefit from our algorithm, and future work should explore its application.

Limitations

Potential biases in ALF availability

Whilst the algorithm improves the number of ALFs allocated, it is important to note that although a secondary outcome of the algorithm was to reduce bias in the CLA population by improving ALF matching, it was not possible to completely alleviate. A successful ALF match is dependent on an individual being present in the dataset through interactions

Figure 1: Data linkage of education, social and healthcare data sources used to identify ALFs



Abbreviations: ALF: Anonymised Linkage Field; CAFW: Children and Family Court Advisory and Support Service (Cafcass) Wales; CINW: Children in Need Wales; CLA: Children looked after; CRCS: Children Receving Care and Support; EDUW: Education Wales; IRN: Individual Record Number; LA: Local authority; LSOA: Lower Layer Super Output Area; NCCHD: National Community Child Health Database; WLGP: Welsh Longitudinal General Practice; WOB: Week of birth

with services. For example, for an individual to be recorded in the WDSD, NCCHD or WLGP, an interaction with a Welsh NHS service must take place. This may not be the case, for example, for infants who are born in England and move into Wales, thus it would not be possible to match an ALF via this method. In addition, children looked after lack stability, which may contribute to limited continuity in primary care and residential addresses. Further, by capturing ALFs using the CAFW dataset we are less likely to identify children who enter care via voluntary arrangements because Cafcass is not involved in these cases [41]. This is potentially problematic as approximately two-thirds of children in Wales initially enter care by voluntary arrangements [42]. Researchers should take these issues into consideration and aim to address any potential biases these may introduce.

Impact of adoption on an ALF

One known limitation of the CLA data is that it does not contain a child's information post-adoption. When an adoption order is granted, an individual's identifier changes because they are assigned a new NHS number and unique identifier within the local authority. It is therefore not possible to link pre- and post-adoption records [14]. This may explain the lower rates of ALF matches in younger children less than four years of age, especially as the average age for adoption is three years old in Wales [43].

Assessing the sensitivity and benchmarking

We were only able to examine the sensitivity (90%) of our ALF matching algorithm in comparison against previously published record linkage methods (routine (pre-assigned) ALFs by SAIL) [19]. As the algorithm used de-identified quasi-identifiers and not the original data, it was not possible to calculate specificity or check for sources of error. Whilst the algorithm was highly sensitive, researchers should be aware that a small proportion of ALFs assigned are potentially erroneous (1.2%). Also, it is important to note that comparing our algorithm to the routine algorithm has its own limitations. First, any incorrect matches are constrained to both algorithms. Second, because we were only able to benchmark against routine (pre-assigned) ALFs this may introduce bias issues, whereby different populations

Figure 2: Percentage of individuals with an Anonymised Linkage Field for children looked after pre- (routine) and post-matching (boosted) algorithm, by age at first entry into care

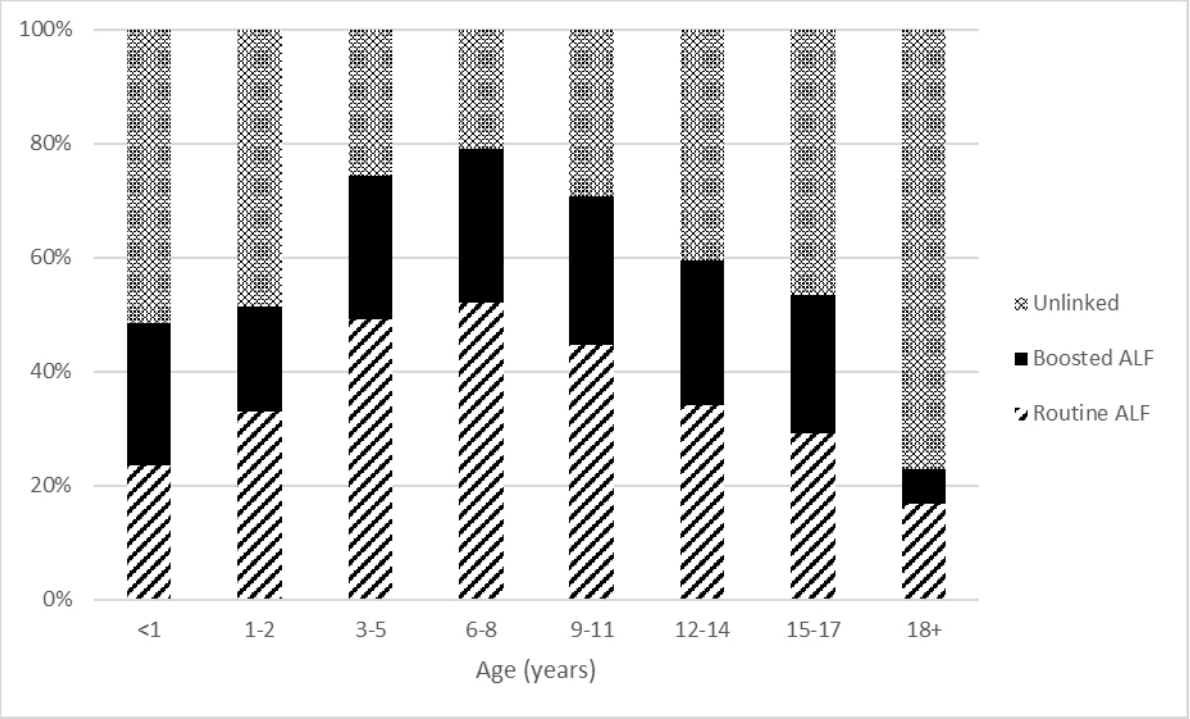
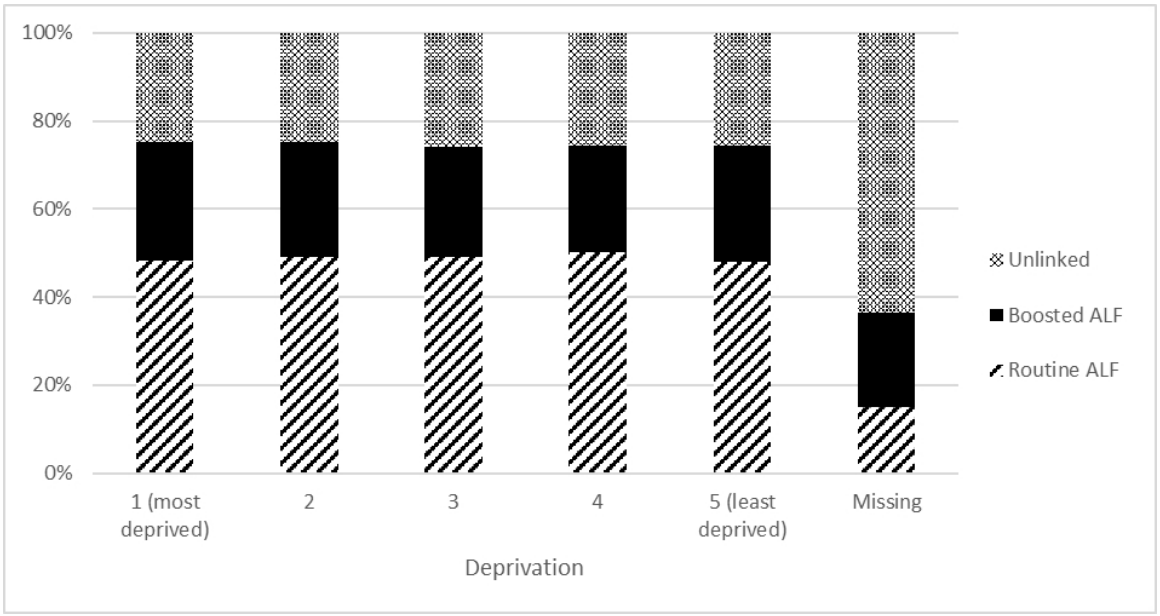


Figure 3: Percentage of individuals with an Anonymised Linkage Field for children looked after pre- (routine) and post-matching (boosted) algorithm, by Welsh Index of Multiple Deprivation (WIMD) based on the home postcode at time of first entry into care

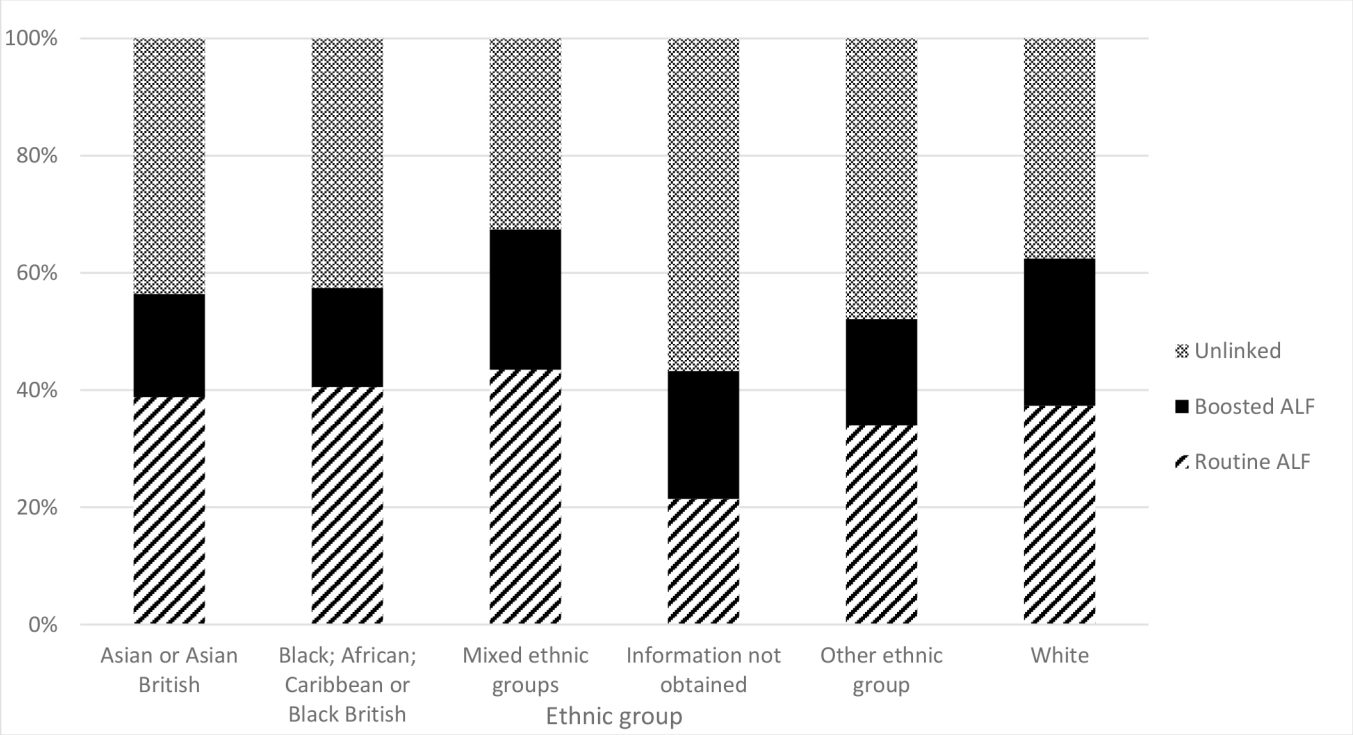


are not equally assigned an ALF. As a result, this may accentuate disparities in social care. Lastly, we cannot directly compare the algorithms because of the influence of data quality; the algorithms use different variables and data sources to assign an ALF, which is likely to affect the matching process.

Future work should aim to validate the algorithm against identifiable data to determine the accuracy of the linkage, or alternatively, to benchmark using synthetic data. Researchers need to consider whether to adjust for missing data amongst vulnerable groups such as those with asylum seeker status, or ethnic minority groups because of missed linkage. This is particularly relevant in this field of research, as those with asylum seeker status represent around 7% of all children looked after in England and figures continue to rise across the UK [44]. Future work should involve collaborations with each of the UK home nations Governments and local authorities to explore how match rates could be improved.

In this work, we have limited our improved ALF matching algorithm to the CLA dataset because of the known poor ALF match rate. It may be possible that with further refinement,

Figure 4: Percentage of individuals with an Anonymised Linkage Field for children looked after pre- (routine) and post-matching (boosted) algorithm, by ethnic group at first entry into care



improved match benefits can also be made to other datasets that contain cohorts of school-aged children without NHS numbers e.g. the Student Health and Wellbeing Survey Dataset.

Conclusion

Our work highlights the benefits of our CLA ALF improvement algorithm, resulting in the potential to explore a wide range of research questions through its linkage to other datasets. We characterise our algorithms' performance by benchmarking against the standard ALF matching algorithm utilised by SAIL, which shows promising results. Our benchmarked algorithm produced high rates of sensitivity (90%) and importantly, greatly improved the ALF match rate from 37% to 61%. Additionally, this algorithm has the potential to overcome some of the bias that researchers may experience due to missing data. However, researchers should be aware of some biases and missing data that may remain. We demonstrate a standardised and reliable matching tool to facilitate the improved allocation of an ALF in the CLA dataset, which has been implemented as part of the CLA dataset. Development and future benchmarking efforts may further refine our algorithm, allowing application to other linkable datasets.

Acknowledgements

This study makes use of anonymised data held in the SAIL system, which is part of the national e-health records research infrastructure for Wales. We would like to acknowledge all the data providers who make anonymised data available for research. We would like to thank our colleagues at Welsh Government for the guidance they have provided.

Statements of conflicts of interest

The authors report no conflicts of interest.

Funding

This work was supported by Health and Care Research Wales and Administrative Data Research (ADR) Wales. LJG is a member of the Children's Social Care Research and Development Centre (CASCADE) partnership, which receives infrastructure funding from Health and Care Research Wales (HCRW) (517199). LEC is a research fellow, funded by Health and Care Research Wales (SCF-22-07).

Ethics statement

In accordance with Health Research Authority guidance, ethical approval is not mandatory for studies using only anonymized data. The project proposal was reviewed by the SAIL independent Information Governance Review Panel, experts in information governance, and members of the public approved this study (reference: 1533).

Data availability statement

Data may be obtained from a third party and are not publicly available. The data used in this study is available

from the Secure Anonymised Information Linkage (SAIL) Databank at the Health Information Research Unit (HIRU) Swansea University, Swansea, UK. All proposals to use SAIL datasets must comply with HIRU's information governance policy and are subject to review by an independent Information Governance Review Panel (IGRP). Before data can be accessed, approval must be given by the IGRP. Requests to access these datasets should be directed to: www.saildatabank.com/application-process.

References

1. NSPCC Learning. Looked after children [Internet]. 2024. Available from: https://learning.nspcc.org.uk/children-and-families-at-risk/looked-after-children#skip-to-content

2. McGhee J, Bunting L, McCartan C, Elliott M, Bywaters P, Featherstone B. Looking after children in the UK—convergence or divergence? Br J Soc Work. 2018;48(5):1176–98. https://doi.org/10.1093/bjsw/bcx103

3. NSPCC. Statistics Briefing: Children in Care. NSPCC Learn. 2024;1–26.

4. Simkiss D. Outcomes for looked after children and young people. Paediatr Child Heal (United Kingdom) [Internet]. 2012;22(9):388–92. https://doi.org/10.1016/j.paed.2012.05.004

5. McAuley C, Davis T. Emotional well-being and mental health of looked after children in England. Child Fam Soc Work. 2009;14(2):147–55. https://doi.org/10.1111/j.1365-2206.2009.00619.x

6. Luke N, O'Higgins A. Is the Care System to Blame for the Poor Educational Outcomes of Children Looked After? Evidence from a Systematic Review and National Database Analysis. Child Aust [Internet]. 2018 Jun 1;43(2):135–51. https://doi.org/10.1017/cha.2018.22

7. Connelly R, Playford CJ, Gayle V, Dibben C. The role of administrative data in the big data revolution in social science research. Soc Sci Res [Internet]. 2016;59:1–12. Available from: https://doi.org/10.1016/j.ssresearch.2016.04.015.

8. Bailey GA, Lee A, Bedford H, Perry M, Holland S, Walton S, et al. Immunisation status of children receiving care and support in Wales: a national data linkage study. Front Public Heal [Internet]. 2023 [cited 2023 Oct 5];11. Available from: /pmc/articles/PMC10423803/https://doi.org/10.3389/FPUBH.2023.1231264/FULL.

9. Doebler S, Broadhurst K, Alrouh B, Cusworth L, Griffiths L. Born into care: Associations between area-level deprivation and the rates of children entering care proceedings in Wales. Child Youth Serv Rev. 2022 Oct 1;141:106595. https://doi.org/10.1016/j.childyouth.2022.106595

10. Department of Health & Social Care. Data saves lives: reshaping health and social care with data - GOV.UK [Internet]. UK Government. 2022. Available from: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data#improving-data-for-adult-social-care

11. Welsh Government. Programme for government 2021 to 2026: update. 2021;(June). Available from: https://gov.wales/programme-for-government-2021-to-2026-html.

12. Mc Grath-Lone L, Jay MA, Blackburn R, Gordon E, Zylbersztejn A, Wiljaars L, et al. What makes administrative data research-ready? Int J Popul Data Sci [Internet]. 2022 Apr 27;7(1):12–7. https://doi.org/10.23889/ijpds.v7i1.1718

13. Holmes L. Use of children's social care data at the local and regional area level. Nuff Fam Justice Obs [Internet]. 2019; Available from: https://www.nuffieldfjo.org.uk/wp-content/uploads/2021/05/Use-of-childrens-social-care-data_final.pdf.

14. Allnatt G, Elliott M, Cowley L, Lee A, North L, Griffiths L. Data Explained: Children Looked After datasets. ADR Wales [Internet]. 2023; Available from: https://adrwales.org/wp-content/uploads/2023/06/Data-Explained-CLA-datasets_final.pdf.

15. Soraghan J, Raab G. Data Explained: Scottish Government's Looked After Children Longitudinal Dataset. ADR Scotl [Internet]. 2023;(April). Available from: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Data_Explained/Data_Explained_Scottish_Government_Looked_After_Children_Longitudinal_Dataset_April_2023.pdf.

16. Jay MA, McGrath-Lone L, Gilbert R. Data Resource: the National Pupil Database (NPD). Int J Popul data Sci [Internet]. 2019;4(1):1101. https://doi.org/10.23889/ijpds.v4i1.1101

17. Mckenna AS, Maguire A, Bunting L, Gleghorne N, Reilly DO. Data Explained: Social Services Client Administration and Retrieval Environment (SOSCARE). ADR Nothern Irel [Internet]. 2024; Available from: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Data_Explained/ADR_UK_Data_Explained_SOSCARE.pdf.

18. Mc Grath-Lone L, Harron K, Dearden L, Nasim B, Gilbert R. Data Resource Profile: Children Looked After Return (CLA). Int J Epidemiol. 2016;45(3):716-717F. https://doi.org/10.1093/ije/dyw117

19. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford D V., et al. The SAIL databank: Linking multiple health and social care datasets. BMC Med Inform Decis Mak [Internet]. 2009 [cited 2021 Jan 20];9(1). https://doi.org/10.1186/1472-6947-9-3

20. Christen P, Ranbaduge T, Schnell R. Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing. Link Sensitive Data Methods Tech Pract Privacy-Preserving Inf Shar. 2020;1–468. https://doi.org/10.1007/978-3-030-59706-1

21. Allnatt G, Lee A, Scourfield J, Elliott M, Broadhurst K, Griffiths L. Data resource profile: children looked after administrative records in Wales. Int J Popul data Sci. 2022;7(1):1752. https://doi.org/10.23889/ijpds.v7i1.1752

22. Alrouh B, Broadhurst K, Cusworth L, Griffiths L, Johnson RD, Akbari A, et al. Born into care: newborns and infants in care proceedings in Wales [Internet]. Nuffield Family Justice Observatory. 2019. Available from: https://www.nuffieldfjo.org.uk/wp-content/uploads/2021/05/Born-into-care-Wales-main-report_English_final_web.pdf.

23. Ford D V., Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res. 2009 Dec 4;9(1):157. https://doi.org/10.1186/1472-6963-9-157

24. Welsh Government. Children looked after census 2021-2022 [Internet]. 2022. Available from: https://www.gov.wales/sites/default/files/statistics-and-research/2022-05/looked-after-children-census-2021-22-guidance.pdf.

25. Wales S for. Welsh Index of Multiple Deprivation 2019 (WIMD 2019): guidance on use [Internet]. 2019 [cited 2021 Jan 21]. Available from: https://gov.wales/sites/default/files/statistics-and-research/2019-11/welsh-index-of-multiple-deprivation-wimd-2019-guidance-on-use.pdf.

26. UK Gov. UK Read Code - data.gov.uk [Internet]. NHS Digital. 2015 [cited 2023 Oct 5]. Available from: https://data.gov.uk/dataset/f262aa32-9c4e-44f1-99eb-4900deada7a4/uk-read-code.

27. Python Foundation. About PythonTM – Python.org [Internet]. About Python. 2016 [cited 2023 Oct 5]. p. 1. Available from: https://www.python.org/about/.

28. de Bruin J. About — Python Record Linkage Toolkit 0.15 documentation [Internet]. [cited 2023 Oct 5]. Available from: https://recordlinkage.readthedocs.io/en/latest/about.html.

29. Application development in Python with ibm_db – IBM Documentation [Internet]. 2022 [cited 2023 Oct 5]. Available from: https://www.ibm.com/docs/en/db2/11.5?topic=framework-application-development-db.

30. ibm-db-sa – PyPI [Internet]. 2023 [cited 2023 Oct 5]. Available from: https://pypi.org/project/ibm-db-sa/.

31. IBM. IBM DB2 Database [Internet]. Https://Www.Ibm.Com/Br-Pt/Products/Db2-Database. 2024. Available from: https://www.ibm.com/products/db2-database.

32. McMillan B, Eastham R, Brown B, Fitton R, Dickinson D. Primary care patient records in the United Kingdom: Past, present, and future research priorities. J Med Internet Res. 2018;20(12):1–7. https://doi.org/10.2196/11293.

33. Torabi F, Harris DE, Bodger O, Akbari A, Lyons RA, Gravenor M, et al. Identifying unmet antithrombotic therapeutic need, and implications for stroke and systemic embolism in atrial fibrillation patients: a population-scale longitudinal study. Eur Hear J Open. 2022;2(6):1–12. https://doi.org/10.1093/ehjopen/oeac066

34. Ridley J, Larkins C, Farrelly N, Hussein S, Austerberry H, Manthorpe J, et al. Investing in the relationship: Practitioners' relationships with looked-after children and care leavers in Social Work Practices. Child Fam Soc Work. 2016;21(1):55–64. https://doi.org/10.1111/cfs.12109

35. Teece L, Gray LJ, Melbourne C, Orton C, Ford D V., Martin CA, et al. United Kingdom Research study into Ethnicity and COVID-19 outcomes in Healthcare workers (UK-REACH): A retrospective cohort study using linked routinely collected data, study protocol. BMJ Open. 2021;11(6). https://doi.org/10.1136/bmjopen-2020-046392

36. Møller H, Henson K, Lüchtenborg M, Broggio J, Charman J, Coupland VH, et al. Short-term breast cancer survival in relation to ethnicity, stage, grade and receptor status: National cohort study in England. Br J Cancer. 2016;115(11):1408–15. https://doi.org/10.1038/bjc.2016.335

37. Libuy N, Harron K, Gilbert R, Caulton R, Cameron E, Blackburn R. Linking education and hospital data in England: Linkage process and quality. Int J Popul Data Sci. 2021;6(1). https://doi.org/10.23889/ijpds.v6i1.1671

38. Hagger-johnson G. Probabilistic linking to enhance deterministic algorithms and reduce linkage errors in hospital administrative data. 2018;24(2). https://doi.org/10.14236/jhi.v24i2.891

39. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699–710. https://doi.org/10.1093/IJE/DYX177

40. Aspinall PJ, Jacobson B. Why poor quality of ethnicity data should not preclude its use for identifying disparities in health and healthcare. Qual Saf Heal Care. 2007;16(3):176–80. https://doi.org/10.1136/qshc.2006.019059

41. Johnson RD, Ford D V., Broadhurst K, Cusworth L, Jones KH, Akbari A, et al. Data Resource: Population level

family justice administrative data with opportunities for data linkage. Int J Popul Data Sci. 2020 Jan;5(1):1339. https://doi.org/10.23889/ijpds.v5i1.1339

42. Elliott M. The backgrounds of children entering public care in Wales [Internet]. 2017. Available from: https://orca.cardiff.ac.uk/id/eprint/125393/1/PhD_Summary_Martin Elliott.pdf

43. StatsWales. Average age, in months, at adoption of looked after children during year ending 31 March by local authority and year [Internet]. Welsh Government. 2022 [cited 2023 Aug 25]. Available from: https://statswales.gov.wales/Catalogue/Health-and-Social-Care/Social-Services/Childrens-Services/Children-Looked-After/Adoptions/averageageatadoptionoflookedafterchildrenduringyearending31march-by-localauthority-year

44. Department for Education. Children looked after in England including adoptions [Internet]. 2022 [cited 2023 Aug 25]. Available from: https://explore-education-statistics.service.gov.uk/find-statistics/children-looked-after-in-england-including-adoptions/2021#releaseHeadlines-charts

45. McKenna S, O'Reilly D, Maguire A. The mental health of all children in contact with social services: a population-wide record-linkage study in Northern Ireland. Epidemiol Psychiatr Sci. 2023;32. https://doi.org/10.1017/S2045796023000276

Supplementary Appendices

Supplementary Appendix 1. The allocation of an Anonymised Linkage Field (ALF)

Demographic data comprising first name, surname, gender, date of birth, and postcode, are separated from content data such as clinical diagnostic testing and results data, or placement details of children looked after. Data is pseudonymised and allocated an ALF in the place of demographic data (by a trusted third party) where a successful match is made. The ALF is then transferred and joined to the personal data (i.e. placement details). Datasets provided to the SAIL Databank are already split at the source organisation into demographic and content data. An individual is allocated an ALF either using their unique 10-digit NHS number (deterministic record linkage, DRL), or a combination of other quasi-identifiers (probabilistic record linkage, PRL) held and maintained by the NHS Administrative Register (name, address (including historical), postcode, gender, date of birth, general practice of registration and NHS number). Within the SAIL Databank, records of individuals with an ALF can be linked to other data sources such as demographic, health, social and education datasets.

Supplementary Appendix 2: Cohort characteristics of individuals and match rate

		Total number of individuals	Routine (pre-algorithm) match rate (n,%)	Boosted (post-algorithm) match rate (n,%)	Percentage change between routine (pre-) and boosted (post-match) rate (%)
Total		38,260 (100)	14,024 (100)	23,409 (100)	24.5
Sex	Male	20,390 (53.3)	7375 (52.6)	12,405 (53)	24.7
	Female	17,870 (46.7)	6649 (47.4)	11,004 (47)	24.4
Age	<1	7842 (20.5)	1839 (13.1)	3797 (16.2)	25.0
	1-2	4568 (11.9)	1501 (10.7)	2348 (10.0)	18.5
	3-5	5357 (14.0)	2635 (18.8)	3976 (17.0)	25.0
	6-8	4752 (12.4)	2471 (17.6)	3755 (16.0)	27.0
	9-11	4457 (11.6)	1988 (14.2)	3147 (13.4)	26.0
	12-14	6186 (16.2)	2112 (15.1)	3684 (15.7)	25.4
	15-17	5032 (13.2)	1467 (10.5)	2687 (11.5)	24.2
	18+	66 (0.2)	11 (0.1)	15 (0.1)	6.1
Year of birth	pre-1988	1569 (4.1)	*	10 (0.0)	*
	1988	699 (1.8)	23* (0.2)	35 (0.1)	1.7
	1989	766 (2.0)	54 (0.4)	151 (0.6)	12.7
	1990	861 (2.3)	95 (0.7)	244 (1.0)	17.3
	1991	946 (2.5)	107 (0.8)	307 (1.3)	21.1
	1992	961 (2.5)	137 (1.0)	444 (1.9)	31.9
	1993	1042 (2.7)	134 (1.0)	590 (2.5)	43.8
	1994	1154 (3.0)	166 (1.2)	724 (3.1)	48.4
	1995	1133 (3.0)	173 (1.2)	767 (3.3)	52.4
	1996	1285 (3.4)	216 (1.5)	897 (3.8)	53.0
	1997	1323 (3.5)	248 (1.8)	941 (4.0)	52.4
	1998	1270 (3.3)	454 (3.2)	877 (3.7)	33.3
	1999	1341 (3.5)	610 (4.3)	927 (4.0)	23.6
	2000	1385 (3.6)	654 (4.7)	944 (4.0)	20.9
	2001	1431 (3.7)	670 (4.8)	933 (4.0)	18.4
	2002	1459 (3.8)	709 (5.1)	970 (4.1)	17.9
	2003	1493 (3.9)	768 (5.5)	1043 (4.5)	18.4
	2004	1447 (3.8)	772 (5.5)	1017 (4.3)	16.9
	2005	1403 (3.7)	734 (5.2)	1056 (4.5)	23.0
	2006	1366 (3.6)	743 (5.3)	1049 (4.5)	22.4
	2007	1300 (3.4)	672 (4.8)	1030 (4.4)	27.5
	2008	1298 (3.4)	691 (4.9)	1008 (4.3)	24.4
	2009	1321 (3.5)	678 (4.8)	998 (4.3)	24.2
	2010	1232 (3.2)	627 (4.5)	902 (3.9)	22.3
	2011	1193 (3.1)	646 (4.6)	894 (3.8)	20.8
	2012	1237 (3.2)	665 (4.7)	878 (3.8)	17.2
	2013	1117 (2.9)	597 (4.3)	783 (3.3)	16.7
	2014	990 (2.6)	519 (3.7)	677 (2.9)	16.0
	2015	892 (2.3)	462 (3.3)	627 (2.7)	18.5
	2016	884 (2.3)	383 (2.7)	559 (2.4)	19.9
	2017	771 (2.0)	175 (1.2)	325 (1.4)	19.5
	2018	716 (1.9)	176 (1.3)	324 (1.4)	20.7
	2019	534 (1.4)	151 (1.1)	252 (1.1)	18.9
	2020	389 (1.0)	101 (0.7)	197 (0.8)	24.7
	2021	52 (0.1)	14 (0.1)	29 (0.1)	28.8
WIMD 2019	1 (most deprived)	11,517 (30.1)	5554 (39.6)	8674 (37.1)	27.1
	2	5827 (15.2)	2856 (20.4)	4385 (18.7)	26.2
	3	3646 (9.5)	1787 (12.7)	2700 (11.5)	25.0
	4	2310 (6.0)	1160 (8.3)	1717 (7.3)	24.1
	5 (least deprived)	1301 (3.4)	626 (4.5)	966 (4.1)	26.1
	Missing	13,659 (35.7)	2041 (14.6)	4967 (21.2)	21.4

Continued

Supplementary Appendix 2: Continued

		Total number of individuals	Routine (pre-algorithm) match rate (n,%)	Boosted (post-algorithm) match rate (n,%)	Percentage change between routine (pre-) and boosted (post-match) rate (%)
Ethnicity	Asian or Asian British	750 (2.0)	291 (2.1)	423 (1.8)	17.6
	Black; African; Caribbean or Black British	585 (1.5)	237 (1.7)	336 (1.4)	16.9
	Mixed ethnic groups	1095 (2.9)	476 (3.4)	738 (3.2)	23.9
	Information not obtained	2067 (5.4)	443 (3.2)	894 (3.8)	21.8
	Other ethnic group	706 (1.8)	240 (1.7)	368 (1.6)	18.1
	White	33,057 (86.4)	12,337 (88.0)	20650 (88.2)	25.1
Asylum seeking status	No	26,713 (69.8)	12,507 (89.2)	194 (0.8)	26.5
	Yes	539 (1.4)	124 (0.9)	19577 (83.6)	13.0
	Missing	11,008 (28.8)	1393 (9.9)	3638 (15.5)	20.4

*numbers combined to prevent disclosure of small numbers (<5).
Demographics are at the time of first CLA episode.
WIMD: Welsh Index of Multiple Deprivation.