International Journal of Population Data Science

Journal Website: www.ijpds.org

Generating synthetic identifiers to support development and evaluation of data linkage methods

Joseph Lam[1,*], Andy Boyd[2], Robin Linacre[3], Ruth Blackburn[1], and Katie Harron[1]

Submission History

Submitted:	24/01/2024
Accepted:	24/05/2024
Published:	01/07/2024

[1]Population, Policy & Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, London, United Kingdom
[2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[3]UK Ministry of Justice, London, United Kingdom

Abstract

Introduction
Careful development and evaluation of data linkage methods is limited by researcher access to personal identifiers. One solution is to generate synthetic identifiers, which do not pose equivalent privacy concerns, but can form a 'gold-standard' linkage algorithm training dataset. Such data could help inform choices about appropriate linkage strategies in different settings.

Objectives
We aimed to develop and demonstrate a framework for generating synthetic identifier datasets to support development and evaluation of data linkage methods. We evaluated whether replicating associations between attributes and identifiers improved the utility of the synthetic data for assessing linkage error.

Methods
We determined the steps required to generate synthetic identifiers that replicate the properties of real-world data collection. We then generated synthetic versions of a large UK cohort study (the Avon Longitudinal Study of Parents and Children; ALSPAC), according to the quality and completeness of identifiers recorded over several waves of the cohort. We evaluated the utility of the synthetic identifier data in terms of assessing linkage quality (false matches and missed matches).

Results
Comparing data from two collection points in ALSPAC, we found within-person disagreement in identifiers (differences in recording due to both natural change and non-valid entries) in 18% of surnames and 12% of forenames. Rates of disagreement varied by maternal age and ethnic group. Synthetic data provided accurate estimates of linkage quality metrics compared with the original data (within 0.13-0.55% for missed matches and 0.00-0.04% for false matches). Incorporating associations between identifier errors and maternal age/ethnicity improved synthetic data utility.

Conclusions
We show that replicating dependencies between attribute values (e.g. ethnicity), values of identifiers (e.g. name), identifier disagreements (e.g. missing values, errors or changes over time), and their patterns and distribution structure enables generation of realistic synthetic data that can be used for robust evaluation of linkage methods.

Keywords
record linkage, data linkage, synthetic data, synthetic identifiers, linkage evaluation, ALSPAC

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

https://doi.org/10.23889/ijpds.v9i1.2389

Introduction

Data linkage facilitates the combination of detailed information on individuals captured in disparate data sources, without the need for new data collection. Linkage is increasingly used as an efficient approach, particularly with existing administrative datasets, and has great potential for social good. Access to identifiable information is crucial when linking multiple datasets, as linkage depends on either the availability of unique identifiers (e.g. a social security number) or a set of individually non-unique variables such as name, sex and date of birth, which in combination, can identify an individual. This is the case for both linkage using identifiers in their natural form or for privacy preserving techniques which mask the identifiers in some form. The level of completeness, uniqueness and accuracy of identifiers recorded in administrative data pose a challenge for linkage, particularly when linking across multiple sectors in countries where unique citizen identifiers are unavailable [1]. Careful development and evaluation of linkage methods is therefore required in order to achieve high quality linkage and robust results [2, 3]. However, methodological development has been constrained by confidentiality concerns and legislative restrictions governing access to personal information for research.

In practice, access to identifiers is usually limited to either the data owners or trusted third parties, who may be unwilling or unable to make use of these identifiers for methodological purposes. Conversely, analysts will typically only have access to the de-identified linked data, with limited information about any uncertainty in linkage, or information with which to assess the quality of linkage [4]. This separation limits opportunities for the development of advanced linkage methods and the assessment of computational performance and/or linkage quality, as the researchers cannot access the identifiable data needed for these evaluations [5]. Even when it is possible to access identifiers, lack of a "ground truth" makes it difficult to evaluate different linkage strategies, as there is no gold standard against which results can be compared.

One solution to this problem is to generate synthetic datasets of identifiers that mimic the characteristics of real identifiers (and so can be used for methodological work) but that do not pose any confidentiality issues [6, 7]. Such synthetic datasets of identifiers would include a "ground truth" to enable evaluation of different linkage methods. Synthetic data generators have been developed in the context of providing realistic *research* datasets, where the aim is to mimic the underlying statistical properties of the original data whilst minimising disclosure risk [8, 9]. A summary of existing synthetic data generation methods is described by Kokosi [9]. Synthetic data that retain the relationships between variables in the original data can provide an accurate representation of the original data and be used for a range of purposes, including evaluation of different methodological approaches [10]. However, these approaches have mainly been developed in the context of 'attribute' data, i.e. variables typically used within an analysis (e.g. social or health status, occupation). In the context of developing linkage methods, we are concerned with 'identifier' data, i.e. variables used for linkage but not necessarily for analysis (e.g. postcode, name). In some cases, there is overlap between the two: date/year of birth and sex can be both attribute variables and personal identifiers.

In most applications of synthetic data, retaining the relationships between different variables helps to replicate the underlying structure of the data and enables users to test and evaluate different methodological approaches. When generating synthetic identifier data, there is also a need to ensure that the data retain dependencies between variables (for example, name might be associated with date of birth). However, there are a number of reasons why an alternative approach to generating synthetic data is required to address the idiosyncrasy of identifiers. Firstly, identifier variables do not always follow standard statistical distributions. Secondly, identifiers are affected by specific types of recording errors and changes that occur within and between datasets, and over time. We refer to these disagreements as 'errors', whilst recognising that in some cases these will be genuine changes (e.g. address change due to migration or surname change following marriage) rather than errors in recording. Such errors are often related to attribute variables (e.g. names may be more often misspelt for particular ethnic groups; address changes are associated with age and changes in socio-economic and potentially health status). There may also be interdependencies between identifier errors, e.g. if name and address change at the same time due to divorce. Accurate replication of identifier errors and their dependencies on attribute variables is important, since these dependencies are directly related to the impact that linkage errors have on analysis [11]. Therefore, these errors and dependencies should be replicated within any synthetic identifier datasets that are used to test linkage methods, so that an assessment of bias resulting from linkage can be conducted [12]. Existing datasets generated to facilitate the development and testing of data linkage algorithms have typically not focussed on preserving these dependencies, and have not been evaluated in terms of their utility for testing linkage algorithms. There is therefore a scientific requirement to develop more robust and realistic synthetic identifier datasets [13].

This paper presents a framework for generating synthetic identifier data that could be used by data owners to enable researchers to develop and test linkage methods in different settings. Use of these data could help overcome the limited capacity for linkage methodology development by providing wider access to realistic identifier data, without disclosure risk, and with a ground truth against which linkage quality can be assessed. In Section 1, we describe a motivating scenario and outline how the steps needed to generate synthetic data can be implemented. Importantly, we consider the need to preserve the dependencies between identifier values, identifier errors, and attribute values. In Section 2, we evaluate the use of synthetic data for assessing linkage quality, based on an exemplar of longitudinal linkage within a large UK cohort study (the Avon Longitudinal Study of Parents and Children; ALSPAC)

Section 1: A framework for generating synthetic identifier data

Motivating scenario

Our motivating scenario is one in which we aim to conduct performance comparisons between different linkage

approaches, in a secure manner with low ethico-legal barriers and no intrusion into personal privacy. The aim of such performance comparisons is to optimise linkage algorithms that would then be applied to specific, real-world linkage projects. We assume that those commissioning the linkage (e.g. a researcher) will not have access to identifiers and that the linkage will be conducted by a trusted third party or data owner. We will examine the utility of synthetic datasets to conduct linkage performance comparisons that would be sufficiently similar to the real data whilst not intrusive of personal privacy. Such datasets could be useful for the development of linkage methods in two settings: 1) by data owners who have access to identifiers but where there are restrictions around using these identifiers for methodological development rather than business-as-usual linkage; 2) by researchers or research infrastructure providers (such as Trusted Research Environments) who cannot access identifiers but for whom synthetic data would be useful for understanding the implications of different linkage methods on their outputs.

Types of variables

First, we distinguish between two types of variables: identifiers and attributes.

i *Identifier variables* (e.g. name, NHS number, postcode, sex, date of birth) that are used within linkage but not necessarily the analysis (though some, e.g. sex, are also attribute variables). Some of these variables may be related to the values of other identifiers and/or attribute variables (e.g. values of name might be associated with sex and ethnicity). Presence of errors in one identifier might be related to errors in other identifiers (e.g. if name is mistyped, it might be more likely that date of birth is also recorded with error).

ii *Attribute variables* (e.g. ethnicity) that are used within analysis but not necessarily the linkage. Some of these variables may be associated with patterns in identifier values and/or identifier errors (e.g. ethnicity might be associated with *values* and also *errors* in name).

We make this distinction because under our motivating scenario, we are mostly interested in generating identifier variables. However, to ensure that the synthetic data are realistic, we need to consider i) the dependencies between identifier values and attributes, and ii) how errors in identifiers are distributed in relation to attribute variables. We often find that errors in linkage (and by implication, in identifiers) are related to differences in the underlying data quality for particular subgroups or to particular events and circumstances. For example, family name may have a higher probability of being typed incorrectly for individuals from minority ethnic groups compared to a majority ethnic group given that the family name may be unfamiliar to the operative recording the data, or that there are cultural differences in the length or structural complexity of names; linkage may be less likely to be successful for individuals following family separation given the tendency for this to result in changes in both address and names. This can therefore lead to dependencies between errors in identifiers – i.e. a change in name may be more likely for individuals who have also changed address. Evidence from the literature suggests that age, ethnic group, sex, deprivation and measures of health and social status may all be related to the risk of linkage error [14, 15]. In order to generate a dataset that is realistic for testing linkage methods, it is therefore crucial to consider whether identifier errors are likely to be related to attribute values. An example of the possible dependencies between identifier values, attribute values, and identifier errors is presented in Figure 1.

Steps in the process of generating synthetic identifier data

This section outlines five steps that are required to generate a realistic set of synthetic identifiers. In summary, the objective of these steps is to generate a 'gold-standard' dataset, i.e. the correct identifiers recorded in the absence of errors or changes over time (Steps 1–3). We then need to customise types and patterns of errors to be introduced to the gold-standard dataset (Step 4). Finally, we create multiple versions of the corrupted data (Step 5). A workflow for this process is presented in Figure 2.

Step 1: Elicit Information

Data linkers or data owners should elicit information on the set of identifiers in each file that are available for linkage, the rates of missingness (percentage of records with missing values for each identifier) and the characteristics of these identifiers (e.g. the range of dates of birth, the percentage of records that have a unique name).

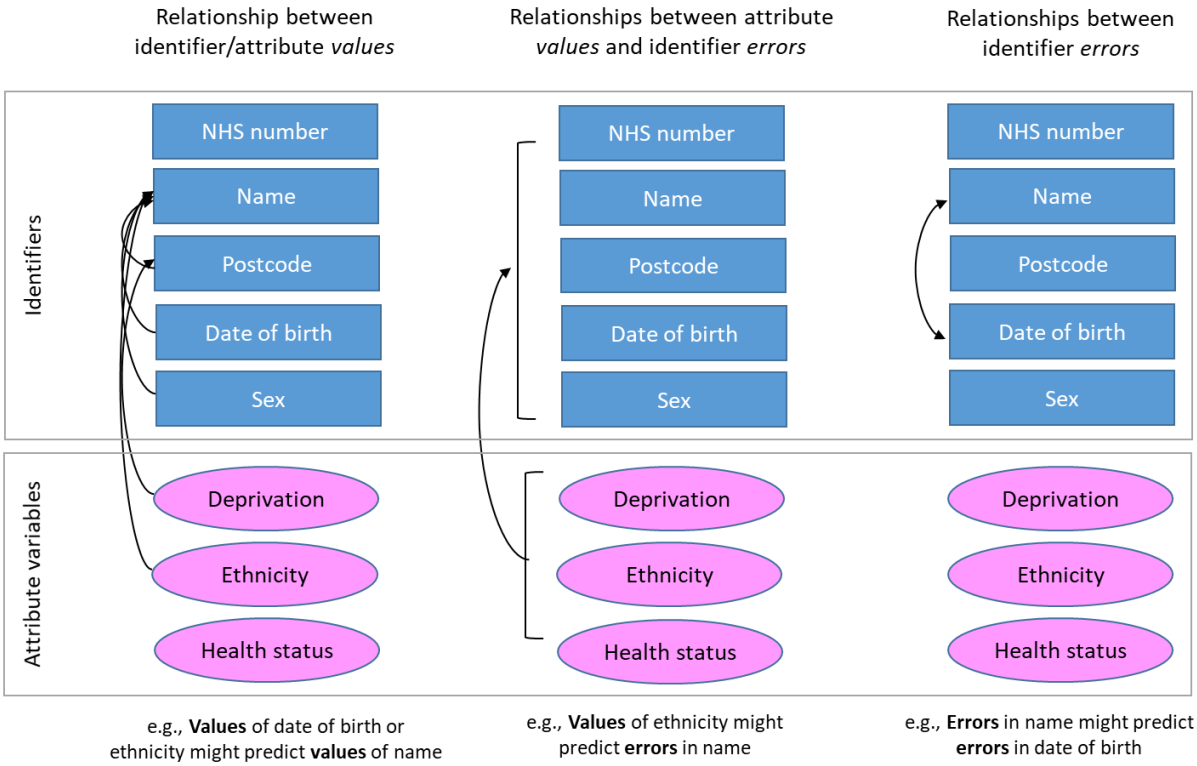
They also need to elicit information about likely rates of errors, types of errors and their patterns of co-occurrence in identifiers, and how these errors are associated with attribute variables. In practice, information on errors may be difficult to obtain, and may need to be based on knowledge about identifier errors (or linkage errors) from other similar data sources or the literature [16]. For the purposes of this paper, we use information on the rates, types and distribution of identifier errors based on analysis of data collected over different waves of the Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort study (see Section 2 for details on ALSPAC) [17, 18].

Information should be obtained on the number of records in each dataset and the joint distribution of key attribute variables (e.g. age and ethnicity), and whether individuals are likely to be recorded multiple times within a dataset (e.g., as they would in hospital admission records).

Step 2: Generate attribute variables

With access to the gold-standard identifiers and attribute data, we can use then use the *Synthpop* package in R to synthesise attribute variables (such as age or ethnicity) [19]. *Synthpop* uses a series of conditional models based on the original data to sequentially predict and impute values of each variable in the synthetic data. This process preserves variable inter-dependency by utilising classification and regression tree models. The output of this step is a 'gold-standard' synthetic dataset of attribute variables replicating those found in the original data.

Figure 1: Dependencies between identifier values, attribute values, and identifier errors



The examples given are not exhaustive but suggestive of the dependencies that might exist between values and errors in different variables.

Step 3: Generate identifiers

Two types of identifiers can be generated: those that are dependent on attribute variables, and those that are independent. Independent identifiers, e.g. NHS number (or social security number, etc.) can be generated according to predefined rules. Identifiers that are dependent on attribute variables will be generated according to the attribute values generated in Step 2. For example, date of birth can be generated according to the distribution of age. Table 1 describes how different types of identifiers might be generated, according to whether or not they are dependent on attribute variables.

Identifiers that are dependent on attribute variables and have high cardinality, such as names, are more challenging to synthesise. There is no existing library that readily generates names and maintains dependencies with other variables. Generation of names should consider the following factors:

1) Privacy and Disclosure Risk: ensuring none of the unique forename-surname combination in the original data appear in the synthesised data
2) Uniqueness
3) Frequency: common names should be synthesised for common names in the original data. For example: "John" (White, Male, common) in ALSPAC surnames or forename could be replaced with "Peter" (White, Male, common) from the name dictionary/look up table.
4) Sharing of surnames between siblings, parents and children and partners.

It is helpful to consider the frequency or uniqueness of different identifier values, as well as their distribution with respect to attribute variables. For example, a male has a higher probability than a female of having a forename of 'Patrick', and the distribution or uniqueness of names may vary according to ethnic group, levels of deprivation and by age (reflecting changing fashions for names).

Step 4: Data corruption

Table 2 provides a summary of different types of errors that may be found in real data and can be introduced during the synthetic data generation.

Data corruption is split into the following steps:

Firstly, error rates, types and co-occurrence patterns are defined and pre-specified.

Secondly, for each row of synthetic data, a corrupted version is generated. There are several approaches available for this data corruption. One approach is to generate multiple rows of corrupted data capturing all combinations of expected errors and patterns. This method retains all pre-specified error type combinations but could be computationally expensive for large datasets. Alternatively, the Splink synthetic data corruptor adapts a likelihood approach to introducing errors, generating multiple rows of corrupted data probabilistically [20]. In Splink's synthetic data corruptor, a baseline probability is assigned for each type of error, and a multiplier is applied based on attribute variables. For example, by following the Zipf distribution, up to 20 rows with varying error types and combinations can be generated for each row of data [21]. This method is less computationally expensive and

Table 1: Generating identifier values

Identifier		Generation process	Example
Identifiers that are independent of attribute variables	Date of birth and sex	Given aggregate information on the distribution of these identifiers, or elements of these identifiers (i.e. year of birth) within the original data, values can be sampled directly from the relevant distributions. In some cases, we might want to reflect dependency between records, for example, there will be a minimum distance between the date of birth of a baby and their mother. In other cases, we might want to allow date of birth to depend on attribute variables, such as place of residence.	Date of birth can be generated from the distribution of year of birth, by assuming a uniform distribution over all possible (or eligible) dates within each year. We could also allow for variations according to day of the week or month of the year (e.g. those born on 31st of December of any year may have a higher probability of being recorded as being born on 1st January of the next year, rather than another random date).
	Unique identifiers	Values of unique identifiers such as a social security number or NHS number can be randomly generated following defined rules. The assumption that unique identifiers are independent does not hold where another identifier is included within in the unique identifiers (e.g. the Community Health Index number in Scotland, which is derived from date of birth and sex).	NHS number is assigned at birth in England and is unrelated to any other personal information [22]. It comprises ten digits, of which the majority are random numbers and the tenth is a check digit to confirm validity: it can therefore be generated using a simple algorithm. If there are multiple unique identifiers (e.g. NHS number and hospital number), these can be generated independently.
	Other identifiers	Personal identifiers such as email addresses, telephone numbers, and social media handles can be generated according to rules. In some cases, we might also want to allow these identifiers to depend on attribute variables: e.g. generating random telephone numbers based on the country and area of residence, or generating random email addresses based on names, date and country of birth.	Fake Mail Generator (https://fakedetail.com/fake-mail-generator) allows the generation of random email addresses given real domains. We can also allow these identifiers to depend on other identifiers or attributes: for example, Fake Number (https://fakenumber.org/united-kingdom) can generate random telephone numbers based on the country and area of residence.
Identifiers that are dependent on attribute variables	Names	First names may be related to age, ethnicity, sex and geography; surnames may also be related to ethnicity. Frequency look-up tables provide a useful tool for sampling names and mapping them to predictor attribute variables. Names can be directly sampled from such frequency tables, and can be allowed to depend on attribute variables such as sex and ethnicity, where these are available.	The Office for National Statistics (ONS) publishes the rank and count of the baby birth names in England and Wales every year, which can be used as the forename frequency table for the England and Wales population [23]. National Records of Scotland also publish popular baby forenames depending on year of birth and gender [24]. Another example is data on forename, gender and ethnicity extracted from the US census and implemented in the R package 'randomNames' [25]. Similar frequency tables, including for surnames, are published in many countries [26].
	Addresses	Addresses may be related to personal social status, income and ethnic background [27]. For example, in 2018, 41% of residents in the London borough of Tower Hamlets were of Asian ethnic background, compared with 5% in the borough of Bromley.	To represent these dependencies in synthetic data, we can start by sampling postcodes from a relevant list. Levels of deprivation can then be assigned to each postcode using the English indices of deprivation (Index of Multiple Deprivation; IMD), and ethnic group distributions can be assigned using ethnic group statistics by geography [28, 29]. Given information on the distribution of ethnic group in the original data, addresses can then be sampled from a frequency table.
	Indirect identifiers	Other non-traditional identifiers used for linkage might include 'indirect' identifiers such as clinical variables or dates [30]. Given sufficient aggregate data on the distributions of these variables, and assumptions about their dependence on attribute variables, these could be generated in a similar way to date of birth and sex (i.e. according to specified distributions).	

Table 2: Types of identifier errors that can be introduced to synthetic data

Type of error/identifier	Description	Manifestations
Typographic error (string variables)	• Occurs during manual typing, e.g. a receptionist types a patient's information for a general practitioner appointment booking. • Depending on the keyboard layout, characters may be *substituted* with neighbouring keyboard characters e.g. 's' instead of 'd'. • New characters or space may be accidentally *inserted* into a field, random characters may be *omitted* from a field, or character positions may be *transposed.* • Errors may result from hitting a key twice, letting eyes move faster than the hand, or misreading [31].	• Typographical errors are more likely to happen in the middle or towards the end of the word, and in longer words [32, 33]. • Over 80% of typographical errors are single instances of substitution, insertion, deletion or transposition [31]. • The likelihood of substituting neighbouring characters differs according to layout as well as personal typing habit, e.g. it is more likely that 'd' is replaced with 's' than with 'x' [34].
Phonetic error (string variables – particularly name)	• Occurs during dictation, where letters may be substituted with letters that are phonetically the same but orthographically incorrect for the intended word, e.g. when a receptionist records information given by a patient, (s)he may mishear information due to the accent of the patient or the pronunciation of similar words or characters, such as 'F' instead of 'Ph' [34].	• Information on phonetic errors can be derived from phonetic algorithms, which apply a range of rules and exceptions to encode words by their pronunciation, instead of spellings. These algorithms have been widely used in applications such as spell checkers and search engines and algorithms have been transformed into look-up tables and rules to group similar-sounding words together [35]. • Soundex is one of the most widely known phonetic algorithms for Anglo-Saxon surname encoding [36]. Extensions to Soundex overcome limitations in recognising different languages and dialects that may have different pronunciations for the same names [37].
Optical Character Recognition error (OCR, any identifiers)	• The OCR system is used to process scanned handwritten documents into electronic versions. • OCR errors occur when the system fails to distinguish two characters that have similar shapes, such as 'l' and '1' or 'm' and 'rn'.	• Error rates in OCR systems can be high if the scanned documents are poorly handwritten, in bad physical condition or have a complex layout [38]. • Error rates in OCR systems are impacted by configuration settings (such as where the threshold is set for manual review). • Look up tables are available that provide around 80 pairs of OCR errors where letters, digits, symbols and combinations of these appear to be similar.
Naming convention inconsistencies	• Some people have two first names (with or without a hyphen), or middle names that are used as first names or vice versa. • Double-barrel surnames may be recorded differently in different datasets (e.g. with or without hyphens) and may include abbreviations (e.g. Saint John as St. John). • First names and surnames may be swapped. • Migrant groups might 'adopt' localised versions of names • Nicknames and diminutives might be provided	• Look up tables of common name variants are available. The software 'Febrl' provides around 350 rules and name variants (e.g. 'Edward' for 'Ted', 'Edwin' and Edwards') [13]. Database of common English diminutives of formal given names are available on Wiktionary. • Table of common surnames with different Romanised representation of the same character are available on Wiktionary (Mandarin Chinese, Cantonese, Hakkan, Korean, Vietnamese, Japanese)

Continued

Table 2: Types of identifier errors that can be introduced to synthetic data

Type of error/identifier	Description	Manifestations
Date errors (date of birth or other date identifiers)	• Format differences, i.e. between countries or people. In the UK, people usually record their date of births in Day-Month-Year format, while in the US it is more often written in the format of Month-Day-Year and in China is Year-Month-Day. • Default/generic values. Some systems have a default value for the date of birth, resulting in those people with missing date of birth automatically being given a default date. • Accidental input of 'today's date'	
Changes over time (e.g. name, sex, postcode)	• Postcode changes occur as people move and if addresses are not updated on a system (e.g. postcodes in healthcare data might only be updated when a patient registers with a new general practitioner, which might be some time after an address change). • Children may have multiple genuine postcodes if they have more than one residence, e.g. mother's or father's address. • Surnames may change following marriage or divorce; recorded sex may change over time. • Postcodes change over time for the same property to reflect changes in the postal system.	• In the UK, evidence suggests that 40% of children move home in the first 5 years of life; 5% move 3 or more times within this time period [39].
Unique identifier errors	• Checksums or other validation methods may be used to prevent invalid identifiers from being recorded. • Intentional use of another person's identifier may lead to errors. • Changes to unique identifiers may occur over time, and some identifiers might be reused, resulting in multiple individuals with the same identifier [40]. • Individuals may be issued many unique IDs (e.g. a pupil moving from one school to another)	• Accurate recording of unique identifiers that depend on interactions with services may be related to how different individuals access those services. For example, completeness of NHS number is often lower for young males [41].

has the capability to introduce some error-attribute variable dependency. However, this method does not necessarily capture all pre-specified error type combinations and co-occurrence patterns.

The final stage is to draw samples from the corrupted data that satisfy the pre-specified error types, co-occurrence patterns, and error-attribute characteristics.

Step 5: Generate linkage files

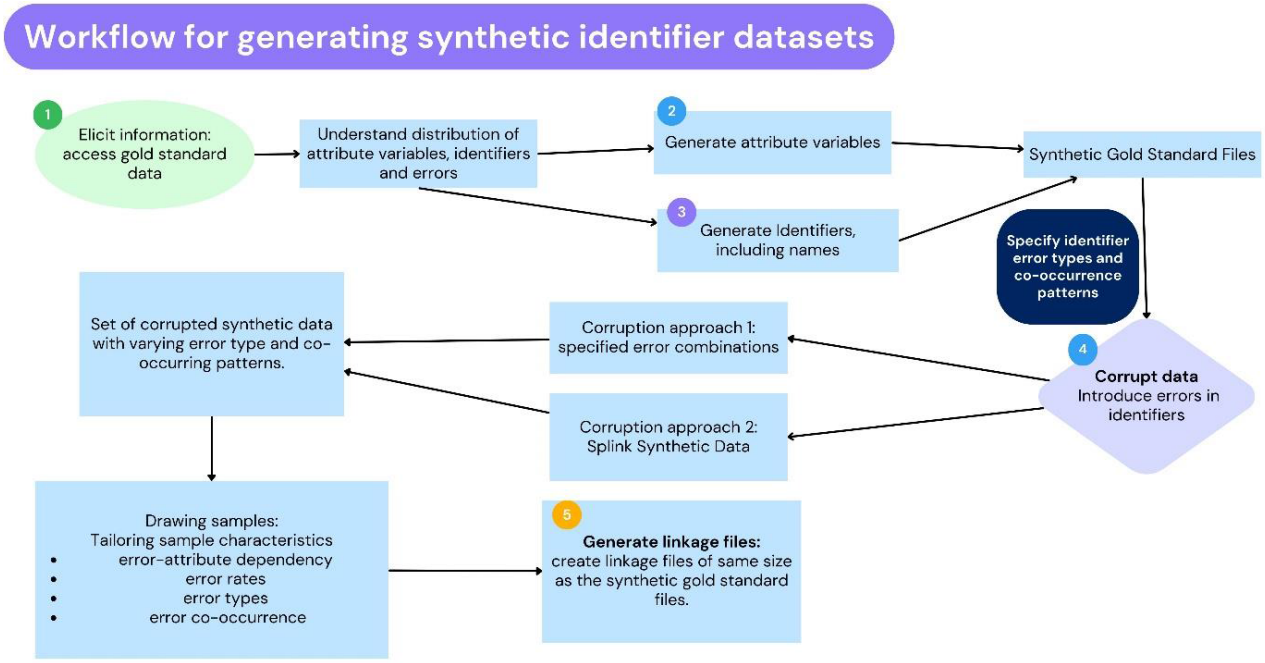
Since the errors selection in Step 4 is probabilistic, we can generate multiple sets of corrupted data files by repeating the step. This gives us several (e.g. 5) different corrupted versions of the same gold standard file, which represent multiple versions of a 'linkage' file. Generating multiple versions of the linkage file is appropriate as it reflects the uncertainty in the process of replicating the original data, in line with the logic of using multiple imputation to model uncertainty.

Section 2: Evaluating synthetic data

Motivating scenario

The following section describes an evaluation of the utility of the data we have generated under our framework. We use an exemplar of data linkage within the ALSPAC birth cohort. In ALSPAC, identifiers for each participant were recorded at multiple time points or data collection waves. For the purposes of evaluating the synthetic data, we used data from a gold-standard list of identifiers held within the ALSPAC administrative database (called ARCADIA, see Appendix Table 1) which contains the 'live' best understanding

Figure 2: Workflow for generating synthetic identifier datasets



of participants current details, and raw records from one data collection wave (the Child Health Database; CHDB), collected when participants were aged 6 years. A unique ALSPAC ID identifies the same individual within ARCADIA and CHDB, but the identifiers collected in each dataset differ. This gives us a gold-standard database that can be used to assess how well synthetic data performs at evaluating different linkage approaches.

We first generate synthetic versions of the identifier data held within ALSPAC, creating a number of 'linkage files' to represent ARCADIA and CHDB. Next, we link the synthetic versions of ARCADIA with synthetic versions of CHDB, and derive metrics of linkage quality. Finally, we compare the linkage quality metrics derived from the synthetic data to the metrics derived from the gold-standard ALSPAC data.

Source data

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a prospective population-based study [17, 18]. Initial recruitment of pregnant women took place in 1990-1992 and the health and development of the children from these pregnancies and their family members have been followed ever since. For this study, we focus on the original parents/carers (Generation 0, G0) and the index children (Generation 1, G1). ALSPAC recruited 14,541 pregnancies by women (G0) who were resident in and around the City of Bristol (South West UK) with expected dates of delivery 1st April 1991 to 31st December 1992. Of these initial pregnancies, there were a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. The eligible sampling frame was constructed retrospectively using linked recruitment and health service records. Additional offspring that were eligible to enrol in the study have been welcomed through major recruitment drives at the ages of 7 and 18 years; and through opportunistic contacts since the age of 7. A total of 913 additional G1 participants have been enrolled in the study since the age of 7 years with 195 of these joining since the age of 18. This additional enrolment provides a baseline sample of 14,901 G1 participants who were alive at 1 year of age.

Linkage methods

Our aim was to determine whether we could use synthetic data to evaluate the quality of different linkage algorithms. Therefore, we used three different linkage strategies to link data for 13,281 individuals in ARCADIA who also had a record in CHDB. We conducted the linkage based on child's forename, surname, date of birth and gender, plus mother's surname, using the following methods, with further details in Appendix 4. Linkage strategies were compared and probabilistic linkage thresholds were chosen to align with the deterministic linkage model, to enable a fair comparison. We estimated false match and missed match rates for each method.

1. Deterministic linkage. We classified records as belonging to the same individual if at least 4 of the 5 identifiers matched exactly.

2. Probabilistic linkage with similarity scores. We calculated probabilistic match weights for agreement/disagreement using the Fellegi-Sunter approach [42]. To allow for typographical errors in names, we calculated probabilistic match weights using the Jaro-Winkler similarity score [33]. Similarity scores were categorised as little agreement (a score of 0–0.8), moderate agreement (0.8–<1), or full agreement (a score of 1). Linkages were accepted at or above the weight threshold of 3.

3. Probabilistic linkage with similarity scores and term frequency adjustments for forenames and surnames. On

top of method 2, we accounted for name frequencies by proportionally adjusting u-probabilities for agreement or disagreement on less common names. Linkages were accepted at or above the weight threshold of 2.

Generating synthetic ALSPAC data

In order to generate realistic synthetic data, we first needed to understand the levels of errors observed in identifiers within ALSPAC. Since we had access to the gold-standard ALSPAC data, we could directly estimate the error rates for each identifier (see Appendix 1, Appendix Tables 2–4).

We generated synthetic data to replicate the two ALSPAC datasets described above (ARCADIA and CHDB). Using "*Synthpop*" in *R*, we generated a 'gold-standard' dataset of identifier and attribute variables (apart from forenames and surnames) to replicate ARCADIA [19]. The dataset contained 13,281 records and was generated using sequential regression modelling based on the original ALSPAC data, using date of birth, gender, maternal age category, ethnic group, and quintile of the Index of Multiple Deprivation (Appendix 3, Appendix Tables 5, 6). Given the small number of people with non-white ethnicity, not all combinations of maternal age and ethnicity exist in the original data. We used a rejection sampling mechanism to ensure synthesised dataset did not generate combinations of attribute variables that did not appear in the original study [43]. Detailed methodology used to synthesise attributes, identifiers and names are described in Appendix 2 and 3.

Data corruption

Four different data corruption approaches were used to examine how results were affected by differences in the types, co-occurrences and dependencies of errors that were introduced to the synthetic data:

1) Error types: We varied whether or not the synthetic data had the same types of errors as original data.

2) Error Field co-occurrence pattern: We varied whether or not the synthetic data had the same pattern of co-occurrence at the field level (e.g. 5% of errors co-occur in G1 forename and surname).

3) Error type co-occurrence pattern: We varied whether synthetic data had the same pattern of co-occurrence of errors at field and type level. For example, 5% of errors co-occurred in G1 forename and G1 surname; 30% of the error co-occurrence is a random name replacement, and 70% of the co-occurrence is a forename variant error with random surname replacement).

4) Error-attribute variable dependency: We varied whether or not the error rates were dependent on attribute variables (in our case, maternal age and ethnic group).

Generating linkage files

Under each of the four scenarios below, we created five synthetic datasets to examine differential impact of error distribution and characteristics on linkage.

1. Scenario 1: Error rates were based on known values derived directly from the original data source. We specified the error rate for each identifier. We allowed identifier error rates to vary according to maternal age and ethnic group. Identifier errors were of the same types as in the original (e.g. 95% surname errors were random replacements). We used the same error co-occurrence patterns as the original data.

2. Scenario 2: Error rates were assumed to be unknown but were assumed to be dependent on maternal age and ethnic group. Identifier errors were restricted to random replacements. We did not allow errors to co-occur in this scenario.

3. Scenario 3: Error rates were assumed to be unknown and were assumed to be independent of attribute characteristics (i.e. constant across maternal age and ethnicity). Identifier errors were of the same types as in the original data but the error co-occurrence pattern was assumed to be unknown.

4. Scenario 4: Error rates were assumed to be unknown but were assumed to be independent of attribute characteristics. In this scenario, we assumed that identifier error rates were constant across maternal age and ethnicity. Identifier error types were randomly assigned. We did not allow errors to co-occur in this scenario.

Deriving linkage quality metrics

Using the three linkage methods described, we linked the five synthetic gold-standard datasets to each of the corrupted synthetic datasets in the four scenarios. Since we had generated these data ourselves, we knew the true match status of each record pair. We were therefore able to evaluate the quality of each linkage method by deriving the rates of missed matches (true links that were matched) and false matches (records that were linked to the wrong individual) for each linkage method. Estimates were averaged over the five synthetic datasets. We then compared these results with linkage error rates derived from the original source data.

Results

Linkage results

There were 13,281 records that linked between the ARCADIA and CHDB datasets based on the gold-standard ALSPAC data. Using deterministic linkage, 12,673 individuals were linked. The number of linked records ranged from 12,920 with probabilistic linkage using similarity scores for comparing names, to 12,962 with probabilistic linkage using term frequency adjustments for comparing names (Table 4). Rates of errors (both missed matches and false matches) were lower using probabilistic compared with deterministic linkage, and lowest with the addition of term frequency adjustment. All results presented for the synthetic linkages were averaged over 5 synthetic datasets.

Table 3: Identifier error rates introduced to synthetic linking files. No errors were introduced to sex or date of birth, and no missing values were introduced

			G1 Surname^	G1 Forename*	G0 Surname^
Scenario 1: Known error rates	**Error Co-occurring Patterns (% of all records)**	**Maternal Age**	% error	% error	% error
	G0 surname & G1 surname & G1 forename (0.30%)	<20	11.1	6.6	36.7
	G1 surname & G1 forename (0.43%)	20-29	6.0	11.5	19.7
	G1 surname & G0 surname (1.90%)	30-39	4.2	14.5	12.0
	G1 forename & G0 surname (1.90%)	40+	5.8	15.6	13.6
		Missing	7.5	11.5	16.5
		Ethnic group			
		White	5.6	13.0	17.5
		Black	8.8	13.6	22.4
		Asian	1.9	13.3	8.6
		Other	12.2	14.9	18.9
		Missing	5.7	9.2	17.7
Scenario 2: Estimated error rates#	No co-occurring errors	**Maternal Age**			
		<20	7.2	13.1	4.7
		20-29	4.6	9.4	5.7
		30-39	4.0	8.4	6.2
		40+	6.1	5.7	6.6
		Missing	6.1	7.7	5.4
		Ethnic group			
		White	4.4	9.2	5.9
		Black	7.2	14.3	9.2
		Asian	4.1	6.8	7.4
		Other	16.4	10.5	7.3
		Missing	5.0	8.5	5.2
Scenario 3: Independent error rates	G0 surname & G1 surname (2.30%)		5.0	10.0	15.0
	G1 forename & G1 surname (1.00%)				.
	G1 forename & G0 surname (0.75%)				
Scenario 4: Independent error rates	No co-occurring errors		5.0	10.0	15.0

*73% of errors were name variants (e.g. Sam for Samuel, Becky for Rebecca); 14% were typographical errors (e.g. insertions/deletions); 7% were due to one dataset recording multiple first names (e.g. Lisa Marie versus Lisa), 6% were completely different names.
^3% of the errors were due to the gold-standard dataset having two surnames (e.g. Harron Kent) and the linking file only having the first name (Harron); 2% were where the gold-standard had two surnames but the linking file only has the second name (Kent).
Error rates presented are based on the relative risk of identifier errors according to attribute variables in Appendix 1, Appendix Table 4, with estimated baseline likelihood of error of 0.1 (G1 Surname), 0.15 (G1 Forename), 0.2 (G0 Surname).

Linkage quality metrics

All of the synthetic datasets broadly replicated the same pattern seen in the original data linkage, i.e. that rates of missed matches were lower than rates of false matches, and that probabilistic linkage with similarity scores and term frequency adjustment had the best performance (Table 4). Scenarios 1 and 3 result in comparable linkage error

Table 4: Comparison of linkage quality metrics based on the original ALSPAC data, and synthetic data generated under three scenarios

		Deterministic linkage	Probabilistic linkage with similarity scores	Probabilistic linkage with similarity scores and term frequency adjustment
Original data	*n linked records*	*12,673*	*12,920*	*12,962*
	*Missed match rate**	*4.59%*	*2.61%*	*2.40%*
	*False match rate***	*0.23%*	*0.12%*	*0.05%*
Synthetic data – known error rates, dependent on attributes, original error co-occurence[1]	n linked records	12,656	12,718	12,712
	Missed match rate	4.72%	4.26%	4.29%
	False match rate	0.29%	0.32%	0.16%
Synthetic data – guessed error rates, dependent on attributes, no error co-occurence[2]	n linked records	13,274	13,276	13,279
	Missed match rate	0.05%	0.04%	0.02%
	False match rate	0.09%	0.12%	0.07%
Synthetic data – guessed error rates, independent of attributes, assumed pattern of error co-occurence[3]	n linked records	12,746	12,817	12,809
	Missed match rate	4.04%	3.50%	3.56%
	False match rate	0.25%	0.24%	0.13%
Synthetic data – guessed error rates, independent of attributes, no error co-occurence[4]	n linked records	13,266	13,277	13,279
	Missed match rate	0.12%	0.03%	0.02%
	False match rate	0.10%	0.10%	0.04%

[1]Scenario 1: Error rates were specified correctly, based on the original data source (Table 3).
[2]Scenario 2: Error rates were guessed, and were allowed to vary according to maternal age and ethnicity (Table 3).
[3]Scenario 3: Error rates were guessed and were assumed to be unrelated to attribute characteristics, with estimated error co-occurrence patterns: 15% G0 surname-G1 surname, 10% G1 forename-G1 surname, 5% G1 surname-G1 forename. Types of error when errors co-occur: G0 surname and G1 surname errors = random replacement, G1 forename and G1 surname errors = random replacement (surname) + 30% typo, 70% forename variant (Table 3).
[4]Scenario 4: Error rates were guessed and were assumed to be unrelated to attribute characteristics, and no restriction on error co-occurring patterns (Table 3).
*missed match rate = % of true matches that were not identified, i.e. 1-sensitivity.
**false match rate = % of linked records that were not true matches, i.e. 1-positive predictive value.

rates compared to the original linkage, and successfully demonstrated that probabilistic linkage was able to reduce both false-matches and missed-matches compared with deterministic linkage.

Across the deterministic linkages, scenarios 1 and 3 had more comparable linkage error rates, with an absolute difference of 0.13–0.55% for missed matches, and 0.00-0.04% for false matches. Linkages for scenarios 2 and 4 had larger variations of linkage errors compared to the original, with a difference of 4.05–4.12% for missed matches, and 0.15–0.16% for false matches.

In scenario 1, linkage error rates were slightly over-estimated in the synthetic data, by 0.55–2.02% for missed matches and 0.04–0.11% for false matches (Table 4). The difference in estimation was similar in scenario 3, at 0.13–1.26% for missed matches and 0.00-0.12% for false matches.

In scenario 2, linkage error rates were under-estimated in the synthetic data, by 2.20–4.12% for missed matches and 0.00–0.16% for false matches (Table 4). Under-estimation was found to a similar extent in scenario 4, by 2.21–4.05% for missed matches and 0.01–0.15% for false matches.

Missed match and false match characteristics

In the original linkage, of the 346 true matches missed by probabilistic linkages with similarity scores, 65.0% were those where there was agreement on forename, date of birth and gender, but disagreement on surname and mother's surname. These missed matches affected people of different ethnicities and genders similarly and affected younger mothers more than older mothers. Use of term frequency adjustment further reduced missed matches to 319. These missed matches appeared to correspond to cases in which both the mother and the child changed their surname between data collection waves (rather than being due to typographical errors). The second most common missed match pattern occurred for records with disagreement on forename and mother's surname, with 13.6% in probabilistic linkage with similarity scores, and 13.8% with term frequency adjustments. These missed matches appeared to correspond to mothers changing their surnames, and children providing alternative names or derivatives at different data collection waves.

Missed match rates were comparable to the original linkage in scenarios 1 and 3. The disagreement pattern of missed matches were also similar to the original linkage (Appendix Table 8).

Compared to the original linkage, missed matches in scenario 3 had similar distributions of disagreement patterns. With probabilistic linkage with similarity scores, 61% of missed matches disagreed on surname and mother's surname; with term frequency adjustment, 58% missed matches disagreed on surname and mother's surname. In both probabilistic linkage with similarity scores and term frequency adjustment, 19.3% missed matches disagreed on forename and mother's surname.

Comparing to the original linkage, missed matches in scenario 1 had a lower proportion of disagreements on surname and mother's surname with 42.3% for probabilistic linkage and 40.2% for term frequency adjustments. Higher proportions of missed matches disagreed on forename and mother's surname, with 37.9% for probabilistic linkage, and 37.8% for term frequency adjustments.

False-match rates were low in the original linkage and synthetic linkages. The higher rate of false-matches with deterministic linkage was predominantly explained by the 54% of record pairs that agreed on surname (both mother and child), sex and date of birth, but disagreed on forename. This was followed by 39% of false-matches in pairs that disagreed on date of birth only (Appendix Table 9). In the deterministic linkages using synthetic datasets, similar patterns and proportions of false-matches records were replicated, where 60.0% of false-matches disagreed only on forename, and 34.9% disagreed on date of birth only. As date of birth was recorded with high accuracy in the ALSPAC data, these pairs were (correctly) not accepted as links by the probabilistic strategies (since disagreement on date of birth conferred a large penalty to the match weight).

In terms of missed match rates, false match rates, and characteristics of missed matches, we were able to best produce linkages most similar to original linkage in scenario 3. This demonstrates that replicating error types and co-occurrence patterns (even if the co-occurrence patterns are estimated) without incorporating dependencies between error and attribute is sufficient to produce realistic synthetic data. Further incorporating information about dependencies between errors and attribute (with true error rates), and error co-occurrence patterns (scenario 1) did not produce substantially more realistic linkages.

Conversely, retaining error and attribute dependency without incorporating error types and co-occurrence (scenario 2) performed similarly to when identifier error rates were assumed to be independent (scenario 4).

Discussion

We provide a generalisable and open-source framework for generating synthetic identifier data that can be used to facilitate development and evaluation of improved linkage methodologies. We show how this framework can be implemented and provide a means of producing corrupted datasets that can be used for linkage development and a complete 'gold standard' file that can be used for linkage validation. We generated synthetic ALSPAC identifier datasets, which are freely available for legitimate users on request to the authors: the intention is that these data, with known characteristics, can be used for the development and comparative benchmarking of different linkage approaches.

Our framework builds on previous methodological work aiming to generate synthetic identifier data for use in data linkage [3, 13]. We extended previous work by overcoming the assumption of independence of identifier errors through explicitly incorporating the associations between identifier errors and attribute variables. If accurate information on the joint distribution of identifiers and identifier errors were available, there would be no need to include information on their dependencies with attributes. However, evaluating linkage quality according to attributes such as age, sex and ethnicity is convenient and intuitive, and knowledge of how linkage errors are typically distributed amongst these subgroups can be easily incorporated into synthetic data generators [44]. Our findings comparing linkage quality metrics for synthetic data generated under different scenarios highlight that preserving error types and co-occurrence patterns is vital for generating a dataset that accurately represents real-world data and that can be meaningfully used to evaluate linkage algorithms, and is useful when incorporating the dependencies between identifier errors and attributes is not easily achievable [16]. This framework can be used to test linkages between more than 2 datasets.

The strengths of our study include the use of gold-standard data from a large cohort study that was used to assess the performance of synthetic data for deriving linkage quality metrics. We compared a range of linkage methods and different scenarios under which the synthetic data were generated. It is likely that the errors observed in these data are representative of those occurring in other administrative and research datasets. We acknowledge that the exemplar ALSPAC dataset is predominately of a White UK population, and recommend that other cultural, geographic and time-point specific alternatives are generated in order to avoid any unintended bias in linkage algorithm development (i.e. to factor in error patterns that exist yet were not observed in the ALSPAC data). However, synthetic data generators such as this one should give users the ability to alter the identifier error rates, types and co-occurrence patterns according to their particular data context. This allows for any uncertainty to be explored, by using a range of error rates and patterns to investigate how results may vary. This could be used to help inform choice of linkage strategy: for example, it could tell us that a simple deterministic approach might generate results of sufficiently high quality if identifier error rates are low, whilst a more sophisticated and resource intensive probabilistic approach might be more suitable in settings where identifier error rates are high. It could also point to possible improvements in algorithms: in our example, all three linkage algorithms failed to identify true matches where there was a disagreement on surname and mother's surname: better handling of name specific characteristics, such as double-barrel surnames, could go some way to mitigating this problem. Using synthetic data could also be used to provide a plausible range of linkage error rates that are likely to arise, given different assumptions about the levels of identifier errors. Under these assumptions, researchers can explore the sensitivity of their linkage approach by assessing the impact of including or

excluding certain error-prone identifiers on linkage rates. This is particularly relevant for longitudinal population data, where richer insight in the variation of identifier errors is more observable, researchers could demonstrate with which data sets the original data could be best linked. Researchers can then use different methods to account for linkage error rates within analysis, e.g. quantitative bias analysis to explore the extent to which results of analyses might be affected by linkage error rates [45]. Synthetic identifier data would be particularly useful for evaluating the quality of privacy preserving linkage techniques, where access to identifiers in the clear is not permitted. Currently, access to real data is needed generate synthetic identifier data. Alternative approaches, such as estimating parameters from existing publications, could provide information sufficient to assess linkage quality to a certain extent (such as Scenario 3, where error rates for each variable were educated guesses). However, this approach might be blind to error characteristics, error co-occurring patterns, and error inter-dependencies that may underlie specific data sources. As these synthetic data would be used to evaluate the validity and utility of the linkages, using mis-specified models, or multiple proposed synthetic models would confer to challenges in data governance. Our proposed framework, while seemingly relying on higher involvement of the data owners, has the advantage of giving more control to data owners, and presents as a more pragmatic approach to drive change.

Limitations of our study are that we only had one gold-standard dataset with which to evaluate the performance of the synthetic data and therefore our testing of dependencies is based on information about a specific population group; further evaluations should be conducted on other datasets with varying proportions of missingness in identifiers. Our name generation mechanism takes advantage of the small sample size and low cardinality of name distributions in ALSPAC (4,000–7,000 distinct forename and surname terms). Replication using the same method would require a more diverse name dictionary. The key advantage of generating realistic names with name dictionaries, (versus string or number sequences), is the potential to better reflect dimensions of name characteristics that are non-metricized and may associate with error distributions by attributes. The current name generation mechanism did not fully preserve name clusters and name-specific characteristics, such as word length, hyphens or number of terms per name [46]. Our framework could be extended in several ways, including by adding in additional variable types, error types and error co-occurrence patterns, by allowing the generation of data at the household level, or for multiple generations to capture between record dependencies. More sophisticated synthetic identifier data might include more nuanced errors (i.e. specifying the most likely letter transpositions based on keyboard strokes, or introducing date-specific errors such as recording today's date). However, these nuances would only be required if the linkage algorithm that was being evaluated was tailored towards resolving these specific sorts of errors. A further problem is on assessing how accurate the error type and co-occurrence pattern has to be for the generated synthetic data to be considered similar enough to reliably test the proposed linkage methods. Our study offers an approach to start investigating this idea more structurally, by contrasting multiple data corruption scenarios. Further investigations on this direction would allow us to be more confident in our comparisons.

Our framework provides a novel and generalisable mechanism for developing and benchmarking record linkage algorithms, which is protective of public privacy and avoids assumptions that errors in personal identifiers are independent of the other identifiers and attribute data. Our findings show that replicating dependencies between attribute values (e.g. ethnicity), values of identifiers (e.g. name), and errors in identifiers (e.g. missing values, typographical errors or changes over time) and its patterns enables generation of realistic synthetic data that can be used to evaluate different linkage methods.

Conflicts of Interest

The authors declare there is no conflict of interest.

Acknowledgements

Harvey Goldstein had a key role in developing this study, but sadly died prior to publication. We are very grateful to his input to this work. We would also like to thank Haoyuan Zhang for his early input to this work.

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We particularly thank Mark Mumme for his time in producing extracts of data for this study. We thank Ruth Gilbert and James Doidge for their input to and feedback on this work.

Ethics

Ethical approval for the ALSPAC cohort study was obtained from the ALSPAC Ethics and Law Committee (a University of Bristol Faculty Ethics Committee) and NHS Local Research Ethics Committee(s). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. This study was approved through the ALSPAC data access mechanism (ALSPAC Reference: B3002, https://proposals.epi.bristol.ac.uk/?q=node/127384). Please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data).

Funding

The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and Katie Harron and Andy Boyd will serve

as guarantors for the contents of this paper. This research was funded in whole, or in part, by the Wellcome Trust [212953/Z/18/Z]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data availability

Synthetic ALSPAC data can be downloaded on UCL Research Data Repository (doi:10.5522/04/25921408). Original ALSPAC data can be requested from ALSPAC website.

References

1. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data & Society. 2017 Dec;4(2):205395171774567. https://doi.org/10.1177/2053951717745678

2. Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. Public Health Res Pr [Internet]. 2015 [cited 2023 Dec 7];25(4). Available from: http://www.phrp.com.au/issues/september-2015-volume-25-issue-4/routinely-collected-data-as-a-strategic-resource-for-research-priorities-for-methods-and-workforce/ https://doi.org/10.17061/phrp2541540

3. Christen P, Vatsalan D. Flexible and extensible generation and corruption of personal data. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13 [Internet]. San Francisco, California, USA: ACM Press; 2013 [cited 2023 Dec 7]. p. 1165–8. Available from: http://dl.acm.org/citation.cfm?doid=2505515.2507815. https://doi.org/10.1145/2505515.2507815

4. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data–a best practice protocol. Aust N Z J Public Health. 2002;26(3):251–5.

5. Harron K, Wade A, Muller-Pebody B, Goldstein H, Gilbert R. Opening the black box of record linkage. J Epidemiol Community Health. 2012 Dec;66(12):1198. https://doi.org/10.1136/jech-2012-201376

6. Christen P. Probabilistic Data Generation for Deduplication and Data Linkage. In: Gallagher M, Hogan JP, Maire F, editors. Intelligent Data Engineering and Automated Learning - IDEAL 2005 [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005 [cited 2023 Dec 7]. p. 109–16. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Lecture Notes in Computer Science; vol. 3578). Available from: http://link.springer.com/10.1007/11508069_15. https://doi.org/10.1007/11508069_15

7. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. J Biomed Inform. 2012 Feb;45(1):165–72. https://doi.org/10.1016/j.jbi.2011.10.006

8. Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 1. Statistical Journal of the IAOS. 2017 Jan 1;33(3):785–96. https://doi.org/10.3233/SJI-150153

9. Kokosi T, De Stavola B, Mitra R, Frayling L, Doherty A, Dove I, et al. An overview on synthetic administrative data for research. IJPDS [Internet]. 2022 May 23 [cited 2022 Jul 7];7(1). Available from: https://ijpds.org/article/view/1727. https://doi.org/10.23889/ijpds.v7i1.1727

10. Raghunathan TE. Annual Review of Statistics and Its Application Synthetic Data. Annual Review of Statistics and Its Application. 2021;8(1):129–40. https://doi.org/10.1146/annurev-statistics-040720-031848

11. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. International Journal of Epidemiology. 2019 Dec 1;48(6):2050–60. https://doi.org/10.1093/ije/dyz203

12. 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. International Journal of Epidemiology. 2017 Oct 1;46(5):1699–710. https://doi.org/10.1093/ije/dyx177

13. Christen P, Pudjijono A. Accurate Synthetic Generation of Realistic Personal Information. In: Theeramunkong T, Kijsirikul B, Cercone N, Ho TB, editors. Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer; 2009. p. 507–14. (Lecture Notes in Computer Science). https://doi.org/10.1007/978-3-642-01307-2_47

14. Bohensky M. Chapter 4: Bias in data linkage studies. In: In: Harron K, Dibben C, Goldstein H, editors Methodological Developments in Data Linkage [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2023 Dec 7]. p. 63–82. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119072454.ch4. https://doi.org/10.1002/9781119072454.ch4

15. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data Linkage: A powerful research tool with potential problems. BMC Health Services Research. 2010 Dec 22;10(1):346. https://doi.org/10.1186/1472-6963-10-346

16. Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. Utilising identifier error variation in linkage of large administrative data sources. BMC Medical Research Methodology. 2017 Feb 7;17(1):23. https://doi.org/10.1186/s12874-017-0306-8

17. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: the Avon

Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013 Feb;42(1):97–110. https://doi.org/10.1093/ije/dys066

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

19. Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software. 2016 Oct 28;74:1–26. https://doi.org/10.18637/jss.v074.i11

20. Linacre R, Lindsay S, Manassis T, Slade Z, Hepworth T. Splink: Free software for probabilistic record linkage at scale. International Journal of Population Data Science [Internet]. 2022 Aug 25 [cited 2023 Jun 5];7(3). Available from: https://ijpds.org/article/view/1794. https://doi.org/10.23889/ijpds.v7i3.1794

21. Piantadosi ST. Zipf's word frequency law in natural language: A critical review and future directions. Psychon Bull Rev. 2014 Oct 1;21(5):1112–30. https://doi.org/10.3758/s13423-014-0585-6

22. CLOSER-resource-NHS-Numbers-and-their-management-systems.pdf [Internet]. [cited 2024 Jan 3]. Available from: https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-NHS-Numbers-and-their-management-systems.pdf.

23. Office for National Statistics. Baby names in England and Wales statistical bulletins. [cited 2024 Jan 3]. Office for National Statistics. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/previousReleases.

24. National Records of Scotland. National Records of Scotland. National Records of Scotland; 2013 [cited 2024 Jan 3]. National Records of Scotland. Available from: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/names/babies-first-names/

25. Betebenner DW. randomNames: Function for Generating Random Names and a Dataset. [Internet]. 2021. Available from: https://cran.r-project.org/package=randomNames.

26. McElduff F, Mateos P, Wade A, Borja MC. What's in a name? The frequency and geographic distributions of UK surnames. Significance. 2008;5(4):189–92. https://doi.org/10.1111/j.1740-9713.2008.00332.x

27. Danesh J, Gault S, Semmence J, Appleby P, Peto R. Postcodes as useful markers of social class: population based study in 26 000 British households. BMJ. 1999 Mar 27;318(7187):843–5. https://doi.org/10.1136%2Fbmj.318.7187.843

28. Ministry of Housing, Communities & Local Government. GOV.UK. 2019 [cited 2024 Jan 3]. English indices of deprivation. Available from: https://www.gov.uk/government/collections/english-indices-of-deprivation

29. Office for National Statistics. Ethnic Groups by Borough - London Datastore [Internet]. [cited 2024 Jan 3]. Available from: https://data.london.gov.uk/dataset/ethnic-groups-borough

30. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. Gebhardt S, editor. PLoS ONE. 2016 Oct 20;11(10):e0164667. https://doi.org/10.1371/journal.pone.0164667

31. Damerau FJ. A technique for computer detection and correction of spelling errors. Commun ACM. 1964 Mar 1;7(3):171–6. https://doi.org/10.1145/363958.363994

32. Pollock JJ, Zamora A. Automatic spelling correction in scientific and scholarly text. Commun ACM. 1984 Apr;27(4):358–68. https://doi.org/10.1145/358027.358048

33. Thomas N. Herzog , Fritz J. Scheuren , William E. Winkler. Data Quality and Record Linkage Techniques [Internet]. New York, NY: Springer; 2007 [cited 2024 Jan 3]. Available from: http://link.springer.com/10.1007/0-387-69505-2. https://doi.org/10.1007/0-387-69505-2

34. Kukich K. Techniques for automatically correcting words in text. ACM Comput Surv. 1992 Dec;24(4):377–439. https://doi.org/10.1145/146370.146380

35. Black PE. Dictionary of Algorithms and Data Structures. NIST [Internet]. 1998 Oct 1 [cited 2024 Jan 3]; Available from: https://www.nist.gov/publications/dictionary-algorithms-and-data-structures.

36. Odell M.K. The profit in records management. Systems (New York). 1956; 20(20).

37. Holmes D, McCabe MC. Improving precision and recall for Soundex retrieval. In: Proceedings International Conference on Information Technology: Coding and Computing [Internet]. Las Vegas, NV, USA: IEEE Comput. Soc; 2002 [cited 2024 Jan 3]. p. 22–6. Available from: http://ieeexplore.ieee.org/document/1000354/. https://doi.org/10.1109/ITCC.2002.1000354

38. Cheriet M, Kharma N, Liu CL, Suen C. Character Recognition Systems: A Guide for Students and Practitioners. John Wiley & Sons; 2007.

39. Gambaro L, Joshi H. Moving home in the early years: what happens to children in the UK? Longitudinal and Life Course Studies. 2016 Jul 18;7(3):265–87. https://doi.org/10.14301/llcs.v7i3.375

40. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number:

possibilities and pitfalls in healthcare and medical research. Eur J Epidemiol. 2009 Nov 1;24(11):659–67. https://doi.org/10.1007/s10654-009-9350-y

41. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. PLOS ONE. 2015 Aug 24;10(8):e0136179. https://doi.org/10.1371/journal.pone.0136179

42. Fellegi IP, Sunter AB. A Theory for Record Linkage. Journal of the American Statistical Association. 1969;64(328):1183–210.

43. Roger Eckhardt. Stan Ulam, John von Neumann and the Monte Carlo Method. Los Alamos Science. 1987;100(15):131.

44. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020 Feb 17;47(2):218–26. https://doi.org/10.1080%2F03014460.2020.1742379

45. Doidge JC, Morris JK, Harron KL, Stevens S, Gilbert R. Prevalence of Down's Syndrome in England, 1998–2013: Comparison of linked surveillance data and electronic health records. International Journal of Population Data Science [Internet]. 2020 Mar 19 [cited 2023 Jun 11];5(1). Available from: https://ijpds.org/article/view/1157. https://doi.org/10.23889/ijpds.v5i1.1157

46. Nanayakkara C, Christen P, Ranbaduge T. An Anonymiser Tool for Sensitive Graph Data. InCIKM (workshops) 2020.

Appendix 1: Estimating identifier error rates in ALSPAC

Methods

To estimate rates of identifier errors and their relationships with attribute variables, we used the ALSPAC administrative database (ARCADIA) as a 'gold standard', given that it is the product of 30 years of intensive management and cleaning. ARCADIA contains up to date information for each participant (including multiple recording of names, e.g. middle name and 'known as', and postcodes as these were changed over time) and also contains the participant ID numbers used by ALSPAC to internally link different data collection waves together. For this study, we compared the identifiers recorded in ARCADIA with those recorded at two data collection waves ('CHDB' and 'PEARL', see Appendix Table 1).

We categorised identifier errors as occurring when there was disagreement between data sources. We refer to these disagreements as errors, whilst recognising that in some cases these will be genuine changes (e.g. for addresses or surnames) rather than errors in recording.

First, we performed some minimal data cleaning on the identifiers, so that obvious differences in formatting between data collection waves were not considered as errors:

- Name: Change to upper case and remove instances of ".", "-", "'" and unnecessary spaces.
- Postcode: Change to upper case and remove internal spaces.
- Sex: Code as a binary variable (only two values for sex were observed in the data)
- Date of birth: Format all dates as dd/mm/yyyy.

We then separately analysed error rates for surname, first name, date of birth, postcode, sex, and mother's surname, comparing data from each data collection wave with ARCADIA.

We estimated error rates stratifying by additional attribute variables: ethnicity, maternal age, sex and deprivation. We derived the level of deprivation from quintiles of Indices of Multiple Deprivation (IMD) which are routinely generated for the UK using census and local authority data. IMD is assigned to participants at postcode level based on participant address, where 1 represents the most deprived area and 5 is most affluent (only evaluated for postcodes within the Avon area).

To evaluate the associations between these predictors and identifier errors, we created a logistic regression model, with identifier error as the outcome, adjusting simultaneously for ethnicity, maternal age, sex and deprivation. This model was based on data only from the CHDB.

Overall, there were very few errors in date of birth and sex (Appendix Table 2).

Postcodes

For the one data collection wave where postcode was available (PEARL), errors occurred in 36% of records. These data were collected when participants were aged between 18 and 24 years, and errors were likely due to genuine address changes rather than data recording: for the majority of discrepancies (70%, see Appendix Table 3), the correct postcode was recorded at a later date in ARCADIA. Of these probable moves, around a quarter retained the same postcode district (first 3 or 4 characters of the full postcode). Of the remaining 30% for whom the correct postcode was not recorded at a later date, the majority (55%) of disagreements were due to single character substitutions, insertions, transpositions or omissions (e.g. recording "L" instead of "1", or "6" instead of "8"). Around 5% were due to incorrectly formatted postcodes (including those from foreign addresses).

Errors in postcodes were clearly related to maternal age (postcodes for younger mothers were more likely to change), were more common for females versus males, and more likely to occur in Black or Asian ethnic groups than Whites (Appendix Table 4, 5). Those living in the most deprived areas were more likely to have errors in postcode (Appendix Table 4).

Names

Overall, recording of G0 surname was more likely to be affected by errors than G1 surname (13% versus 10%, Appendix Table 2). We also observed that errors in G0 surname were much more common in G0 women aged $<$20 years compared with older G0 women (Appendix Table 4), which may be related to changes in name following marriage. Errors or changes in the G1 surname were also related to G0 maternal age, which could indicate that both G1 and G0 surnames were changed following marriages. This was supported by further exploration of one dataset (CHDB), which revealed that G1 surname was more likely to change if G0 surname had also changed (13.6% errors in G1 surname if there was an error in G0 surname, compared with 3.7% if there was no error in G0 surname). G1 surname errors were much more likely to occur in G1 females (10.1%) than in G1 males (4.4%, Appendix Table 4).

G1 forename contained more errors than G1 surname (10% versus 7%, Appendix Table 2). Of the 1569 forename errors comparing ARCADIA with CHDB, the majority (73%) were due to nicknames or shortened name variants (e.g. Sam for Samuel, Becky for Rebecca; Appendix Table 3). The remainder were typographical errors (14%, e.g. William versus Willlam), errors due to recording of single versus multiple first names (8%, e.g. Lisa versus Lisa Marie), swapping of first and middle names (3%), or completely different names (3%).

Those living in the most deprived areas were more likely to have errors in G0 and G1 surname. However, for G1 forename, the pattern was reversed: those in more affluent areas were more likely to have errors.

Errors in identifiers were not independent: the probability of a postcode error increased from 36% to 45% if there had also been an error in surname.

Error co-occurring patterns in names

We found 0.32% of records with errors in all G1 forename, G0 and G1 surname, 0.43% of records with errors in G1 forename and G1 surname, 1.94% of records with errors in G0 and G1 surname, and 1.93% of records with errors in G1 forename and G0 surname. Given the small number of co-occurring errors, we did not explore the type of these errors.

Appendix Table 1: Data collection waves in the ALSPAC extract

Dataset	Method of recording	Age at data collection
ARCADIA (gold-standard)	The master study administrative database containing continually updated records (i.e. the 'gold-standard' of participants' identifiers).	Ongoing
The local Child Health database (CHDB)	ALSPAC received an extract of patient identifiers from the Child Health Database (CHDB) when participants were aged 5–7 years old. The CHDB was an electronic database maintained by the regional NHS for the administration of Child Health services (e.g. school-based health checks and immunisations). The CHDB record was established from birth records and then maintained by the NHS. The records were linked to ALSPAC using the internal CHDB patient ID number ('SYSNUM') which had been linked to the ALSPAC administrative database at the time of birth by trained operators using daily birth notification records [1]. The identifiers from this CHDB extract have been filtered to exclude information on ALSPAC participants who have subsequently objected to the study's use of their linked NHS records.	Extract captured at index child age between 5 and 7
Pearl: Identifiers from the 'PEARL' record linkage consent forms	The Project to Enhance ALSPAC through Record Linkage (PEARL) is a Wellcome Trust funded study that aims to develop generalizable methods for cohort studies to incorporate routine records into study databanks using data linkage techniques. The identifiers from the PEARL record linkage onsent forms were scanned and input into electronic records using OCR (using the OpenText Teleform system) ith manual review of all values exceeding an uncertainty threshold determined by the system.	Extract captured at index child age between 18 and 24

Appendix Table 2: Identifier error rates, comparing gold-standard ARCADIA data with identifiers captured in CHDB and PEARL

	G1: Child					G1: Mother
	Surname	Forename	Postcode	Sex	Date of birth	Mother's surname
Data collection wave	% (n errors/total)	% (n errors/total)	% (n errors/total)	% (n errors/total)	% (n errors/total)	% (n errors/total)
CHDB (total = 17,086)	5.1 (878/17,086)	9.3 (1569/16,905)	–	<0.1 (9/17,086)	0.1 (9/17,086)	14.7 (2507/13,086)
PEARL (total = 5680)	8.1 (459/5675)	16.1 (914/5680)	36.3 (1733/4769)	–	–	–

The denominator is the number of records with at least one completed value for each identifier.

Appendix 2: Generating synthetic names

Synthesising names that retain dependency with other variables is not straightforward. We outlined the considerations in the main article, some of which are unique to the current dataset.

For this study, we decided to use a 1:1 direct replacement of names from an existing dictionary that preserves name-sex and name-ethnicity relationship, and ordering of name frequency. The process of the name synthesis is divided into the following steps.

Step 1: Assigning ethnicity to names

We used ONS released baby forename and surname lists ordered by frequency from 1996 to 2021, and established a name dictionary. The forename list was separated by sex, but neither list provided ethnicity information. We could not find publicly available lists of names according to ethnic group in the UK.

To retain dependency between names and ethnicity, we used the NamePrism API to prescribe ethnicity based on forenames and surnames separately [2]. NamePrism is a name-based classifier that is trained on 74 million labelled name sets, developed in the United States [2]. NamePrism provides the likelihood of a certain name being correctly classified as White, Black, Asian and Pacific Islander (API), American Indian and Alaska Native (AIAN) or Hispanic. To match the ethnicity terminology used in ALSPAC, I grouped AIAN and Hispanic to "Other", and renamed API as "Asian". The ethnic group with the highest likelihood for each name was taken.

From step 1, we assigned an ethnicity to each name in the name list.

Appendix Table 3: Rates of identifier errors and relationship with attribute variables comparing ARCADIA and CHDB (names) ' and Pearl (postcode)

	G1: Child			G0: Mother
	Errors in Surname	Errors in Forename	Errors in Postcode	Errors in Mother's Surname
	% (n/total)	% (n/total)	% (n/total)	% (n/total)
Maternal Age				
<20	11.1 (62/561)	6.6 (37/561)	74.7 (59/79)	36.7 (206/561)
20–29	6.0 (431/7241)	11.5 (833/7241)	37.5 (812/2168)	19.7 (1424/7241)
30–39	4.2 (194/4604)	14.5 (668/4604)	29.6 (582/1969)	12.0 (552/4604)
40+	5.8 (9/154)	15.6 (24/154)	26.1 (26/72)	13.6 (21/154)
Missing	7.5 (54/721)	11.5 (83/721)	50.6 (222/439)	16.5 (119/721)
Ethnic group[1]				
White	5.6 (601/10756)	13.0 (1398/10756)	33.6 (1367/4067)	17.5 (1878/10756)
Black	8.8 (11/125)	13.6 (17/125)	60.9 (14/23)	22.4 (28/125)
Asian	<5% (<5/105)	13.3 (14/105)	45.8 (11/24)	8.6 (9/105)
Other	12.2 (9/74)	14.9 (11/74)	31.6 (6/19)	18.9 (14/74)
Missing/Withdrawn	5.7 (127/2221)	9.2 (205/2221)	51.0 (303/594)	17.7 (393/2221)
Sex				
Female	10.4 (535/6498)	10.0 (652/6498)	39.4 (1106/2811)	18.0 (1169/6498)
Male	3.2 (215/6783)	14.6 (993/6783)	31.0 (595/1916)	17.0 (1153/6783)
Index of Multiple Deprivation quintile[2]				
Most deprived	6.7 (128/1904)	8.0 (153/1904)	43.7 (153/350)	21.3 (405/1904)
2	5.3 (92/1740)	9.8 (171/1740)	40.7 (190/467)	20.7 (360/1740)
3	5.7 (101/1785)	11.2 (199/1785)	38.2 (225/589)	19.2 (342/1785)
4	4.9 (125/2544)	12.3 (313/2544)	32.1 (328/1021)	15.6 (398/2544)
Most affluent	5.0 (156/3093)	13.6 (419/3093)	29.2 (413/1414)	13.9 (431/3093)
Outside Avon/Missing	6.7 (148/2215)	17.6 (390/2215)	44.2 (392/886)	17.4 (386/2215)

[1]Asian: Bangladeshi, Chinese, Indian, Pakistani; Black: Black African, Black Caribbean, Other Black; [2]IMD only evaluated for postcodes within the Avon area.
Records with no attribute data were excluded. Denominator N is the number of records with a completed value for each identifier.

Step 2. Creating name dictionaries

Name lists from ONS were deduplicated by gender and ethnicity, and across forenames and surnames. To avoid names in original data appearing in the synthesised dataset, all terms appearing in the original data were removed from the forename and surname lists. For co-occurring forenames across male and female, duplicated names were removed from the female forename list since there were more female names than male names in the ONS forename lists (Female = 21,958, Male = 16,777), leaving 19,634 unique female forenames. For co-occurring terms across forenames and surnames (for example, Woods is used both as a surname and a forename), duplicates terms were removed from the surname list, leaving 8,395 unique surnames. From the above processes, we created a unique male forename list, female forename list, and surname list. For names that had a missing ethnicity, replacement names were drawn from "White" ethnic group that is the least common (occurred once) in the ONS lists.

For names that co-occurred across gender or ethnicity, the combination with the highest frequency was retained. Forenames were then ranked by gender and ethnicity, and surnames ranked by ethnicity. In the original data, individuals may have provided multiple surnames and forenames. For example, mother's surname (g0_surname) often contains multiple terms, with one of them duplicating the child's surname (g1_surname). This is likely due to the mothers including the fathers' surname in the data. We split all names by spaces, such that all terms would be taken into consideration for term frequency. This meant that in cultures where surnames are changed after marriage, their surnames (father's surname) would be double-counted in g0_surname and g1_surname, hence strengthening certain ethnicity-name associations and over-representing ethnically ambiguous names as "White".

Step 3. Combing synthesised names with synthetic data

Synthetic data were created using the R package Synthpop [3]. Synthetic data created in Synthpop does not follow a 1:1 structure to the original data. The distribution of people of different gender and ethnicity varies across the synthesised datasets. The gender-ethnicity matched names created in the data dictionary cannot fully match all datasets. However, to retain the cardinality and uniqueness of the name variables, we decided not to further sample new names that would fit the gender-ethnicity association for each sample. We used the same set of synthesised names for all synthesised datasets, matching with gender and ethnicity where possible, and inspected the average mismatch by gender and ethnicity.

Appendix Table 4: Relative risk of identifier errors according to attribute variables (N = 14,142 records)

	G1: Child			G0: Mother
	ESurname	**Forename**	**Postcode**	**Mother's Surname**
	Relative risk (95% CI)	Relative risk (95% CI)	Relative risk (95% CI)	Relative risk (95% CI)
Maternal Age				
<20	2.60 (1.96, 3.44)	0.57 (0.42, 0.79)	2.43 (2.11, 2.81)	2.89 (2.51, 3.32)
20–29	1.43 (1.21, 1.69)	0.85 (0.77, 0.94)	1.26 (1.16, 1.38)	1.61 (1.47, 1.77)
30–39	Reference	Reference	Reference	Reference
40+	1.44 (0.76, 2.74)	1.09 (0.75, 1.58)	1.24 (0.91, 1.69)	1.14 (0.76, 1.71)
Missing	2.13 (1.48, 3.08)	1.24 (0.94, 1.62)	1.68 (1.50, 1.89)	1.50 (1.21, 1.84)
Ethnic group[1]				
White	Reference	Reference	Reference	Reference
Black	1.43 (0.81, 2.51)	1.19 (0.76, 1.85)	1.76 (1.27, 2.43)	1.05 (0.76, 1.45)
Asian	0.34 (0.09, 1.33)	0.99 (0.61, 1.61)	1.39 (0.90, 2.14)	0.46 (0.25, 0.86)
Other	2.05 (1.12,3.74)	1.13 (0.66, 1.95)	0.91 (0.47, 1.77)	1.08 (0.68, 1.71)
Missing/Withdrawn	0.77 (0.61, 0.98)	0.72 (0.60, 0.86)	1.49 (1.36, 1.63)	0.89 (0.79, 1.00)
Sex				
Female	2.55 (2.19, 2.98)	0.68 (0.62, 0.75)	1.27 (1.17, 1.37)	1.05 (0.98, 1.13)
Male	Reference	Reference	Reference	Reference
Index of Multiple Deprivation quintile[2]				
Most deprived	1.13 (0.90, 1.43)	0.66 (0.55, 0.79)	1.46 (1.27, 1.69)	1.27 (1.12, 1.45)
2	0.96 (0.75, 1.24)	0.77 (0.65, 0.91)	1.38 (1.20, 1.58)	1.32 (1.17, 1.51)
3	1.03 (0.81, 1.32)	0.86 (0.73, 1.01)	1.29 (1.13, 1.47)	1.28 (1.12, 1.45)
4	0.94 (0.75, 1.18)	0.93 (0.81, 1.06)	1.10 (0.98, 1.24)	1.08 (0.96, 1.23)
Most affluent	Reference	Reference	Reference	Reference
Outside Avon/Missing	1.29 (1.04, 1.61)	1.30 (1.15, 1.48)	1.50 (1.35, 1.68)	1.22 (1.07, 1.38)

[1] Asian: Bangladeshi, Chinese, Indian, Pakistani; Black: Black African, Black Caribbean, Other Black; [2]IMD only evaluated for postcodes within the Avon area.
Estimates for name are adjusted for all variables in the table; estimates for postcode are adjusted only for sex (due to small numbers of records with postcode available).

Gender and ethnicity average mismatch rates are less than 1% across all datasets. Gold-standard synthetic ALSPAC data with identifier and attribute variables were produced. No forename-lastname combinations in the original data is present in the synthetic data.

Limitations of using NamePrism

The NamePrism ethnicity classifier is trained using mainly United States data and race categories. Race and ethnicity terms are country and context specific and should not be used interchangeably. Naming practices also vary across countries and regions. Names more frequently associated with certain populations in the United States may not hold the same association in the UK. Our current approach has the risk of inducing and reducing name-ethnicity associations in the ALSPAC cohort.

However, we estimate that the extent of the impact would be rather limited. ALSPAC has a predominantly White cohort, with predominantly anglicised European names. There is a shared cultural naming heritage between White British and White north Americans. Future studies could be improved by using a name dictionary that is properly labelled with ethnicity. We sent a data request to the ONS Census team for an ethnicity labelled name dictionary, with frequency, but our request was not approved due to confidentiality concerns.

R and Python codes used to synthesise names, identifiers and attributes are available here on GitHub: https://github.com/UCL-CHIG/ALSPAC_synthetic_identifiers.

Appendix 3: Generating synthetic identifiers and attributes

Synthpop utilises sequential imputation models for data synthesis. The order of included variables depends on the level of completeness of the variable. We included identifier variables (gender, date of birth), along with attribute variables (ethnicity, index of multiple deprivation, maternal age), in the following sequence:

Gender, date of birth, maternal age, ethnicity, index of multiple deprivation.

We implemented a rejection sampling mechanism to ensure synthesised dataset did not generate combinations of attribute variables that did not appear in the original study. Synthesised attribute and identifiers were appended with synthesised names.

Appendix Table 5: Distribution of attribute characteristics used to generate the synthetic data, based on aggregate data from ALSPAC

Attribute variable	% of records
Sex	
Female	51.1
Male	48.9
Index of Multiple Deprivation quintile	
Most deprived	14.3
2	13.1
3	13.4
4	19.2
Most affluent	23.3
Outside Avon/Missing	16.7

Appendix Table 6: Joint distribution of maternal age and ethnic group used to generate the synthetic data, based on aggregate data from ALSPAC

	Ethnic group				
	White	**Black**	**Asian**	**Other**	**Missing**
Maternal Age					
<20	69.0	2.0	0.5	0.5	28.0
20–29	84.0	1.0	1.0	0.5	14.0
30–39	90.0	1.0	1.0	0.5	7.5
40+	86.5	0.0	0.5	0.5	12.5
Missing	0.0	0.0	0.0	0.0	100.0

Figures are rounded to prevent statistical disclosure of small numbers.

Appendix 4: Data Linkage settings

U-probabilities were estimated using random sampling. M-probabilities for name variables were estimated from labelled data. The M-probability for date of birth and gender were set at 0.999. Prior match weights were calculated from the probability that 2 records drawn at random were a match, in the 13,281 records pairs, which is equivalent to a starting matching weight of -13.697. JW refers to Jaro-Winkler similarity scores. When using Jaro-Winkler similarity scores to compare names, we categorised outcomes into three categories, and m- and u-probabilities were derived for each of these score categories: 0–<0.8, 0.8–<1, and exact match (Appendix Table 7).

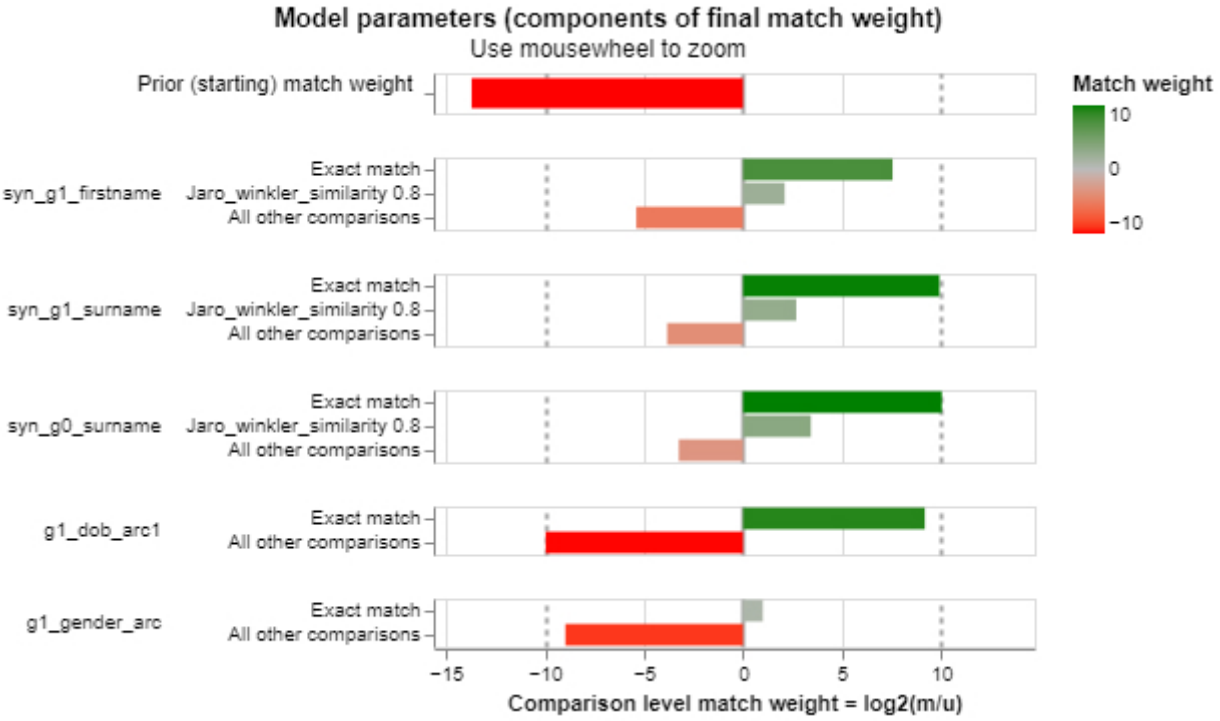
For example, a Jaro-Winkler score of <0.8 when comparing surname would have a m-probability of 0.051 and a u-probability of 0.995; a score of 0.8–<1 would have a m-probability of 0.01 and a u-probability of 0.004, and a score of 1 (exact agreement) would have a m-probability of 0.9435 and a u-probability of 0.001.

Figure A8 is an illustrative depiction of the model parameters and match weights for each variable at each comparison level.

Appendix Table 7: Weights used in probabilistic linkage

	m-probability		u-probability		Identifier match weight	
	Agreement	**Disagreement**	**Agreement**	**Disagreement**	**Agreement**	**Disagreement**
Surname (exact)	0.905	0.095	0.001	0.999	9.91	-
Surname (JW $\geq$ 0.8)	0.025	0.975	0.004	0.996	2.69	-
Surname (else)	0.070	0.930	0.995	0.005	-3.83	-
Forename(exact)	0.905	0.095	0.005	0.995	7.55	-
Forename (JW $\geq$ 0.8)	0.027	0.973	0.006	0.994	2.10	-
Forename (else)	0.024	0.976	0.988	0.012	-5.39	-
Sex	0.999	0.001	0.5000	0.5000	1.0	-8.97
Date of Birth	0.999	0.001	0.0015	0.9985	9.17	-9.96
Mother's surname (exact)	0.857	0.143	0.0008	0.9992	10.0	-
Mother's surname (JW $\geq$ 0.8)	0.383	0.617	0.0036	0.9964	3.42	-
Mother's surname (else)	0.105	0.895	0.996	0.004	-3.24	-

Figure A8: Model Parameters and match weights



Appendix Table 8: False matches, missed matches in original linkage, scenario 1 and 3

Scenario	Data	Type	Number false matches	Number missed matches	Pattern	Count per pattern	Proportion
–	Original	probabilistic	15	346	10011	225	0.65
					00111	47	0.14
					00011	33	0.10
					01011	23	0.07
					01110	8	0.02
					01101	4	0.02
					10110	3	0.01
					11010	1	0.00
					11011	1	0.00
					01001	1	0.00
–	Original	term frequency adjustment	6	319	10011	198	0.62
					00111	44	0.14
					00011	32	0.10
					01011	21	0.07
					01110	8	0.03
					11110	6	0.02
					01101	4	0.01
					10110	3	0.01
					11010	1	0.00
					11011	1	0.00
					01001	1	0.00
1	1	probabilistic	37	467	10011	283	0.61
					00111	93	0.20
					01011	79	0.17
					01110	8	0.02
					01101	2	0.00
					00011	1	0.00
					11110	1	0.00
1	1	term frequency adjustment	24	474	10011	272	0.57
					00111	95	0.20
					01011	80	0.17
					01110	15	0.03
					11110	10	0.02
					01101	1	0.00
					00011	1	0.00
1	2	probabilistic	34	463	10011	280	0.60
					00111	90	0.19
					01011	79	0.17
					01110	8	0.02
					01101	3	0.01
					11110	3	0.01
1	2	term frequency adjustment	15	473	10011	274	0.58
					00111	92	0.19
					01011	79	0.17
					11110	13	0.03
					01110	11	0.02
					01101	4	0.01

Continued

Appendix Table 8: Continued

Scenario	Data	Type	Number false matches	Number missed matches	Pattern	Count per pattern	Proportion
1	3	probabilistic	31	464	10011	285	0.61
					00111	87	0.19
					01011	79	0.17
					01110	7	0.02
					01101	3	0.01
					11110	3	0.01
1	3	term frequency adjustment	16	471	10011	278	0.59
					00111	88	0.19
					01011	81	0.17
					11110	12	0.03
					01110	10	0.02
					01101	2	0.00
1	4	probabilistic	27	465	10011	285	0.61
					00111	92	0.20
					01011	76	0.16
					01110	7	0.02
					11110	3	0.01
					01101	2	0.00
1	4	term frequency adjustment	13	473	10011	276	0.58
					00111	94	0.20
					01011	78	0.16
					11110	13	0.03
					01110	10	0.02
					01101	2	0.00
1	5	probabilistic	27	460	10011	287	0.62
					00111	85	0.18
					01011	77	0.17
					01110	8	0.02
					11110	3	0.01
1	5	term frequency adjustment	14	467	10011	279	0.60
					00111	86	0.18
					01011	78	0.17
					01110	12	0.03
					11110	12	0.03
3	1	probabilistic	48	578	10011	257	0.44
					00111	221	0.38
					01011	44	0.08
					00011	37	0.06
					01110	17	0.03
					11010	1	0.00
					10110	1	0.00
3	1	term frequency adjustment	26	586	10011	256	0.44
					00111	222	0.38
					01011	42	0.07
					00011	36	0.06
					01110	19	0.03
					11110	7	0.01
					01101	2	0.00
					11010	1	0.00
					10110	1	0.00

Continued

Appendix Table 8: Continued

Scenario	Data	Type	Number false matches	Number missed matches	Pattern	Count per pattern	Proportion
3	2	probabilistic	40	579	10011	259	0.45
					00111	219	0.38
					01011	44	0.08
					00011	32	0.06
					01110	22	0.04
					01101	2	0.00
					10110	1	0.00
3	2	term frequency adjustment	19	594	10011	256	0.43
					00111	219	0.37
					01011	39	0.07
					00011	30	0.05
					01110	29	0.05
					11110	16	0.03
					01101	4	0.01
					10110	1	0.00
3	3	probabilistic	28	534	10011	251	0.47
					00111	204	0.38
					01011	37	0.07
					00011	24	0.04
					01110	17	0.03
					11010	1	0.00
3	3	term frequency adjustment	8	546	10011	247	0.45
					00111	207	0.38
					01011	37	0.07
					00011	23	0.04
					01110	20	0.04
					11110	10	0.02
					01101	1	0.00
					11010	1	0.00
3	4	probabilistic	31	557	10011	255	0.46
					00111	214	0.38
					01011	41	0.07
					00011	26	0.05
					01110	17	0.03
					01101	3	0.01
					00110	1	0.00
3	4	term frequency adjustment	15	564	10011	248	0.44
					00111	214	0.38
					01011	41	0.07
					01110	26	0.05
					00011	22	0.04
					11110	7	0.01
					01101	4	0.01
					10001	1	0.00
					00110	1	0.00

Continued

Appendix Table 8: Continued

Scenario	Data	Type	Number false matches	Number missed matches	Pattern	Count per pattern	Proportion
3	5	probabilistic	56	569	00111	209	0.37
					10011	175	0.31
					01110	102	0.18
					11100	36	0.06
					01011	21	0.04
					00011	12	0.02
					01100	4	0.01
					10010	3	0.01
					10110	2	0.00
					00110	2	0.00
					01010	2	0.00
					01001	1	0.00
3	5	term frequency adjustment	35	553	00111	211	0.38
					01110	146	0.26
					10011	137	0.25
					01011	19	0.03
					00011	14	0.03
					11100	11	0.02
					01100	4	0.01
					10110	2	0.00
					00110	2	0.00
					01101	2	0.00
					01010	2	0.00
					10010	1	0.00
					01001	1	0.00
					11110	1	0.00

Matching pattern corresponds to:
Forename, mother's surname, surname, gender, date of birth.

Appendix Table 9: False matches in deterministic linkages, original and all scenarios

Scenario	Data	Type	Number false matches	Number missed matches	Pattern	Count per pattern	Proportion
–	original	deterministic	28	01111	15	0.54	
				11110	11	0.39	
				11111	1	0.04	
				10111	1	0.04	
1	1	deterministic	45	01111	27	0.60	
				11110	18	0.40	
1	2	deterministic	42	11110	29	0.69	
				01111	13	0.31	
1	3	deterministic	33	01111	18	0.55	
				11110	15	0.45	
1	4	deterministic	30	01111	18	0.60	
				11110	11	0.37	
				11111	1	0.03	
1	5	deterministic	36	11110	18	0.50	
				01111	17	0.47	
				11111	1	0.03	
2	1	deterministic	14	01111	14	1.00	
2	2	deterministic	16	01111	16	1.00	
2	3	deterministic	12	01111	12	1.00	
2	4	deterministic	12	01111	11	0.92	
				11111	1	0.08	
2	5	deterministic	9	01111	8	0.89	
				11111	1	0.11	
3	1	deterministic	38	01111	20	0.53	
				11110	17	0.45	
				10111	1	0.03	
3	2	deterministic	34	01111	20	0.59	
				11110	14	0.41	
3	3	deterministic	31	01111	17	0.55	
				11110	14	0.45	
3	4	deterministic	27	11110	14	0.52	
				01111	12	0.44	
				11111	1	0.04	
3	5	deterministic	31	01111	17	0.55	
				11110	14	0.45	
4	1	deterministic	14	01111	14	1.00	
4	2	deterministic	18	01111	18	1.00	
4	3	deterministic	13	01111	13	1.00	
4	4	deterministic	10	01111	9	0.90	
				11111	1	0.10	
4	5	deterministic	8	01111	7	0.88	
				11111	1	0.13	

Matching pattern corresponds to:
Forename, mother's surname, surname, gender, date of birth.

References for Appendix

1. Mummé M, Boyd A, Golding J, Macleod J. The STORK dataset: Linked midwifery and delivery records of the mothers and index children in the Avon Longitudinal Study of Parents and Children (ALSPAC). Wellcome Open Res. 2020;5:229. https://doi.org/10.12688/wellcomeopenres.16247.1

2. Ye J, Han S, Hu Y, Coskun B, Liu M, Qin H, et al. Nationality Classification Using Name Embeddings. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management [Internet]. New York, NY, USA: Association for Computing Machinery; 2017 [cited 2024 Jan 2]. p. 1897–906. (CIKM '17). Available from: https://dl.acm.org/doi/10.1145/3132847.3133008. https://doi.org/10.1145/3132847.3133008

3. Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software. 2016 Oct 28;74:1–26. https://doi.org/10.18637/jss.v074.i11