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).

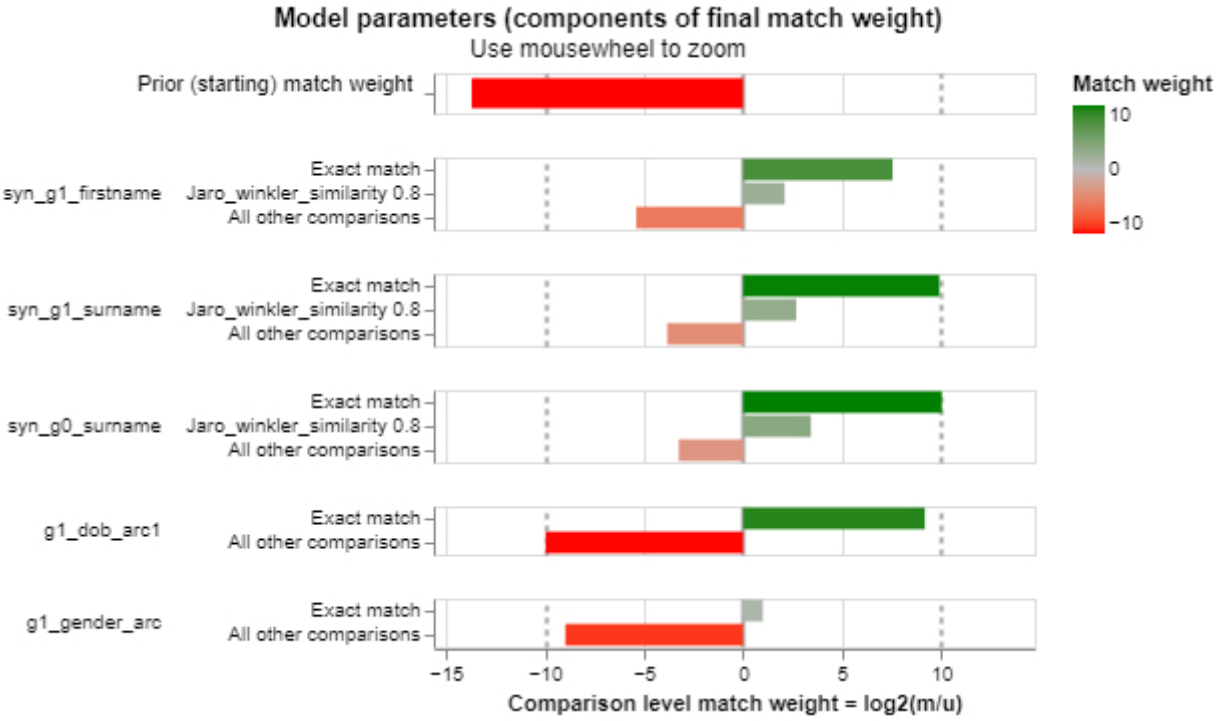
For example, a Jaro-Winkler score of <0.8 when comparing surname would have a m-probability of 0.051 and a u-probability of 0.995; a score of 0.8–<1 would have a m-probability of 0.01 and a u-probability of 0.004, and a score of 1 (exact agreement) would have a m-probability of 0.9435 and a u-probability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