This report summarises the outputs from the matching of the Service Leavers dataset to the NRS Population Spine. Outcomes will be used for research projects obtaining the relevant approvals.

NRS Indexing provides a trusted third party indexing service for linkage projects in Scotland. You can find more information about the team on our website (https://www.nrscotland.gov.uk/statistics-and-data/national-records-of-scotland-indexing-team).

In order to facilitate de-identification and linkage NRS Indexing identifies individuals in common across datasets by matching them to the Population Spine. This is done using Personally Identifiable Information (PII). PII is Personal Data that can be used to identify a person, and it includes variables like Name, Surname, Postcode, Date of Birth, Gender, Community Health Index (CHI). NRS Indexing uses one or more of these variables to find people on the spine.

For most projects the Indexing phase of the Linkage involves:

1. Spine Matching: Finding people on the Population Spine. Datasets are matched to the spine separately. Individuals across datasets pointing to the same ID in the spine are considered to be the same person.

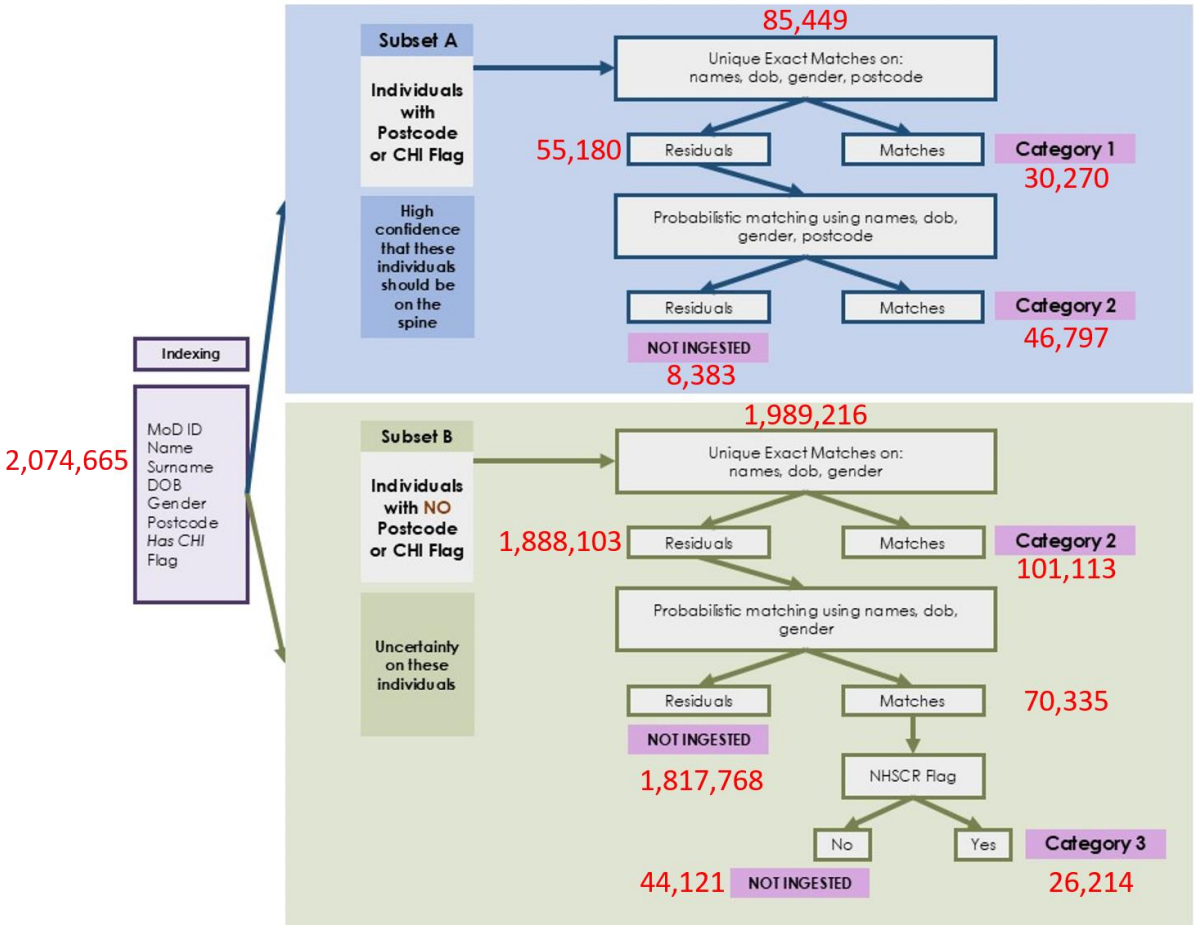
2. Indexing (de-identification): This includes creating indexes that allow the de-identified datasets to be linked.

Datasets matched to the spine covered in this report	
Service Leavers	Matched (both exact- and probabilistic-) to the spine. Details in this report: 1. Matching flow 2. Pre-processing 3. Matching 4. Bias Analyses 5. Output: Summary information for future use in linkage projects Annex 1. Matching quality (subset A) Annex 2. Matching quality (subset B) Annex 3. Probabilistic matching process and glossary

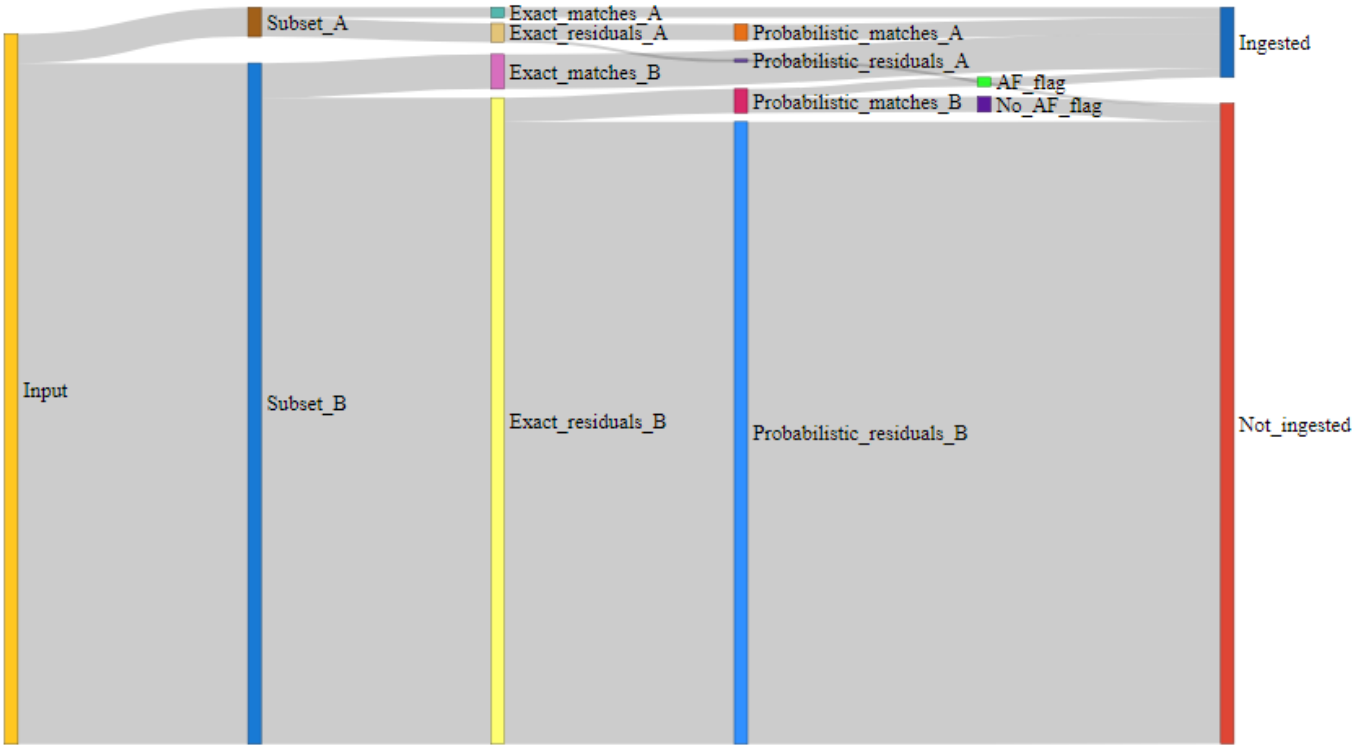
Dataset name: Service Leavers

Stage 1: Matching flow

Matching was performed following the flow represented in the diagram below. The red numbers show the number of records at each stage.



Sankey diagram showing the matching flow.



Dataset name: **Service Leavers**

Stage 2: Preprocessing

Total number of Input Records:	**2,074,665**	
Unique individuals	1,872,973	
Variable completion	**N. Records**	**%**
First name	1,754,845	84.6%
Middle name	1,182,351	57.0%
Surname	2,061,350	99.4%
Date of Birth	2,074,633	100.0%
Valid Date of Birth	2,061,350	99.4%
Sex	2,074,665	100.0%
Postcode	83,810	4.0%
Valid Postcode	82,354	4.0%
CHI flag	2,074,665	100.0%
Subset A (with a postcode or a positive CHI flag)	85,449	4.1%
Subset B (neither a postcode or a positive CHI flag)	1,989,216	95.9%
Records with completed PII	81,614	3.9%
valid DOB & filled names & sex	1,673,098	80.6%
Duplicates	0	0.0%
Total unique input records	**2,074,665**	**100.0%**

DOB = Date of Birth
Completed PII = First name, surname, date of birth, sex and postcode present

Dataset name: Service Leavers

Stage 3: Matching against the Indexing Spine

Subset A:	**N. Records**	**Ingest category**
Records which exactly match the spine on names, DOB, postcode and sex	30,270	Category 1
Records which probabilistically match the spine on names, DOB, postcode and sex	46,797	Category 2
Non-matches	8,382	Not ingested
Total	**85,449**	

Subset B:	**N. Records**	**Ingest category**
Records which exactly match the spine on names, DOB and sex	101,113	Category 2
Records which probabilistically match the spine on names, DOB and sex	70,335	
Probabilistic matches with an Armed Forces NHSCR flag	26,214	Category 3
Probabilistic matches without an Armed Forces NHSCR flag	44,121	Not ingested
Non-matches	1,817,768	Not ingested
Total	**1,989,216**	

Clerical review of probabilistic matching	**N. Unique Matches**	**Number Sampled**	**Estimated Precision**
Subset A	47,602	1,222	96.4%
Subset B	97,906	943	98.3%

Dataset name: Service Leavers

Stage 4: Bias Analyses

The tables below indicate the number (and %) of matches in different categories based on completion, gender, and postcode type

In the tables below, records for lower categories have been removed where another record for the same individual has matched to a higher category. The number of matches, and the match rates, in Categories 2 and 3 is therefore slightly lower than in previous sheets.

Data Completion

Subset A	Spine matching				
PII Completion	**Category 1**	**Category 2**	**Unmatched**	**Total**	**%Match**
Incomplete	0	2,308	1,526	3,834	60.2%
Complete	30,270	43,676	7,669	81,615	90.6%
Total	**30,270**	**45,984**	**9,195**	**85,449**	**89.2%**

Subset B	Spine matching				
PII Completion	**Category 2**	**Category 3**	**Unmatched**	**Total**	**%Match**
Incomplete	0	0	319,147	319,147	0.0%
Complete (except postcode)	99,942	24,477	1,545,650	1,670,069	7.4%
Total	**99,942**	**24,477**	**1,864,797**	**1,989,216**	**6.3%**

Gender

Subset A	Spine matching				
Gender	**Category 1**	**Category 2**	**Unmatched**	**Total**	**%Match**
Male	28,458	42,165	8,178	78,801	89.6%
Female	1,812	3,819	1,017	6,648	84.7%
Total	**30,270**	**45,984**	**9,195**	**85,449**	**89.2%**

Subset B	Spine matching				
Gender	**Category 2**	**Category 3**	**Unmatched**	**Total**	**%Match**
Male	86,122	22,304	1,681,029	1,789,455	6.1%
Female	13,820	2,173	183,768	199,761	8.0%
Total	**99,942**	**24,477**	**1,864,797**	**1,989,216**	**6.3%**

Military Postcode (Subset A only)

	Spine matching				
Military Postcode	**Category 1**	**Category 2**	**Unmatched**	**Total**	**%Match**
Yes	1,709	9,599	5,592	16,900	66.9%
No	28,561	35,596	2,753	66,910	95.9%
Postcode missing	0	789	850	1,639	48.1%
Total	**30,270**	**45,984**	**9,195**	**85,449**	**89.2%**

"Yes" indicates that the record's postcode is one of the most common in the dataset, which correspond to a military base, reserve centre or HQ. "No" is for other postcodes.

Year of birth

Subset A	Spine matching				
YOB	**Category 1**	**Category 2**	**Unmatched**	**Total**	**%Match**
before 1930	144	341	18	503	96.4%
1930 to 1939	947	2,250	107	3,304	96.8%
1940 to 1949	4,035	4,936	262	9,233	97.2%
1950 to 1959	5,560	5,380	532	11,472	95.4%
1960 to 1969	6,668	10,999	1,577	19,244	91.8%
1970 to 1979	5,834	11,523	1,967	19,324	89.8%
1980 to 1989	5,346	8,886	2,887	17,119	83.1%
1990 to 1999	1,611	1,556	1,696	4,863	65.1%
after 1999	125	113	149	387	61.5%
Total	**30,270**	**45,984**	**9,195**	**85,449**	**89.2%**

Subset B	Spine matching				
YOB	**Category 2**	**Category 3**	**Unmatched**	**Total**	**%Match**
before 1930	202	41	35,992	36,235	0.7%
1930 to 1939	989	334	80,202	81,525	1.6%
1940 to 1949	4,492	1,788	205,874	212,154	3.0%
1950 to 1959	15,706	6,554	398,426	420,686	5.3%
1960 to 1969	37,905	7,243	540,945	586,093	7.7%
1970 to 1979	20,771	4,884	296,082	321,737	8.0%
1980 to 1989	11,606	2,837	203,381	217,824	6.6%
1990 to 1999	7,287	665	89,576	97,528	8.2%
after 1999	984	131	14,287	15,402	7.2%
Total	**99,942**	**24,477**	**1,864,765**	**1,989,184**	**6.3%**

Dataset name: Service Leavers

Stage 5: Spine matching summary

Total number of Input Records:		**2,074,665**	
		N. Records	**%**
Records Matched to Population Spine	Category 1	30,270	1.5%
	Category 2	147,910	7.1%
	Category 3	26,214	1.3%
	Total	**204,394**	9.9%

Total number of unique person IDs		**1,872,973**		
		N. Person IDs	**%**	**N. Spine IDs**
Person IDs Matched to Population Spine	Category 1	26,886	1.4%	26,790
	Category 2	128,171	6.8%	126,469
	Category 3	22,195	1.2%	21,743
	Total	**177,252**	9.5%	**175,002**

Some individuals have records in multiple categories. The individual is then assigned to the highest Category.
Some spine IDs matched to multiple person IDs, suggesting that these person IDs correspond to the same individual.

Stage 5: Outputs

Individuals matched to the population spine will be ingested in the ADR infrastructure so they are linkable to other datasets in projects with the relevant approvals in place

Dataset name: Service Leavers

Annex 1: Matching quality (subset A)

Subset A: Probabilistic matching used the following hierarchical blocking criteria :

Block	Block description (exact matches on)	N. Candidate Matches	N. Unique Matches	N. Quality Matches
1	YOB, postcode area, surname, first name, middle name	5,733	5,691	5,691
2	YOB, postcode area, surname-hyphen, first name, middle initial	9,749	9,654	9,651
3	DOB, surname-hyphen, first initial, sex	31,296	30,323	30,246
4	YOB, surname, first initial, postcode initial	8,273	820	424
5	YOB, surname-hyphen, first initial, middle name, sex	668	85	59
6	YOB, surname-hyphen, first initial, middle initial	1,532	96	10
7	Postcode, surname, first name, sex	571	28	22
8	YOB, MOB, surname, first name, sex	4,601	116	8
9	MOB, postcode, surname, sex	1,366	623	552
10	DyOB, postcode, surname-hyphen, sex	332	25	6
11	DOB, previous postcode, surname initial, first initial, sex	121	107	106
12	YOB, postcode, surname first 2 letters, first name	32	16	4
13	DOB, previous postcode, first name, middle initial	20	18	18
Total		**64,294**	**47,602**	**46,797**

Clerical review results

Subset A

Block	N. Unique Matches	% Unique Matches	Number Sampled	Estimated Precision
1	5,691	12.0%	128	100.0%
2	9,654	20.3%	139	98.9%
3	30,323	63.7%	167	95.6%
4	820	1.7%	253	83.8%
5	85	0.2%	47	80.4%
6	96	0.2%	96	65.0%
7	28	0.1%	38	92.5%
8	116	0.2%	113	72.2%
9	623	1.3%	107	88.8%
10	25	0.1%	37	83.3%
11	107	0.2%	65	100.0%
12	16	0.0%	17	100.0%
13	18	0.0%	15	100.0%
All	**47,602**	**100.0%**	**1,222**	**96.4%**

DOB = Date of Birth
YOB = Year of Birth
MOB = Month of Birth
DyOB = Day of Birth

Dataset name: Service Leavers

Annex 2: Matching quality (subset B)

Subset B: Probabilistic matching used the following hierarchical blocking criteria :

Block	Block description (exact matches on)	N. Candidate Matches	N. Unique Matches	N. Quality Matches	N. Quality Matches with AF flag
1	DOB, surname, swapped middle and first names	152	152	152	48
2	DOB, surname, middle name, sex	4,230	4,134	2,173	701
3	DOB, surname, middle initial, first name	25,395	25,326	25,326	7427
4	DOB, surname, first name	29,337	28,789	27,064	9895
5	YOB, MOB, surname, middle name, first name	1,767	1,657	976	424
6	MOB, DyOB, surname, middle name, first name	1,371	1,204	549	38
7	YOB, DyOB, surname, middle name, first name	443	395	395	48
8	DOB, surname, middle initial, first initial, sex	13,542	13,437	13,437	7566
9	DOB, middle name, first name	28,718	22,812	263	67
Total		**104,955**	**97,906**	**70,335**	**26,214**

Clerical review results

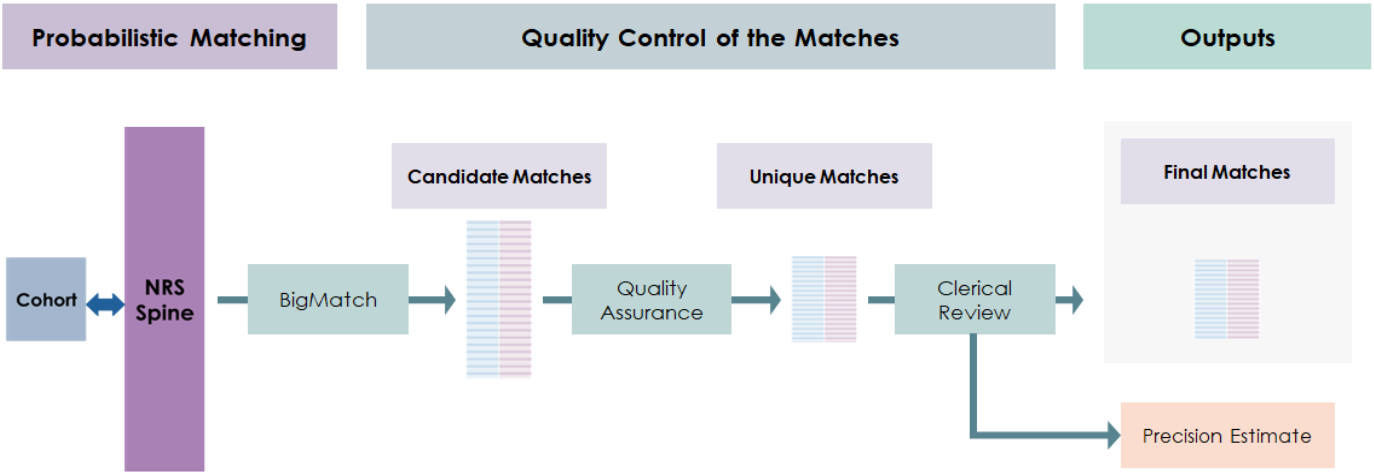
Subset B

Block	N. Unique Matches	% Unique Matches	Number Sampled	Estimated Precision
1	152	0.2%	9	100.0%
2	4,134	4.2%	120	100.0%
3	25,326	25.9%	37	100.0%
4	28,789	29.4%	113	100.0%
5	1,657	1.7%	63	100.0%
6	1,204	1.2%	79	50.0%
7	395	0.4%	36	50.0%
8	13,437	13.7%	136	99.0%
9	22,812	23.3%	350	97.5%
All	**97,906**	**100.0%**	**943**	**98.3%**

DOB = Date of Birth
YOB = Year of Birth
MOB = Month of Birth
DyOB = Day of Birth

Dataset name: Service Leavers

Annex 3: Probabilistic matching process and glossary



Probabilistic matching was carried out using BigMatch. BigMatch is a linkage software program developed by the Statistical Research Division, U.S. Bureau of Census. More information:

https://www.census.gov/library/working-papers/2002/adrm/rrc2002-01.html

The program is a linkage engine and implements traditional probabilistic record linkage methodology following the Fellegi-Sunter model for record linkage

BigMatch is designed to extract plausible matches from a large file using several blocking criteria without having to sort the file before each blocking run. Blocking is a commonly used technique in record linkage to minimise the number of comparisons between records. Records are grouped into blocks based on specified values that agree, for example instead of comparing all records, only records with the same sex are compared. Indexers select the most efficient blocks to perform the matching. This document contains results stratified by blocks. More information about blocking:

https://usc-isi-i2.github.io/papers/michelson06-aaai.pdf

Glossary

Candidate Matches	Total number of matches between external records and Spine IDs identified by BigMatch. This can include several records matching to the same Spine ID before any quality control.
Unique Matches	Total number of external records matched to the Spine after removing competing matches (ensuring only one record matches to each Spine ID)
Quality Matches	Total number of external records matched to the Spine after removing competing matches and after clerical review. For Subset A this reflects the final matches considered true.
Quality matches (with AF flag)	Total number of external records matched to the Spine after removing competing matches and after clerical review, and which have an NHSCR Armed Forces flag. For Subset B this reflects the final matches considered true.