NRS Project ID: 2324-A001
Pupil Census update 2020-2022 - data ingest

Stage 1: preprocessing - data quality of matching fields

Pupil Census records		**2,122,103**	
Pupil Census individuals		**831,412**	
valid postcode		2,122,103	100.0%
valid gender		2,122,103	100.0%
valid date of birth		2,121,783	99.98%
where year of birth < 2002	(oldest is 1900)	318	0.0%
where year of birth > 2018	(youngest is 2021)	2	0.0%
where month of birth = 1		174,568	8.2%
- January month of birth expectation based on uniform distribution across calendar year		180,110	8.5%
where day of birth = 1		70,233	3.3%
- day of birth = 1 expectation based on uniform distribution across months		69,720	3.3%
Pupil Census 2019 read-through indexes - records		1,759,417	
Pupil Census 2019 read-through indexes - individuals		**644,624**	
new records		362,686	
PII for new pupils		**205,678**	
new pupils		**186,788**	

NRS Project ID: 2324-A001
Pupil census update, new pupils 2020-2022 - data ingest

Stage 2: BigMatch linkage against the population spine

BigMatch is a linkage software program developed and used in-house by the Statistical Research Division, U.S. Bureau of Census. It has been designed to undertake timely matching of very large files (e.g. linking the US census, 300 million x 300 million).

The program is strictly a linkage engine and implements traditional probabilistic record linkage methodology.

The Bigmatch program is designed to extract plausible matches from a large file using several blocking criteria without having to sort the file before each blocking run.

Further details at https://www.census.gov/srd/papers/pdf/rrc2007-01.pdf

In this run, probabilistic weights and match categories similar to those piloted in the pupil census linkage were generated in SAS on the plausible matches generated by BigMatch - see http://www.isdscotland.org/Products-and-Services/eDRIS/Docs/20150421-Linking-ScotXed-Data.pdf

The BigMatch parameters file was set up with the following heirarchical blocking criteria :

Block number	Block description
0	Exact matches on Postcode, Sex, DOB
1	Matches on Postcode & DOB
2	Matches on 1st 6 characters of Postcode, Sex, DOB
3	Matches on Postcode, Sex, Year & Month of Birth
4	Matches on Postcode, Sex, Year & Day of Month of Birth
5	Matches on Postcode, Sex, Month & Day of Birth
6	Matches on 1st 5 characters of Postcode, Sex, DOB
7	Matches on 1st 4 characters of Postcode, Sex, DOB
8	Matches on Postcode, Sex & Year of Birth

Number of pairs above threshold score output from all blocks per batch:

Batch Number	censusID in batch	Number of pairs	Unique censusID/spineID combinations above threshold(s)	Unique censusID above threshold(s)	Unique spineID above threshold(s)	Unique censusID/spineID combinations at best match score
1	**205,678**	755,975	405,580	69,259	287,261	189,191
TOTAL	**205,678**	**755,975**	**405,580**	**69,259**	**287,261**	**189,191**

Stage 3: deduplication

Identify where there are duplicate censusID across multiple spineID

Number of censusID/spineID combinations at best match score (per censusID)	**189,191**
Number of censusID matched to single spineID at best match score	176,311
Number of unique censusID	184,032

An automated process is carried out in order to ensure that each censusID can appear a maximum of only once in the final linked dataset.

Step 1: Where censusID spans>1 spineID in same block retain lowest ordered spineID	184,033
Step 2: Where censusID spans>1 spineID in different blocks, drop higher numbered block(s)	184,032
Pupil Census records with best matches to the spine	**184,032**
Percentage of Pupil Census records with best matches to the spine	**98.5%**
Pupil Census records with minimum threshold best matches after re-categorisation (see Step4)	**178,081**
Percentage of Pupil Census records with best matches after re-categorisation (see Step4)	**95.3%**
Unique seeded spineID numbers amongst minimum threshold best matches	**176,040**

Stage 4: match categories

Degree of specificity for EAS datasets *(ADR-S Project Request)*

For Education datasets that are ADR-S ingested, **Exact / Safe / Minimum** are categories referring to the quality of the matches:

Unique Exact is the most stringent category, but with the highest precision. It only includes unique exact matches for sex, postcode, and date of birth (DEFAULT for projects). It excludes same sex twins.

Safe includes also high confidence matches: exact match for sex, exact match for date of birth, and partial match for postcode, this slightly increases the match rate, but also increases the risk of some incorrect matches

Minimum* matches are those which meet the minimal threshold: exact match for sex, exact match for postcode, and two out of three matches from the three fields of date of birth (dd/mm/yy) - this offers the highest match rate, but introduces many false positives.

Competing matches (e.g., same sex twins) are included only as minimum. NRS strongly advises against their inclusion because it introduces many duplicate matches, involving incorrect links, that in many cases will be impossible for the researcher to resolve.

If a match is **Exact**, it'll be included as **Safe** and **Minimum** too; if a match is **Safe**, it will also be considered **Minimum**

National Records of Scotland

* *Minimum* matches have been called *Optimal* in the past. The name was changed as the word *Optimal* was misleading to represent the lowest quality of matches

Number of initial best matches - by BigMatch blocking strategy

BestBlock	Description	Frequency	Percent
0	Exact matches on Postcode, Sex, DOB	176,417	95.9%
1	Matches on Postcode & DOB (missing gender)	0	0.0%
2	Matches on 1st 6 characters of Postcode, Sex, DOB	527	0.3%
3	Matches on Postcode, Sex, Year & Month of Birth	646	0.4%
4	Matches on Postcode, Sex, Year & Day of Month of Birth	312	0.2%
5	Matches on Postcode, Sex, Month & Day of Birth	1,060	0.6%
6	Matches on 1st 5 characters of Postcode, Sex, DOB	1,936	1.1%
7	Matches on 1st 4 characters of Postcode, Sex, DOB	2,591	1.4%
8	Matches on Postcode, Sex & Year of Birth	543	0.3%
Overall		184,032	100.0%

Number of best matches by linkage criteria

	N	% of cohort	Precision* Crude Estimate
Minimum threshold links	178,081	95.3%	98.0%
Safe links	172,065	92.1%	99.8%
Unique exact links	170,982	91.5%	99.9%

*Precision estimate based on ScotXed linkages - see http://www.isdscotland.org/Products-and-Services/eDRIS/Docs/20150421-Linking-ScotXed-Data.pdf

Number of initial best matches - by broad match categories and linkage criteria

Broad Category	Description	Minimum threshold links		Safe links		Unique exact links	
1	Exact Match (including ties).	175,244	93.8%	170,982	91.5%	170,982	91.5%
2	Mis-match on last character of standardised 7-character postcode.	1,097	0.6%	1,083	0.6%	0	0.0%
3	Mis-match on one of either year, month or day of date of birth.	298	0.2%	0	0.0%	0	0.0%
4	ISD MRL linkage weight >24.0.	1,442	0.8%	0	0.0%	0	0.0%
5	Non-links	8,708	4.7%	14,724	7.9%	15,807	8.5%
Overall		186,789	100.0%	186,789	100.0%	186,789	100.0%

National Records of Scotland

NRS Project ID: 2324-A001
Pupil Census update, new pupils 2020-2022 - data ingest

Stage 5: Pupil Census 2020-22 - linkage rates by demography

Sex

			Minimum threshold links				Safe links				Unique exact links			
					Total	Row %			Total	Row %			Total	Row %
Sex		Frequency	Sex				Sex				Sex			
1	Male	95,728	1		91,310	95.4%	1		88,223	92.2%	1		87,670	91.6%
2	Female	91,060	2		86,771	95.3%	2		83,842	92.1%	2		83,312	91.5%
Total		186,788	Total	Frequency	178,081	95.3%	Total	Frequency	172,065	92.1%	Total	Frequency	170,982	91.5%

Year of birth

			Minimum threshold links				Safe links				Unique exact links			
					Total	Row %			Total	Row %			Total	Row %
YOB		Frequency	YOB				YOB				YOB			
<=2001		100	<=2001		0	0.0%	<=2001		0	0.0%	<=2001		0	0.0%
2002		37	2002		31	83.8%	2002		31	83.8%	2002		30	81.1%
2003		274	2003		192	70.1%	2003		180	65.7%	2003		177	64.6%
2004		742	2004		482	65.0%	2004		459	61.9%	2004		449	60.5%
2005		1,222	2005		940	76.9%	2005		909	74.4%	2005		896	73.3%
2006		1,617	2006		1,313	81.2%	2006		1,252	77.4%	2006		1,232	76.2%
2007		1,782	2007		1,452	81.5%	2007		1,410	79.1%	2007		1,386	77.8%
2008		1,980	2008		1,659	83.8%	2008		1,587	80.2%	2008		1,563	78.9%
2009		2,116	2009		1,829	86.4%	2009		1,744	82.4%	2009		1,719	81.2%
2010		2,186	2010		1,924	88.0%	2010		1,859	85.0%	2010		1,836	84.0%
2011		2,351	2011		2,106	89.6%	2011		2,002	85.2%	2011		1,970	83.8%
2012		2,621	2012		2,329	88.9%	2012		2,239	85.4%	2012		2,203	84.1%
2013		2,668	2013		2,418	90.6%	2013		2,316	86.8%	2013		2,282	85.5%
2014		4,338	2014		4,039	93.1%	2014		3,854	88.8%	2014		3,805	87.7%
2015		52,587	2015		51,112	97.2%	2015		49,446	94.0%	2015		49,199	93.6%
2016		55,659	2016		53,896	96.8%	2016		52,144	93.7%	2016		51,900	93.2%
2017		50,634	2017		48,593	96.0%	2017		46,993	92.8%	2017		46,722	92.3%
2018		3,872	2018		3,766	97.3%	2018		3,640	94.0%	2018		3,613	93.3%
>=2019		2	>=2019		0	0.0%	>=2019		0	0.0%	>=2019		0	0.0%
Total		186,788	Total		178,081	95.3%	Total		172,065	92.1%	Total		170,982	91.5%

SIMD 2020 decile

			Minimum threshold links				Safe links				Unique exact links			
					Total	Row %			Total	Row %			Total	Row %
SIMD 2020 decile		Frequency	SIMD 2020				SIMD 2020				SIMD 2020			
1	Most deprived	22,309	1		21,247	99.1%	1		20,439	91.6%	1		20,257	90.8%
2		20,394	2		19,529	99.1%	2		18,844	92.4%	2		18,729	91.8%
3		18,124	3		17,285	99.1%	3		16,796	92.7%	3		16,703	92.2%
4		17,872	4		17,020	99.2%	4		16,451	92.0%	4		16,269	91.0%
5		16,720	5		15,921	99.1%	5		15,366	91.9%	5		15,274	91.4%
6		17,004	6		16,227	99.1%	6		15,661	92.1%	6		15,562	91.5%
7		18,235	7		17,315	99.1%	7		16,757	91.9%	7		16,628	91.2%
8		20,586	8		19,669	99.2%	8		19,007	92.3%	8		18,928	91.9%
9		18,774	9		17,947	99.2%	9		17,337	92.3%	9		17,283	92.1%
10	Least deprived	16,620	10		15,877	99.2%	10		15,377	92.5%	10		15,325	92.2%
99	Missing	150	99		44	82.5%	99		30	20.0%	99		24	16.0%
Total		186,788	Total		178,081	95.3%	Total		172,065	92.1%	Total		170,982	91.5%

NRS Project ID: 2324-A001
Pupil Census 2007-19 and 2020-2022 - data ingest
Stage 6: Indexing Summary

Input: pupil census 2007-19 and 2020-22

Records for previously recorded pupils 2007-19	10,525,879
Records for previously recorded pupils 2020-22	1,759,417
Records for new pupils 2020-22	362,686
Total records 2007-22	**12,647,982**
Previously recorded pupils 2007-19	1,457,198
Previously recorded pupils 2020-22	644,624
New pupils 2020-22	186,788
Total pupils 2007-22	**1,643,986**
Demographic keys for previously recorded pupils 2007-19	8,907,450
Demographic keys for previously recorded pupils 2020-22	1,759,417
Demographic keys for new pupils 2020-22	362,686
Total demographic keys 2007-22	**11,029,553**
Storage keys for previously recorded pupils 2007-19	8,908,491
Storage keys for previously recorded pupils 2020-22	1,759,417
Storage keys for new pupils 2020-22	362,686
Total storage keys 2007-22	**11,030,594**

Output: pupil census 2007-19		%Match Rate
Records matched to the spine in Pupil Census dataset	12,195,733	99.3%
Individuals matched to the spine in Pupil Census dataset	1,436,256	98.6%
Distinct seeded spineID numbers	1,412,137	

Output: pupil census 2020-22

Records matched to the spine in Pupil Census dataset	348,106	96.0%
Individuals matched to the spine in Pupil Census dataset	178,081	95.3%
Distinct seeded spineID numbers	176,040	

Output: pupil census 2007-22

Records matched to the spine in Pupil Census dataset	12,543,839	99.2%
Individuals matched to the spine in Pupil Census dataset	1,614,337	98.2%
Distinct seeded spineID numbers	1,588,176	