Changes over time to Pupil Matching Reference Numbers in the National Pupil Database

Author: Dr Matthew Jay
Date: November 2024

This report summarises findings from an analysis examining the extent to which the Pupil Matching Reference numbers (PMRs) within the National Pupil Database (NPD) change over time. The project also investigated which demographic groups of individuals were more or less likely to be affected by these changes. This work was carried out as part of the ECHILD (Education and Child Health Insights from Linked Data) project.

How PMRs are used to securely link data about children

The NPD contains longitudinal information on children enrolled in state schools and nurseries in England. Data are generated by schools, exam boards and children's social care departments and sent in mandatory submissions to the UK Government's Department for Education (DfE). Across several modules, data are available on, among other things:

- pupil characteristics, including deprivation, special educational needs and social care provision
- the school each child is enrolled in
- their absences and exclusions
- exam results.

Each census (conducted once in each of the academic year's three terms) contains approximately 7 to 8 million enrolled pupils, around 93% of all school-aged children in England. NPD data have been used in a range of studies on children's education and social care. More recently, linkage to the Hospital Episode Statistics as part of the Education and Child Health Insights using Linked Data (ECHILD) project has enabled powerful, whole population studies examining the relationships between child health and education. Find out more on the ECHILD website.

Although the DfE holds identifiable data that can be used for linkage, external, accredited researchers can apply to access to de-identified extracts in trusted research environments such as the Office for National Statistics Secure Research Service. In such cases, children's longitudinal records can be linked together using the anonymised Pupil Matching Reference (PMR)—an encrypted identifier that is applicable within and across each of the NPD's modules. Accuracy of the PMR is therefore essential as it is usually the only key that researchers can use to link records together.

Why understanding changes to PMRs is important

Linkage rates between NPD and Hospital Episode Statistics in ECHILD, particularly in more recent years, are around 99%, suggesting that both sources contain high quality identifier information. This in turn implies that a high degree of confidence in the PMR is warranted. NPD data are subject to a host of cleaning and validation rules prior to submission to DfE, especially around child identifiers, whose accuracy are also implicated in funding allocation. However, errors do occur and DfE allows submitting bodies to amend previously submitted records. As a result, PMRs are liable to change over time and such changes may be applied retrospectively. This means that two separate extracts of the NPD covering the same years but created at different time points are liable to contain different PMRs. This in turn could affect comparability of results across different studies. Likewise, where researchers obtain refreshes (for example, supplementing an existing extract with more recent data), links between children's records between the pre-existing and updated data may be lost. It is therefore important to understand:

- the extent to which PMR allocation changes over time
- what child characteristics predict this.

In turn, we can understand the extent of possible biases that may result in analyses using NPD.

The ECHILD team are fortunate to be able to investigate this issue through comparisons of two separate NPD extracts:

- The first (made available in December 2020 and designated the "2020 extract") contains NPD data until the 2018/19 academic year for children born from 1 September 1995.
- The second (available in May 2023 and called the "2023 extract") contains records until 2021/22 for children born from 1 September 1984.

By restricting the 2023 extract to children born from 1 September 1995, we were able to calculate the number of PMRs appearing in one but not the other, thereby estimating the rate of change in PMRs between the two. We also calculated relative risks of a PMR not appearing in the other extract according to key demographic variables. In doing so, we aimed to quantify changes over time in PMRs within the NPD and to identify which groups of individuals are most likely to be affected by these changes.

What we did

We examined each of the spring censuses (conducted each January) from 2005/6 to 2018/2019. For each census in both the 2020 and 2023 extracts, we calculated the number of rows (= enrolments) and the number of unique PMRs (= unique pupils: pupils can be dually-enrolled). We then calculated the number and percentage of PMRs that appeared in the 2020 extract that did not appear in the 2023 extract, and vice versa.

For both extracts, for all years combined, we then calculated the unadjusted relative risk of not appearing in the other extract, according to the following factors as recorded in the same census:

- male gender (reference: female)
- major ethnic group (reference: white)
- region (reference: London)
- any special educational needs (SEN) provision (reference: none)
- composite deprivation (reference: least deprived & no free school meals). The composite deprivation metric, which we have used before, was defined as a combination of the area-based income domain affecting children index (IDACI, using earlier versions for earlier activity) and the family-based free school meals eligibility (which indicates low family income). Children were assigned IDACI quintiles (5, least deprived, to 1, most deprived) and whether or not they were eligible for free school meals.

What we found

The number of rows and PMRs in each extract, and the number and percentage of PMRs that appear in the other extract, are shown in Table 1. The number of PMRs not in the other extract was very low (≤5,100) with the percentage as low as 0.005% in 2006, rising to 0.061% in 2019. These results were almost equal whether examining PMRs in the 2020 or the 2023 extract.

Table 1. Numbers of rows and PMRs in each extract and the number and percentage of PMRs that appear in the other extract

	2020 Extract			2023 Extract		
Year*	Rows	PMRs	PMRs not in 2023 Extract	Rows	PMRs	PMRs not in 2020 Extract
2006	3680197	3679132	188 (0.0051%)	3680196	3679120	176 (0.0048%)
2007	4223017	4221936	299 (0.0071%)	4223016	4221923	286 (0.0068%)
2008	4771508	4770523	392 (0.0082%)	4771507	4770519	388 (0.0081%)
2009	5336540	5336019	498 (0.0093%)	5336539	5336012	491 (0.0092%)
2010	5925405	5924861	1005 (0.0170%)	5925404	5924855	999 (0.0169%)
2011	6528521	6527998	1576 (0.0241%)	6528520	6527979	1556 (0.0238%)
2012	7150240	7149672	2201 (0.0308%)	7150238	7149663	2192 (0.0307%)
2013	7440789	7440107	2458 (0.0330%)	7440786	7440095	2446 (0.0329%)
2014	7723793	7722609	2693 (0.0349%)	7723790	7722593	2677 (0.0347%)
2015	7913998	7851835	3082 (0.0389%)	7913993	7851824	3070 (0.0388%)
2016	8038412	7973633	3412 (0.0424%)	8038408	7973624	3403 (0.0423%)
2017	8150378	8083357	3776 (0.0463%)	8150373	8083338	3755 (0.0461%)
2018	8219215	8150895	4156 (0.0506%)	8219211	8150878	4135 (0.0503%)
2019	8305796	8236732	5098 (0.0614%)	8305790	8236732	5100 (0.0614%)

* The first few years have a far lower number of enrolments and pupils than expected because we restricted to children born from 1 September 1995.

PMRs for children with special educational needs provision were more likely to be consistent

The relative risks of a PMR not appearing in the other extract are shown in Table 2. Again, results were the same regardless of extract. PMRs assigned to children with SEN provision were *less* likely to *not* appear in the other extract compared to those without SEN; in other words, PMRs for children with SEN provision were more likely to be consistent. A similar pattern was observed for PMRs assigned to children living outside of London compared to those living in London.

Boys, pupils of ethnic minority status, and those experiencing greater levels of deprivation had a higher risk of inconsistency

By contrast, PMRs assigned to boys, pupils of ethnic minority status, and those experiencing greater levels of deprivation were *more* likely to *not* appear in the other extract. There was a clear interaction between IDACI and free school meals eligibility: within each IDACI quintile, there was

a higher relative risk of a PMR not appearing in the other extract for children with free school meal eligibility than for those without.

Table 2. Relative risk of PMRs not appearing in the other extract (spring censuses 2006 to 2019 combined)

Variable		Reference	PMR in 2020 Extract not in 2023	PMR in 2023 Extract not in 2020
			RR (95% CI)	RR (95% CI)
Gender	Male	Female	1.21 (1.18, 1.23)	1.20 (1.18, 1.23)
Ethnicity	Black	White	3.06 (2.96, 3.18)	3.07 (2.96, 3.18)
	Asian		2.52 (2.44, 2.60)	2.52 (2.44, 2.59)
	Chinese		2.72 (2.39, 3.09)	2.72 (2.39, 3.09)
	Mixed		2.10 (2.01, 2.20)	2.10 (2.01, 2.19)
	Other		2.94 (2.76, 3.13)	2.95 (2.77, 3.14)
	Unknown		2.14 (2.01, 2.28)	2.11 (1.98, 2.25)
Region	South East	London	0.57 (0.54, 0.59)	0.57 (0.55, 0.59)
	South West		0.41 (0.39, 0.43)	0.41 (0.39, 0.43)
	East of England		0.69 (0.66, 0.72)	0.69 (0.66, 0.72)
	East Midlands		0.52 (0.50, 0.55)	0.52 (0.49, 0.54)
	West Midlands		0.59 (0.57, 0.62)	0.60 (0.57, 0.62)
	Yorkshire & The Humber		0.48 (0.46, 0.50)	0.48 (0.46, 0.50)
	North East		0.33 (0.31, 0.36)	0.33 (0.31, 0.36)
	North West		0.53 (0.51, 0.55)	0.53 (0.51, 0.56)
	Unknown		0.43 (0.39, 0.48)	0.43 (0.39, 0.48)
SEN provision	Any	None	0.93 (0.91, 0.96)	0.93 (0.90, 0.96)
IDACI,FSM*	5,1	5,0	1.82 (1.61, 2.06)	1.84 (1.62, 2.08)
	4,0		0.89 (0.85, 0.93)	0.89 (0.85, 0.93)
	4,1		1.36 (1.23, 1.51)	1.33 (1.20, 1.48)
	3,0		0.97 (0.93, 1.02)	0.97 (0.93, 1.01)
	3,1		1.37 (1.26, 1.48)	1.37 (1.26, 1.48)
	2,0		1.19 (1.14, 1.24)	1.19 (1.14, 1.24)
	2,1		1.45 (1.37, 1.54)	1.45 (1.37, 1.54)
	1,0		1.54 (1.48, 1.60)	1.54 (1.48, 1.60)
	1,1		1.57 (1.50, 1.65)	1.56 (1.49, 1.63)
	Unknown		1.95 (1.65, 2.29)	1.94 (1.65, 2.28)

CI confidence interval; FSM free school meals; IDACI income deprivation affecting children index; RR relative risk; SEN special educational needs. * IDACI 5 = least deprived quintile; IDACI 1 = most deprived quintile; FSM 0 = not getting FSM; FSM 1 = getting FSM.

What these results mean

The extent to which PMRs changed between our two extracts was extremely small (≤0.0614% for any given census). While it is striking that records are still being updated at least 14 to 17 years after being submitted (for example, there were 188 PMRs in the 2006 census of the 2020 extract that were not in the same census of the 2023 extract), it was more recent years that were most affected. This recency effect may be due to schools being more likely to update more recent records.

The fact that the numbers of enrolments and PMRs were virtually identical between the two extracts indicates either that PMRs are being replaced outright, or that there is splitting and merging of PMRs in roughly equal measure. Unfortunately, as there is no third persistent identifier (such as the episode key in the Hospital Episode Statistics), and because we do not have access to natural identifiers, it was not possible for us to examine this further.
PMRs assigned to children with SEN were less likely to not appear in the other extract (i.e., these PMRs were more likely to persist). This could be due to better quality data being held for children to the whom the school and other agencies have had to make extra provision and resource available.

We observed that PMRs assigned to pupils of ethnic minority status and those experiencing greater levels of area-based and familial financial deprivation were most likely to be affected. Similarly, PMRs of children in London were more likely to be affected than PMRs of children outside of London. These findings reflect poorer linkage rates in ECHILD for children with non-White ethnicity, children living in more deprived areas, and children in London, all of which overlap to some degree. Non-English names, for example, are more likely to be mis-spelt and require correction. Children living in more deprived circumstances may be more likely to transfer schools, for example due to managed moves or off-rolling; it is therefore more likely that schools would hold poorer quality information for these children.

Nonetheless, these are relative risks calculated against a very small baseline. Results from this analysis indicate that only very few children are affected by changes in PMRs, at least over a 2 ½ year period (we are unable to conduct this analysis over longer periods or estimate cumulative effects). Whilst some groups are more at risk, which may induce some bias in results, the extent of such bias is likely no more than minimal overall.

Acknowledgement

We are grateful to the Office for National Statistics (ONS) for providing the trusted research environment for the ECHILD Database. This work was undertaken in the Office for National Statistics (ONS) Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners. This work contains data cleared by the Office for National Statistics (STATS19927 and STATS20145). This work uses data provided by children and young people collected by the NHS or the Department for Education as part of their services. We gratefully acknowledge all children and families whose de-identified data are used in this analysis. The DfE does not accept responsibility for any inferences or conclusions derived by the authors. Permissions to use linked, de-identified data from Hospital Episode Statistics and the National Public Database were granted by DfE (DR200604.02B) and NHS Digital (DARS-NIC-381972). Ethical approval for the ECHILD project was granted by the National Research Ethics Service (17/LO/1494), NHS Health Research Authority Research Ethics Committee (21/SW/0159) and UCL Great Ormond Street Institute of Child Health's Joint Research and Development Office (20PE06). This work is supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council (part of UK Research and Innovation) programme (ES/V000977/1, ES/X000427/1 and ES/X003663/1).