Data Explained

Exploring context, factors and approaches to educational exclusions and absences

Author: Dr Patricio Troncoso, Professor Morag Treanor, Dr Lee Williamson, and Cecilia Macintyre
Date: October 2023

This Data Explained summarises learning from working with education, health, and census datasets for the linkage and research project, "Exploring context, factors and approaches to educational exclusions and absences". This publication is intended to help guide future researchers using this data and to provide feedback into future dataset development and documentation.

The administrative data discussed and used in this research was made securely available through the Scottish Centre for Administrative Data Research (SCADR) Understanding Children's Lives and Outcomes project. The data comes from Educational Analytical Services (EAS) in Scottish Government, National Records Scotland, Public Health Scotland (PHS) and National Services Scotland (NHS Scotland). It was accessed through the Scottish National Safe Haven. Given that the data was not originally collected for research, gaps and inconsistencies are expected and we detail some of them in the following sections.

Introduction

The education data used for this project included person-level data for all children and young people who have been recorded in the education data demographics database from the academic year 2007/08 to 2018/19. It covers the whole of Scotland. In this project, it formed the main spine and was linked to the following datasets[1]:

- School exam qualifications
- School leaver destinations and attainment
- Exclusions
- Attendance and absence
- Pupil census
- Census 2001 and 2011
- Prescribing records (PIS – Prescription Information System)
- Hospital admissions: Scottish Morbidity Records (SMR01 – acute inpatient; SMR00 – outpatients).

National Records Scotland (NRS) conducted the linkage. To identify individuals in common and facilitate linkage, all datasets were individually matched to the NRS population spine by the NRS indexing team. Education data, Census 2001, and Census 2011 were matched to the spine via probabilistic matching using the BigMatch software. The cohort was defined by Pupil Census, including only unique exact matches to the spine for postcode, sex, and date of birth. NRS indexing used the spine to identify cohort members' Community Health Index (CHI) numbers to extract their health records.

Table 1: Overall matching rates of the datasets to the spine and cohort completion

Dataset	Matching Rate	Cohort %	Identifiers used for the spine matching
Pupil census	95%	94%	Postcode + Sex + DOB
Attendance	95%	94%	Postcode + Sex + DOB
Exclusions	95%	6%	Postcode + Sex + DOB
Leavers	96%	45%	Via pupil census
Qualifications	95%	54%	Via pupil census
Census 2001	95%	34%	Name + Surname + Postcode + Sex + DOB
Census 2011	95%	80%	Name + Surname + Postcode + Sex + DOB
Health datasets	99%	94%	CHI

All the datasets were separately indexed (pseudonymised). The master index files, i.e. files indicating how pseudonymised individuals correspond across datasets, were transferred to the Edinburgh Parallel Computing Centre (EPCC). Dataset-specific indexes were transferred to Public Health Scotland's Electronic Data Research and Innovation Service (eDRIS) and NRS Census. EPCC used Administrative Data Research Scotland (ADR Scotland) Hash rules and salts and eDRIS/NRS Census used the index files provided to extract the approved variables from the education, health, and census datasets. The EPCC linkage agent uses the master index files to link

[1] It is not possible to report a precise number of unique pupils recorded in the datasets. However, the total number of pupils in schools in Scotland from 2007 to 2019 ranged between approximately 670,000 and 700,000, according to official statistics (Scottish Government, 2020).

the cohort and payload data across the different datasets. eDRIS then transferred the linked data into the National Safe Haven's project area and made them available to researchers.

These data sources have all been linked together for the first time for this project and are only available to the research team. However, researchers interested in this linkage are welcome to apply similar linkages following the regular procedures. All data was pseudonymised (ICO, 2022) prior to being shared with researchers.

How is the data collected?

This linkage project uses administrative data from four sources. The data that the research team has access to is standalone, and covers the period 2007 to 2019. However, all these sources are linkable and can be requested in further updates of each of the below sources.

Education data: The pupil census covers all publicly-funded schools in Scotland. All local authorities, grant-aided schools and school centres in Scotland submit information relating to all pupils on roll to the Scottish Exchange of Data Unit (ScotXed). This is part of the Education Analytical Services Division within the Education and Justice Directorate of the Scottish Government. Full details of the process by which data is collected are available on the Scottish Government website.

Census data: Scotland's Census is run by the National Records of Scotland (NRS) and normally takes place once every ten years. This project uses data from the Scotland censuses that took place on 29 April 2001 and 27 March 2011. View detailed information on the 2001 Census and the 2011 Census.

Prescribing Information System (PIS): The PIS is the authoritative information hub containing comprehensive data on the prescription and cost details of all medications prescribed and dispensed in the community in Scotland. This database includes prescriptions written by GPs and prescriptions written in hospitals, but dispensed in the community – excluding prescriptions dispensed in hospitals. For this project, we requested all prescriptions that fall under the category of medicines used in mental health of the Legacy British National Formulary (BNF) classifications. Detailed information about the PIS database can be found on the NHS Scotland website.

Hospital admissions: This data is collected as part of the Scottish Morbidity Records (SMR). SMR00 records new and follow-up outpatient attendance, excluding A&E, ward attenders and bedside consultations. SMR01 records general/acute inpatient attendance and day cases. Examples include discharge from hospital, transfer, change of specialty or death. For both SMR00 and SMR01, we requested mental health-related records. Detailed information of the SMR Datasets can be found on the NHS Scotland website.

For other linkage projects, researchers should in the first instance contact Research Data Scotland, who support researchers wishing to make use of administrative datasets. They provide a single-entry point and end-to-end support to researchers throughout the process.

We advise prospective users of linked administrative datasets to consider that this process from ideation to data access can be lengthy. Thus, engaging in discussions and incorporating realistic timeframes into project proposals and planning from an early stage is highly recommended.

Key variables

The main focus of our research is on educational outcomes and trajectories, thus the main variables come from the education datasets. These are complemented with variables from the census and health datasets, which serve as contextual information.

From the education datasets, we will be focusing our analyses on the following variables:

- **Presence of exclusions:** A binary indicator of the occurrence of an exclusion within an academic year can be derived by the presence of the pupil record in the exclusions dataset. This is then linked to the pupil census. All pupils who do not appear in the exclusions dataset will have a derived variable with a value of zero, indicating no exclusions in the academic year. Those who do appear will have a value of one for that derived variable. This process can be repeated by each academic year, by year group or by educational stage (e.g. primary or secondary).
- **Length of exclusions:** Another variable of interest to our research is "length of exclusion", which appears in the exclusions dataset. As described above, pupils who appear in the exclusions dataset can have one or more exclusions and the length in half-days (called sessions) will be recorded. A variable indicating the total number of days or sessions can be derived by adding those lengths per pupil. This is again linked back to the pupil census, where those who do not appear in the exclusions dataset will have a value of zero for the derived variable. This process can be repeated by each academic year, by year group or by educational stage (e.g. primary or secondary).
- **Absences:** The attendance and absence dataset contains information for every pupil on roll. Each record (row) in this dataset represents one of 18 attendance codes indicating the number of half-days in each category of attendance or absence. The dataset contains one record per pupil per code. We will collapse reasons for absence into two main categories: authorised and unauthorised. Our analyses will focus on those two main categories and the total number of absences. This dataset also contains a code corresponding to the total number of possible sessions, which serves as the offset to account for in-year admissions, transfers, and other reasons that might prevent a pupil to be on roll for the full academic year.
- **Educational achievement:** A breakdown of Scottish Qualifications Authority (SQA) national qualifications achieved by each pupil, including grades, are recorded in the Qualifications and Leaver Destinations and Attainment datasets. The qualifications dataset allows the researchers to track down progress across the senior phase of secondary school. It provides information for qualifications achieved at any level up to the academic year of the dataset. In contrast, the Leaver Destinations and Attainment dataset allows the researchers to determine the highest level of qualifications achieved at the

point the pupil leaves school. Qualification grades will be converted to numerical values using tariff points as per the Universities and Colleges Admissions Service (UCAS) methodology.

- **Identifiers:** All pupil records in the education datasets contain an anonymised pupil identifier, which allows the tracking of pupils over time. They also contain an anonymised school identifier and a local authority code for the location of the school and the residence of the pupil. These variables allow us to estimate the amount of variability in exclusions (in terms of whether a pupil has been excluded or not and for how long) that is exclusively attributable to each of these sources (e.g. when fitting a multilevel model).
- **Household characteristics:** Information from the census will be used to derive contextual variables – potentially an index of socioeconomic disadvantage. A selection of variables we will use are:
    - Economic activity
    - Ever worked indicator
    - General health
    - Highest level of qualification
    - Hours worked
    - Long-term health condition: nature of condition
    - Long-term health condition: number of conditions
    - Long-term health problem or disability
    - Marital and civil partnership status
    - National Statistics Socio-economic classification (NS-SeC)
    - Occupation
    - Number of People in the household with limited by a long-term health problem or disability
    - Unemployment history.
- **Health contextual information:** Health-related circumstances are another important source of contextual information that will shed light on the reasons behind different patterns of absence and exclusions. Mental health indicators can be derived from the Prescribing Information System (PIS) by comparing with the pupil census in the same way as described in the case for exclusions. That is, we will derive mental health indicators from the PIS by assigning a pupil a value of one if they have a record in the PIS, comparing the data with the pupil census, and assigning them a zero for mental health if they are present in the pupil census but not the PIS. It is important to note that a value of zero in this derived mental health variable does not necessarily indicate absence of mental health issues, as it can also indicate that no issues have been recorded or that no healthcare has been sought.
- **Additional contextual information:** The pupil census contains demographic information that can be used as another source of information to contextualise the patterns of absence and attendance. Some of the variables to be used are: free school meal registration, sex, special education needs, and disabilities.

What can the data be used for?

This project examines the variation between schools and local authorities in relation to temporary exclusions and authorised/unauthorised absences, providing a natural experiment to be carried out between local authorities. The research will also explore the role of family characteristics and circumstances, including both material poverty and associated factors (e.g. unemployment, disability) and household living arrangements. This will allow us to explore the extent to which it is variability at the level of the local authority, and not the demographics of the catchment or the characteristics of the family, that influence absences, exclusions, and outcomes.

More specifically, we aim to address the following research questions:

1. What impacts do absences and exclusions have on children and young people's educational outcomes and post-school destinations?
2. How do exclusions and absences vary by schools and local authorities?
3. What impacts do these school and local authority-level variations have on children and young people's educational outcomes and post-school destinations?
4. To what extent may absences be masking unofficial exclusions, and what effects do these have?
5. How do these impacts and effects vary for particular groups of children, e.g. those living in poverty and those with additional support needs/special educational needs and disabilities?
6. How do family characteristics and circumstances interact, mediate, or moderate the prevalence and the impacts of young people's exclusions and absences?

Existing research or examples of previous research

This large-scale linkage project is the first of its kind in Scotland. Therefore, we will be producing the first ever research outputs.

Data limitations encountered

Administrative data is not collected for research, hence there are numerous challenges to overcome when working with this data. First and foremost, this type of data normally requires extensive manipulation to make it suitable for research, and careful planning and attention must be given to how data is restructured, selected, merged, and analysed.

Despite it being a record linkage of administrative data and not suffering from attrition in the same way as for sample surveys, missing data can and does occur in the linked datasets. It is not always possible to determine a reason for this. Another non-negligible challenge of this high volume of administrative data is the processing power required for fairly simple tasks. Complex statistical modelling in large datasets such as these is time-consuming. , which is why sampling should be considered, at least for the model exploration phase of the research. Inconsistencies

across datasets is also possible and researchers need to make decisions around them, such as what sources should supersede contradictory records in other sources.

In terms of the health datasets, while it is a reasonable assumption that children living in Scotland will have access to NHS healthcare and will be engaging with healthcare during childhood, there may still be self-selection processes at play. This is especially the case given our focus on mental health. Young people have records on the health datasets that have been linked to the education datasets, when they themselves or their parents have sought medical help and have been referred, assessed, and treated. This means that pupils whose needs do not reach set thresholds might be missing completely from the health datasets. Furthermore, those who do not seek or who are unable to access the requisite healthcare will not be present. Only those who successfully received healthcare in the period under study will be present.

As previously noted, the education datasets encompass all publicly-funded schools in Scotland. Hence, it is important to acknowledge that this scope excludes independent schools. Consequently, all conclusions derived from the analyses of these datasets refer specifically to the population of publicly-funded schools in Scotland.

This is not an exhaustive list of challenges, but a set of examples of relevant issues that are common to find in linked datasets.

Suggested improvements & recommendations to data owners

We have mentioned before that administrative data is not collected for research purposes and extensive data management and manipulation is expected. Nevertheless, data owners can still improve the way in which the data is delivered to researchers. One salient example from our data linkage project is the attendance dataset, which is made available to researchers in the same structure as specified in the ScotXed data collection. This means that multiple attendance codes are stored as rows which are nested within unique pupils, resulting in a complex data structure that requires careful restructuring by the researchers.

This could be solved if data were restructured prior to sharing. Attendance codes could be reshaped into variables (columns) and keep one record (row) per pupil. This could significantly reduce research processing times.

Suggested future data linkages

In the future, a further linkage to the Achievement of Curriculum for Excellence Levels dataset would be extremely beneficial. This would allow tracking school achievement in core subjects from primary school to secondary school, which would crucially help to inform the achievement of qualifications at secondary school.

Another linkage that would enhance research to understand the factors associated with absences and exclusions from school would be to the Health and Wellbeing Census Scotland. This is because the health data we have access to provides information on those cases who have sought healthcare, hence it suffers from selection bias. Having access to health and wellbeing data for the entire school population would enable multiple comparisons at different levels of wellbeing.

Finally, further linkage to data sources about post-compulsory education would be beneficial to understand the potential (if any) long-term impacts of absences and exclusions from school. Potential sources would be the Higher Education Statistics Agency (HESA), the Scottish Funding Council and the University and Colleges Admission Service (UCAS).

Conclusion

In this Data Explained, we have described how we created the linked dataset that will be used for the project "Exploring context, factors and approaches to educational exclusions and absences". Even though the resulting dataset from this complex linkage is only available for this particular research, this can inform and inspire future research in similar topics. Working towards the principles of FAIR (findability, accessibility, interoperability, and reusability) at the end of the project, key parts of data management and analysis code will be made available within an online code repository in GitHub. This document will be updated with further information as the research progresses.

Version 1 – November 2023

References

Information Commissioner's Office. (2022). Chapter 3: pseudonymisation. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Available at: https://ico.org.uk/media/about-the-ico/consultations/4019579/chapter-3-anonymisation-guidance.pdf

National Records Scotland (2001). Scotland's 2001 Census. https://www.scotlandscensus.gov.uk/about/2001-census/

National Records Scotland (2011). Scotland's 2011 Census. https://www.scotlandscensus.gov.uk/about/2011-census/

NHS Scotland. (2019). Medicines used in Mental Health: Years 2009/10-2018-19. Available at: https://www.isdscotland.org/Health-Topics/Prescribing-and-Medicines/Publications/2019-10-22/2019-10-22-PrescribingMentalHealth-Report.pdf

NHS Scotland. (2023). SMR Datasets: ISD Scotland Data Dictionary. Available at: https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/

NHS Scotland. (2023). National Data Catalogue: Prescribing Information System (PIS). Available at: https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=9

Public Health Scotland. (2023). eDRIS: Products and Services. Available at: https://www.isdscotland.org/products-and-services/edris/

Scottish Government. (2020). Pupils in Scotland 2020. Official Statistics available at: https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2019/07/pupil-census-supplementary-tables/documents/pupil-census-2020-supplementary-statistics/pupil-census-2020-supplementary-statistics/govscot%3Adocument/Pupils%2Bin%2BScotland%2B2020%2BV2.xlsx

Scottish Government. (2022). Achievement of Curriculum for Excellence levels: 2021/22. Available at: https://www.gov.scot/publications/achievement-curriculum-excellence-cfe-levels-2021-22/

Scottish Government. (2023). Health and Wellbeing Census Scotland 2021- 2022. Available at: https://www.gov.scot/publications/health-and-wellbeing-census-scotland-2021-22/

Scottish Government. (2023). Scottish Exchange of Data: school-pupil census. Available at: https://www.gov.scot/publications/scottish-exchange-of-data-school-pupil-census/

Universities and Colleges Admissions Service. (2022). UCAS Tariff Tables. Available at: https://www.ucas.com/file/167761/download?token=jHp8Krb1

Glossary

ACEL: Achievement of Curriculum for Excellence Levels
ADR-S: Administrative Data Research Scotland
EAS: Educational Analytical Services
CHI: Community Health Index
eDRIS: Electronic Data Research and Innovation Service
EPCC: Edinburgh Parallel Computing Centre
HESA: Higher Education Statistics Agency
NHS Scotland: National Health Services Scotland
NRS: National Records Scotland
PIS: Prescription Information System
SCADR: Scottish Centre for Administrative Data Research
ScotXed: Scottish Exchange of Data Unit
SMR: Scottish Morbidity Records
SQA: Scottish Qualifications Authority
UCAS: University and Colleges Admission Service

Disclaimer

This work was produced using administrative data accessed through the Scottish National Safe Haven. The use of the data in this work does not imply the endorsement of the Trusted Research Environment or data owners in relation to the interpretation or analysis. This work uses research datasets which may not exactly reproduce National Statistics aggregates. National Statistics follow consistent statistical conventions over time and cannot be compared to linked datasets.

Acknowledgements

This work is supported by ADR Scotland which is part of Administrative Data Research UK. ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. It is funded by the Economic and Social Research Council [Grant number: ESRCES/W010321/1].

Contact

Name: Patricio Troncoso, Research Fellow, SCADR, University of Glasgow
Email: Patricio.Troncoso@glasgow.ac.uk