Data Explained

Data First: Ministry of Justice & Department for Education linked dataset - England

Education and social care predictors of offending trajectories: A UK administrative data linkage study

Author: Dr Hannah Dickson

Date: April 2023

This Data Explained output summarises experiences and learning from working with the Data First: Ministry of Justice & Department for Education linked dataset in the course of producing research into education and social care predictors of offending trajectories. 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 in this Data Explained was made securely available through Data First programme: a ground-breaking data linkage initiative, led by the Ministry of Justice (MoJ) and funded by ADR UK. The data used in this research project comes from the MoJ and Department for Education (DfE) and was accessed through the Secure Research Service (SRS) hosted by the Office for National Statistics (ONS). The data was not originally collected for research and it is expected that there are gaps and inconsistencies in its recording, some of which are detailed in the following.

Project details

The long-established age-crime curve shows that criminal offending peaks in adolescence and decreases in adulthood. However, longitudinal studies following individuals over the lifespan suggest that this curve conceals a number of distinct patterns of (re)-offending or trajectories. The first well-established trajectory is called ***'life-course persistent'*** offending where individuals begin to behave antisocially in childhood and continue into adulthood (Moffit, 2018). The majority of criminal offences are conducted by this group of offenders with an associated individual lifetime cost of £1.1- £1.9 million in the UK (MoJ, 2019). The second pattern is known as '***Adolescent-limited'*** offending (Moffit, 2018). Here, individuals commit criminal offences mostly during adolescence, with a minority continuing to offend into adulthood (Piquero et al., 2013). Another trajectory is that of ***no or low densities*** of offending behaviours.

However, longitudinal studies are often limited by small sample sizes, selection bias and infrequency of data collection which may limit the identification of lesser-known offending trajectories (Bosick et al., 2015; Jolliffe et al., 2017). It is important to examine offending trajectories because some (re)-offending patterns are associated with more negative outcomes, like poor physical and mental health, lower educational attainment and unemployment than others (Piquero et al., 2011; Reising et al., 2019; Van der Geest et al., 2014).

In the first part of my ADR UK Research Fellowship, I have been using information on offending from the Ministry of Justice & Department for Education linked dataset to develop trajectories of known offending (i.e., offences recorded by the criminal justice system). This work is the focus of this Data Explained publication. In the next stages of my fellowship, I will be using education and social care information contained within the linked dataset to see if we can discriminate between the different offending trajectories identified. This next stage has the potential to inform early targeted interventions to reduce criminal justice system involvement, thereby reducing criminal offending and its associated social and economic costs (Moffit, 2018).

Initial research questions

My research fellowship set out to examine the following two research questions. As this work is ongoing, this Data Explained publication will focus primarily on the first research question.

1. What are the offending trajectories of individuals born on or after 31 August 1985 up to 31 August 1999?
2. Which administrative education and social care record data is most helpful in predicting these specific offending trajectories?

Research methodology

I am using offending information up to the end of December 2017 taken from the Police National Computer (PNC) for individuals born on or after 31 August 1985 up to 31 August 1999. The PNC contains information on criminal offences recorded in the criminal justice system. To develop trajectories of recorded offences, I have used a statistical analytical technique called 'Latent Class analyses' (LCA), which is where you try and identify *'subgroups'* of individuals based on patterns of observed information (Masyn, 2013). Using existing research and previous work by the Ministry of Justice (MoJ) on prolific offending (MoJ, 2019), I developed a series of variables to use in LCA:

- *Offence type*: Ever committed a violent offence; or non-violent offences only
- *Age of first conviction or caution*: 10-13 years; 14-17 years or 18 years and over
- *Age of last offence in PNC*: 10-17 years or 18 years and over
- *Offending History* : Number of offences committed as a juvenile (10-17 years); as a young adult (18-20 years); and as an adult (21 years and over). At each of these stages, an individual was categorised as being '***prolific***' if they had committed more offences than the median, or ***'low-density'***, if they had not[1].

Once the trajectories have been identified, it is then possible to assign each individual to a 'class' (trajectory,) based on their probability of being in that particular 'class' given the pattern of scores they have on the indicator variables described above.

Key variables

To develop the variables listed above, I used the variables in the table below to address my first research question. Note that variables with an asterisk (*) were not used for initial model development but were used as covariates once the best class solution (i.e., optimal number of trajectories to fit the data) had been identified. These covariates were used because there is evidence that the trajectories of female offenders will differ from males, and because if someone has served a custodial sentence, they will have less opportunity to commit offences. Information on Sex was primarily taken from the spring census of the National Pupil Database (NPD), but I used the PNC to fill in missing information, if appropriate.

Police National Computer (PNC)			
MoJUID	DisposalDays*	OffenceStartDate	AdjudicationCode
DisposalDuration*	HOOffencecode	OffenceStartage	Sex*

Note: *variables used to develop covariates in LCA.

[1] Note that this definition is different to the MoJ's. See below for more details.

How you dealt with data limitations

Any data not collected for research purposes will always have limitations. The main limitations observed whilst using the PNC are reported below:

The lack of detail for some variables in the metadata
Some offence codes had a letter 'prefix'. Understanding what these letters meant was important because it impacted upon which offences were included in an individual's offending history and which were not. It would be helpful to have information on what the letters mean in the metadata for future users.

Recording errors/missing data and implications for checking inconsistencies
The PNC contains all the information required to establish individuals' criminal histories. The PNC does not contain information on month or year of birth but does have a variable indicating the individual's age in years when the offence took place. Age at the time of offence was key to developing trajectories because it meant I could develop an *'Age at First Offence'* variable. However, some individuals were given zero or negative numbers for offence start age.

I also found that some offence codes were missing. Offence codes were important because they provided information on offence type which I used in LCA. In cases where offence ages were not plausible and/or offence codes were missing, I dropped those offenders, as opposed to individual offences, from the dataset to reduce the potential for bias. As I was using the linked PNC-NPD data, I was able to double-check '*Age at First Offence'* by using offence start date from the PNC and month and year of birth taken from the census information in NPD. A very small proportion appeared to have an incompatible *'Age at First Offence'* according to offence start date and month/year of birth. A slightly higher proportion of individuals in the PNC did not have information on month/year of birth in the NPD, which meant I was unable to check if their *'Age at First Offence'* was correct.

All analyses will be conducted across the whole sample and then run again removing the small number of individuals with apparently incorrect information on age at first offence and those missing month/year of birth information in the NPD. Although recording errors are a limitation, the small number observed point to good data quality for some important PNC variables.

Defining prolific offending
The MoJ (2019) re-defined prolific offending as the previous definitions did not take into account criminal career duration (i.e., a longer criminal career provides more opportunity for offending and a higher total number of offences). The new definition of a prolific offender considered three age groups of offenders (e.g., juvenile, young adult and adult) by using information on age of first offence, age last seen in the PNC and offending history. Offending history here can be thought of as the total number of convictions or cautions with the amount of time someone has had to offend is taken into consideration. How offending history for each of the three age groups is defined by the MoJ is briefly described below:

A **juvenile prolific offender** meets one of the following criteria:

- Aged 21 years or older at last appearance in the PNC and has fewer than 16 convictions or cautions in total: fewer than 4 when aged 18-20 years, but more than 4 when aged 10-17 years;
- Aged 18-20 years on last appearance in the PNC and has 8 or more convictions, and less than 4 of which occurred when aged 18-20 years;
- Last seen in the PNC when aged 10-17 years with 4 or more convictions or cautions.

A **young adult prolific offender** meets one of the following criteria:

- Someone who on last appearance was 21 years or over with 16 or more convictions, but fewer than 8 occurred when over 21 years and 4 or more occurred when aged between 18-20 years;
- Someone who on last appearance was between 18-20 years with 8 or more convictions/cautions, with at least 4 occurring between ages 18-20 years.

An **adult prolific offender** is defined as someone who on last appearance was 21 years or over with 16 or more convictions, 8 of which occurred when aged over 21 years.

Initially, I had planned to use the MoJ's definitions of prolific offending in order to increase the relevance of my project findings to the UK criminal justice system. However, I had wanted to be able to ascertain whether an individual had been prolific or not as a juvenile, young adult and an adult. Using this approach meant I would potentially be able to identify *'life-course persistent'* and *'adolescent-limited'* trajectories that have been widely reported in existing research (Moffit, 2018).

Using the MoJ definition of prolific offending described above, an individual was either a juvenile prolific offender, young adult prolific offender or an adult prolific offender. Using the MoJ definition it was also unclear how to define a non-prolific offender across the three age categories. A further issue for my project was that because the MoJ definitions were 'trajectories' in themselves, LCA model non-convergence was problematic. Therefore, I opted to develop my own definitions of criminal history (see '**Research Methodology'**) but using the same information from the PNC used in the MoJ's definition of prolific offending.

Data coverage

In order to address research question 1, I selected the oldest birth years available for my project. I did this to maximise the length of time someone could possibly appear in the PNC. For example, for someone born in academic year 1985/1986, I had offending information from 10-32 years. For someone, born in 1986/1987, I had offending information from 10-31 years. However, selecting this older cohort has meant that I have faced two key issues in my research.

Firstly, depending on academic year of birth everyone had slightly different lengths of possible offending follow-up information, although everyone except those born in academic year 1997/1998 and 1998/1999 could have potentially offended at each of the three age time points (i.e., as a juvenile, young adult and adult). It was difficult to account for differences in length of follow-up according to academic year of birth due to model non-convergence, issues with interpretability of results, limitations of statistical software used and lack of computational power.

Instead, I compared (re)-offending trajectories between individuals in the oldest five years relative to the rest of the cohort to see whether length of follow-up impacted upon the types of trajectories I found[2].

Secondly, selecting an older cohort means that I will have less data coverage or available NPD data for the next stage of the analyses (research question 2), which is looking at the educational and social care predictors of the observed offending trajectories. To address this limitation, I will undertake a sensitivity analysis by re-running models removing those offenders born in the earlier academic years. Similarly, it should also be noted that the PNC contains information on crimes committed in England and Wales. The NPD only contains information for school children in England. Therefore, in the next of stage of my analyses I will need to undertake further sensitivity analyses to address this limitation.

Overall, I encountered very few limitations with the PNC, the recording errors and missing data I observed are similar to problems encountered in data collected for research purposes. All future users are urged to undertake the necessary data accuracy checks on random smaller samples to support double-checking of code. Users are also encouraged to document all decision making in 'read me' documents and in heavily annotated coding files when cleaning variables for analyses for transparency and consistency. Finally, linked data provides opportunities to examine data quality and accuracy across individual datasets and should be utilised as early as possible in the analyses pipeline.

[2] In my original project proposal, I had requested data for individuals born up from 1985-2007 but I was advised that this was an unnecessarily large cohort. Reducing the years of birth requested helped minimise potentially bigger issues arising from differences in length of follow-up.

Suggested improvements recommended to data owners

The definitions of prolific offenders set out by the MoJ accurately identify individuals who have committed a disproportionately large number of offences. However, these definitions have limited applicability outside of policy-related reports, as comparison to the wider evidence base is problematic. One recommendation is that the data owners, in consultation with other relevant stakeholders, consider whether the current definitions of prolific offending are overly complex and could be simplified by looking at prolific offending over the life course, or at multiple age points. Clearly defining non-prolific offending would also be of benefit to future users examining patterns of (re)-offending.

Working with administrative data involves making multiple decisions about variables, which can have implications further down the analyses pipeline. Sometimes guidance on making these decisions is not immediately obvious or available. The DfE have a number of methodology reports related to key NPD variables (e.g., student absence and exclusions), which can be really helpful when making decisions on how best to develop variables for analyses. Also of use would be a PNC (or PNC-NPD) user group that is hosted within the ONS SRS, where users can share code and advice on variable quality and development. This could potentially encourage consistency across policy and research reports, where appropriate. Both these resources could be used alongside the metadata and would be invaluable to new users.

Additional data which would help to further develop the research

The PNC-NPD linked dataset, which this project utilises, is an important research resource with huge potential to improve outcomes for young people in the criminal justice system. In the longer term, new linkages between the PNC and administrative health data, or even a linkage between the PNC-NPD and health data would be a welcome addition to the Data First programme and ADR UK's data catalogue. This would enable researchers external to government to examine the intersectionality between education, health and crime among hard-to-reach and disadvantaged populations, to produce research with a high level of external validity and applicability to UK policy making.

References

Bosick, S.J., Bersani, B.E. & Farrington, D.P. (2015). Relating Clusters of Adolescent Problems to Adult Criminal Trajectories: A Person-Centered, Prospective Approach. *Journal of Developmental Life Course Criminology*, *1***:** 169–188.

Jolliffe, D., Farrington, D. P., Piquero, A. R, (2017). Prevalence of life-course persistent, adolescence-limited, and late-onset offenders in prospective longitudinal studies. *Aggression and Violent Behavior*, *33*: 4–14.

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Statistical analysis* (pp. 551–611). Oxford University Press

Ministry of Justice (2019). Prolific Offenders Criminal Pathway: Prison Events & Offender Needs. London.

Moffitt, T. E. (2018). Male antisocial behaviour in adolescence and beyond. *Nature Human Behavior*, *2*: 177–186.

Piquero, A. R., Diamond, B., Jennings, W. G., & Reingle, J. M. (2013). Adolescence-limited offending. In C. L. Gibson & M. D. Krohn (Eds.), Handbook of life-course criminology: Emerging trends and directions for future research (p. 129–142). Springer Science + Business Media. New York.

Piquero, A. R., Shepherd, I., Shepherd, J. P. & Farrington, D. P. (2011). Impact of offending trajectories on health: disability, hospitalisation and death in middle-aged men in the Cambridge Study in Delinquent Development. Criminal Behaviour and Mental Health, 21: 189-201

Reising, K., Ttofi, M. M., Farrington, D. P. & Piquero, A. R. (2019). Depression and anxiety outcomes of offending trajectories: A systematic review of prospective longitudinal studies. Journal of Criminal Justice, 62: 3-15.

Van der Geest, V. R., Bijleveld, C. J. H, Bokland, A. J. & Nagin, D. S. (2014). The Effects of Incarceration on Longitudinal Trajectories of Employment. Crime & Delinquency, 62 (1): 107-140

Disclaimer

This work was produced using administrative data accessed through the ONS SRS. The use of the data in this work does not imply the endorsement of the ONS SRS 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 Data First linked datasets.

Acknowledgements

This work is supported by ADR UK (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. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). [Grant number: ES/W002647/1]

Contact

Dr Hannah Dickson
Email: hannah.dickson@kcl.ac.uk