Data Explained

Linking probation and criminal courts datasets

Enforced alcohol abstinence: does it reduce reoffending?
Author: Dr Carly Lightowlers
Date: September 2024

This Data Explained summarises experiences and learning from work linking the Data First probation (Jan 2014 - Dec 2020) and magistrates' courts datasets (Jan 2011 – Dec 2020). This was done while producing research into how alcohol treatment and monitoring requirements are being used and whether they are reducing reoffending. 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 the 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 Ministry of Justice (MoJ) and was accessed through the Office for National Statistics (ONS) Secure Research Service. The data was not originally collected for research, and it is expected that there are gaps and inconsistencies in its recording, a number of which are detailed in the following.

Project details

In cases where alcohol consumption has played a role in a defendant's offending, courts in England and Wales may make use of alcohol treatment or abstinence requirements as part of a community sentence.

Alcohol treatment requirements are aimed at offenders who are dependent on alcohol, whereas alcohol abstinence monitoring requirements are considered suitable for non-dependent drinkers whose drinking contributed to their offending. Compliance with the latter can be backed up with electronic monitoring devices (and thus the threat of further court hearings, fines or imprisonment). Given the considerable social and economic costs associated with alcohol-related crime, the project's insights will contribute to our understanding of how to effectively respond to this problem.

This Data Explained will focus on lessons learned in linking probation (Jan 2014 - Dec 2020) and magistrates' courts datasets (Jan 2011 – Dec 2020) to explore whether those receiving the above outlined requirements reappear before the magistrates' courts again (as a proxy for reoffending). It will reflect on linking these datasets and generating proxy measures of reoffending prevalence, incidence, and the duration (in days) elapsed between receiving an alcohol requirement and any further offence.

This is the second of two Data Explained reports associated with this project. The first focused solely on the probation dataset and its utility for identifying alcohol-related offences and community sentence requirements therein. This second report expands on the first to consider further insights made available upon linking the probation data with the magistrates' courts data.

Initial research questions

The study is guided by two overarching research questions as follows:

- What is the profile of offenders for whom alcohol related treatment and abstinence requirements are used?
- Are alcohol related treatment and abstinence requirements effective?

Research methodology

The study adopts both cross-sectional and longitudinal analysis to:

- Assess the profile of offenders issued with alcohol treatment or abstinence orders and for whom these are effective.
- Compare whether those given alcohol treatment or abstinence orders fare better than those who do not receive such (e.g., for similar (alcohol-defined) offence and demographic profiles[1]).

As well as examining successful completion of alcohol treatment or abstinence orders, success will be evaluated based on three related proxy reoffending outcome measures (because there is no information on crimes that were not detected and proceeded against). Namely, a binary indicator and count of reappearances before court as well as a duration variable capturing the time elapsed to any such reappearance.[2]

Analyses will comprise descriptive interrogation of the data as well as bivariate tests and regression modelling (to include logistic and negative binomial/Poisson regression as well as event history analysis). Further detail of the analytical strategy and detail of the statistical methods to be deployed can be accessed on OSF Registries.

This Data Explained report focuses on the how the linked (longitudinal) analysis of the probation and magistrates' courts data can be used to identify reoffending. Specifically, how the data was linked, and how the proxy outcome measures of reoffending, to be deployed in the analyses, were derived.

[1] See my Data Explained for further detail on criteria being used to identify comparisons in offence groups, including the identification of alcohol-*defined* offences (owing to limitations in not being able to specify offences as alcohol-*related*).
[2] A common metric of intervention effectiveness is reoffending and one way of measuring this is by reconviction (Merrington and Stanley, 2007). Here reconviction is operationalised by reappearance before the court.

Datasets and key linking variables

This study deployed three datasets:

1. **Magistrates' courts defendant data**
   This data comprises an extract from the magistrates' courts management information system (Libra) for January 2011 to December 2020 (based on appearance date). It captures adults and young people appearing as defendants before the magistrates' courts in England and Wales. The data is structured with one record per defendant per case: that is, defendants who have appeared twice in different cases before the magistrates' courts will feature twice in the data.

2. **Probation data**
   The probation dataset comprises an extract from a system used for the management of offenders subject to probation supervision (National Delius (nDelius)) for 1 January 2014 to 31 December 2020 (based on the referral date). It captures adults under supervision of the probation service in England and Wales while serving a community sentence, or released from prison on licence or on parole. The main data file is structured with one record per offender-event with information about offender characteristics, offence and disposal types, to which additional data about any community requirements[3] can be appended.

3. **Cross-justice system linking dataset**
   This linking dataset can be used in conjunction with other datasets shared as part of the Data First programme to join up information from across the justice system (criminal, family and civil courts, prisons and probation). Records relating to individual justice system users can be linked using unique identifiers provided for people involved. It acts as a lookup to identify where records in other datasets are believed to relate to the same person, using the probabilistic record linkage package, Splink (Linacre et al., 2022).

Both the magistrates' courts and probation datasets have been deduplicated using the open-source statistical Splink software. The deduplication process allows data users to identify which defendants have re-entered the criminal courts or probation caseload more than once by establishing defendant records that are believed to belong to the same person, using a probabilistic linkage method.

Records in the two datasets can be linked (at the individual level) by the cross-justice system linking dataset, which acts as a lookup table to identify where records in the criminal courts datasets (probabilistically) refer to the same people in the probation dataset. This allows identification of records in the two datasets that belong to the same defendant/offender, and so repeat appearances can be investigated.

Having previously prepared the probation data by linking the probation flatfile and requirements datasets (see my first Data Explained), the ambition was to link these person-events of probation

[3] Information on license and post-sentence supervision as well as pre-sentence report details are also available, but not deployed in this study.

supervision back to cases in the magistrates' courts data. This means matching one court case to one associated period of probation supervision where this was part of any resulting sentence.

However, this has (to date) not been made possible, as the existing linking dataset does not facilitate the linking of cases within people across the probation and criminal courts datasets. Rather, linking information is only available at the person level – that is, the linking file only provides a unique *person* ID (as opposed to case ID) which serves as a statistical estimate of which defendants are the same person in both datasets.[4] This can be used to join records based on the unique *person* ID and will result in joining (and duplicating) ***all*** probation records relating to that person (as opposed to all records relating to a case).

As such, several alternative approaches were considered, including:

1. Filtering cases (such as only those in receipt of treatment or abstinence requirements to keep fewer in scope for matching). However, this would compromise the ability in further analyses to compare these subsamples against similar cases not in receipt of treatment or abstinence requirements .
2. Matching person-events based on the unique person identifier as well as key information such as the offence dates in each dataset. However, this additional deterministic linking approach risks losing genuine matches (owing to errors in data entry or missing entries).
3. Developing a (probabilistic) 'best match solution' based on information in first table and a set of rules to select one magistrates' courts record for probation row_id_hash and other duplicates (based on how expected information like the magistrates is). However, developing this method was beyond the scope of the project.

Ultimately, an approach akin to #2 was adopted, whereby the magistrates' courts and probation data was matched based on the unique person ID as well as the offence date (main_offence_date in the probation data and offence_date in the magistrates' courts data).[5] This yielded a linked person-events[6] datafile and resulted in a cohort of probation service users with matched records in the magistrates' courts data for further analysis.[7]

[4] Although linking across cases within individuals (identification of records related to the same case and individual) has been made possible using the existing Splink methodology across the magistrates' and Crown Court datasets, a similar link has not been generated for linking cases across probation and criminal court data (see https://www.gov.uk/guidance/ministry-of-justice-data-first).
[5] This could deploy a fuzzy matching process whereby the absolute difference between these dates was no more than a given number of days. However, this would reduce confidence in the quality of the matches yielded.
[6] Events in each dataset which were thought to be associated with the same person.
[7] Matching on offence types was not attempted as it was thought to prove unreliable and result in further loss of genuine matches, noting the earlier outlined lack of correspondence in offence classifications/categories across the court and probation datasets (Lightowlers 2024).

The key variables used from each dataset were as follows:

Table 1: Variables used in the probation dataset

Variable name(s)	Comments
row_id_hash	A unique row identifier (containing information about a particular offender).
main_offence_date	The date of the main offence.
referral_date	A referral date for every unique event.

Table 2: Variables used in the magistrates' courts dataset

Variable name(s)	Comments
row_id_hash	A unique row identifier (containing information about a particular defendant, case, and outcome).
offence_date	Date on which offence committed or earliest date if a range specified.
case_id_hash	A unique case identifier.

Table 3: Variables used in the cross-justice system linking dataset

Variable name(s)	Comments
source_dataset	Whether the defendant_in_case_id_hash variable is from the magistrates' courts (mags_hocas), Crown Court (crown_xhibit), probation (probation_delius) prison (prison_nomis), family court (family_dedupe) or civil court (caseman_pcol) datasets.
estimated_xjs_id	A unique identifier for the defendant or offender. This is a derived field resulting from a process of probabilistic record matching using personal information not shared in this dataset. It therefore represents a statistical estimate of which defendants and offenders are the same person.

The steps taken to achieve the above outlined matching were as follows:[8]

Steps 1-3: Load in magistrates' courts data, probation data and cross-justice system linking dataset

The magistrates' courts data comprised 13,433,308 cases relating to 6,341,591 distinct defendants. These were sub-setted to retain only those entries relating to the most serious offence in any given case in the analyses that follow.

Probation data comprised 1,873,228 cases relating to 880,193 unique offenders. These observations were linked to probation requirements data to generate a wide format dataset used in the analyses that follow (see my first Data Explained for the approach on linking probation requirements and flatfile).

Step 4: Append linking information to magistrates' courts data

Step 5: Append probation linking information to magistrates' courts data

Upon linking this information only unique appearances in the magistrates' courts data were retained to match with individuals.

Step 6: Link probation data to magistrates' courts data: Event matching on date and person identifier

To offer more confidence in matching, cases were matched on the estimated person identifier variable (estimated_xjs_id) as well as the offence dates.[9]

This produces a dataset (n=9,657,787) with all magistrates' courts data entries and corresponding probation details where the person is the same (comprising records for 6,030,725 unique defendants).[10]

Step 7: Subset matched data for only those with probation records results for further analysis

Sub-setting this data for only those with probation records results in n=1,198,780 (unique defendants, n=656,729).

---

[8] Accompanying R code can be found here: https://github.com/CarlyLL/Enforced-alcohol-abstinence.
[9] As part of this process, some matches were also omitted owing to date formatting errors (e.g. in the magistrates' courts data, some dates were not in the standard (ymd) date format).
[10] NB: as the linkage is by people (as opposed to case), probation records are appended each time the person appears in the magistrates' courts data. As such, information about a person's probation supervision is appended to a person before the magistrates' courts, even if that case did not result in the period of supervision. To mitigate against this problem, matching is performed using offence date as well as the person identifier and step 7 is subsequently performed.

Step 8: Subset matched data to remove invalid or historical offence dates[11]

This left n=1,188,349 records, with n=647,559 unique defendant events with which to pursue further analysis.

Calculating court reappearances

Having linked the data, it was then possible to examine whether and how long after a referral to probation before an individual comes back before the courts again – reappearances, as a proxy for reoffending.

Having performed the previous linkage, it was possible to create a (binary) identifier of whether an individual has reappeared in court for a subsequent offence. This was achieved by identifying all those in the probation data that have a record in the court data where the offence_date (magistrates' courts) is greater than the initial referral_date to probation: **prevalence**.

It was also possible to calculate how many times an individual reappeared before the courts for subsequent offences following a referral to probation: **incidence**.

And the time (in days) elapsed between the original probation referral and any subsequent further offence: **duration/time elapsed**.

These three measures form the key outcome variables for this study, which will be reported on upon its completion.[12]

[11] The substantive outcome of interest is whether/how long after referral to probation before individuals come back before the courts again for a subsequent offence. In each case someone could reappear before the courts for historic offences, so it makes sense to remove these to align with window of data collection within the magistrates' courts, e.g. from 1st January 2011, and remove cases with invalid (e.g. 0010-04-21 and 5010-05-16) or historical offence dates outside of the study window of data collection.
[12] Accompanying R code can be found here: https://github.com/CarlyLL/Enforced-alcohol-abstinence.

Dealing with data limitations

In this study, Crown Court data was not additionally matched. This is owing to trade-offs in the minimal likely additional matches of cases heard in Crown Court[13] that did not get logged as originating in the magistrates' courts and quality of matching thresholds. As such, matching on only the magistrates' courts and probation datasets results in more confidence in the quality of the matches yielded.

Hitherto, no studies have linked the magistrates' courts and probation datasets to study person-events over time as represented in both. While a methodology for linking people in each dataset has been developed (*cf*. Linacre et al., 2022), there has been no approved approach to how one might link person-events across the two datasets. I adopted unique person identifiers as generated via Splink, as well as offence dates. The most relevant event dates upon which to match is not clear and a variety of options exist. For example, matching on offence date or administrative processing dates were both options. While offence dates can be multiple, historic, or the earliest date committed in a date range, and contain implausibly formatted entries, offence dates resulted in the most matches (when compared to administrative date fields) and were thought to represent the most intuitive dates upon which to match.

Suggested improvements recommended to data owners

Future users wishing to link probation and criminal court data would benefit from further metadata and documentation about the respective date fields contained in each. This should include what they represent, how they are populated and by whom, as well as which yield the most reliable matches for cases across the criminal justice datasets.

There should also be a baseline set of instructions in the guidance accompanying the data on how to link person-events across these criminal justice datasets. This would assist new users of this data and make it more accessible to a wider range of prospective users. Any future probabilistic methodology for linking person-events across these datasets would also be useful.

Likewise, providing further guidance on which dates yield the most reliable reconviction or reappearance measures would be useful. It would also ensure some consistency in approach across users of these datasets where this is appropriate.

[13] 95% of cases are dealt with at magistrates' court (HM Courts and Tribunals Service, 2023) and analyses only pertain to sentences served in the community (as opposed to immediate custody which are likely to have been sentenced at Crown Court). Moreover, all cases start in the magistrates' courts regardless of seriousness.

Additional data which would help to further develop the research

Although the probation data has been made available for research purposes via the Data First programme, details on risk assessment and criminogenic needs held by probation in the Offender Assessment System (OASys) are not, at the time of writing, currently available. Of specific relevance to this study would be information held on alcohol use captured by the Alcohol Use Disorders Identification Test (AUDIT). This would allow for further assessment of the association between drinking patterns, alcohol orders and reoffending or reconviction. OASys data is being added to the Data First Cross-Justice linked dataset and is anticipated to be available for research via the ONS Secure Research Service and the SAIL Databank in 2024. Interested researchers can contact Data First (datafirst@justice.gov.uk) to discuss and submit research proposals.

References

HM Courts and Tribunals Service (2023). About magistrates' courts. https://www.judiciary.uk/courts-and-tribunals/magistrates-courts/magistrates-court/ [Accessed 18/09/2023]

Jackson, T., Greyson, C., Rickard, I. and Tseloni A. (2022). Data First: Criminal Courts Linked Data. Ministry of Justice. Available at: https://www.gov.uk/government/publications/data-first-criminal-courts-linked-data

Linacre, R., Lindsay, S., Manassis, T., Slade, Z. and Hepworth, T. (2022). Splink: Free software for probabilistic record linkage at scale. *International Journal of Population Data Science*, 7(3). doi: 10.23889/ijpds.v7i3.1794.

Merrington, S., and Stanley, S., (2007). Effectiveness: who counts what? In (Ed.s) Loraine Gelsthorpe and Rod Morgan Handbook of Probation. Abingdon: Routledge. Chapter 15.

Ministry of Justice, released 09 November 2021, ONS SRS Metadata Catalogue, dataset, Ministry of Justice Data First Magistrates' Court Defendant - England and Wales, https://doi.org/10.57906/de97-0m89

Ministry of Justice, released 09 April 2024, ONS SRS Metadata Catalogue, dataset, Ministry of Justice Data First Probation - England and Wales, https://doi.org/10.57906/hvz4-m857

Ministry of Justice, released 16 April 2024, ONS SRS Metadata Catalogue, dataset, MoJ Data First cross-justice system linking dataset – England and Wales, https://doi.org/10.57906/0y39-4s34

Disclaimer

This work was produced using administrative data accessed through ONS Secure Research Service. The use of the data in this work does not imply the endorsement of the ONS Secure Research Service or data owners (e.g., HM Courts and Tribunals and the Ministry of Justice) 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/X003566/1].

Many thanks are also owed to Professor Ian Brunton-Smith who also acted as a mentor and advisor to me during this fellowship and advised on my data linkage strategy in particular.

Contact

Name: Dr Carly Lightowlers
Email: c.lightowlers@liverpool.ac.uk