

# Data Explained

---

## **Data First: Cross-justice system – England and Wales**

**Constructing re-offending measures for prison leavers**

Authors: Markus Gehrsitz, Sam Grant, and Stuart McIntyre

Date: February 2026

---

This Data Explained summarises experiences and learning from working with the Ministry of Justice Data First Cross-Justice System datasets while producing research on the effects of offender supervision on re-offending. This publication is intended to help guide future researchers using these data. Specifically, we outline how re-offending measures for prison leavers can be constructed from the data. We also provide feedback into future dataset development and documentation.

The administrative data discussed in this Data Explained was made securely available through project “Understanding offender rehabilitation and supervision” funded by ADR UK. The data used in this research project comes from the Ministry of Justice and were accessed through the ONS Secure Research Service via the University of Strathclyde. The data were 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

We analysed whether the introduction of post-release supervision for this offender group by way of the 2014 Offender rehabilitation Act (ORA) led to lower reoffending rates than what otherwise would have occurred. ORA mandated that all prisoners with a sentence of less than 12 months would now receive 1 year of supervision following their release.

Following their release from prison, offenders had to abide by conditions stipulated by their probation officer. These conditions placed restrictions, among other things, on offender's liberties, prohibiting who they could meet, where they could go, what they could possess and, in some cases, mandating compliance with drug testing.

We found that post-release supervision works well in both the short-run and the long-run. Both the share of prisoners who re-offend and the number of offences committed by supervised offenders is substantially and statistically significantly lower than for an all but identical unsupervised group. An overview of our findings is provided in a forthcoming Data Insights report.

In this Data Explained, we provide a walkthrough of how we calculated reoffending outcomes for prison leavers. We believe this will be of interest to other Data First users, since reoffending is typically the core research outcome of interest.

### Initial research questions

1. How do reoffending and reconviction rates change over time and by sentence type?
2. Which groups of offenders are at high risk of reoffending?
3. Are license conditions and offender supervision effective in reducing reoffending and reincarceration?
4. Do sanctions during probation periods act as a deterrent and thus improve public safety?
5. Which individual or contextual factors boost or diminish the effectiveness and deterrent effects of offender supervision?

### Research methodology

This study draws on detailed offender-level data from magistrates' and Crown courts, a prisoner custodial journey dataset, and probation data. The project combines these data sources to construct offender journeys through the criminal justice system.

We estimate the impact of offender supervision by using a natural experiment created by the 2014 Offender Rehabilitation Act (ORA). ORA introduced license conditions and close supervision for offenders released from short custodial sentences. Crucially, ORA only applied to offenders who were incarcerated for offences committed on or after 1 February 2015. These offenders form the treatment group in our study. Offenders who committed an offence that led to a prison spell just before this cutoff date were released unconditionally. These offenders form the control group in our study.

Our treatment and control group turn out to be all but indistinguishable in terms of their characteristics except that the treatment group was in fact supervised. A comparison of both groups within the framework of a regression discontinuity design, thus recovers the causal effect of supervision and license conditions on re-offending – where re-offending is defined as a further offence that results in a conviction by a court.

To implement our research strategy, we had to construct offender journeys. Among other things, we needed to determine whether and when ex-prisoners re-offended. Below we outline our approach in constructing these recidivism measures.

## Datasets and key linking variables

Calculating reoffending with [Data First: Cross-justice system data](#) is a complex process, requiring users to combine insights from 6 administrative datasets and a lookup table:

### *Cross Justice Person Look-Up*

This dataset acts as a lookup to identify which records in separate Data First datasets refer to the same individual. It consists of one row per record – e.g. a prison spell, a magistrates' case, a probation spell etc – with each one assigned a unique ID, called "row\_id\_hash". Appearances that MoJ have estimated reflect the same individual are grouped together under the same offender ID: "estimated\_xjs\_id".

<b>Key variable(s):</b>	row_id_hash	Unique record ID
	estimated_xjs_id	Unique Offender ID

### *Prisoner Flatfile*

Contains one row per prison spell, combining periods of remand and recall events. It provides insights into the type and length of a prison sentence, and some offender characteristics.

<b>Key variable(s):</b>	offender_book_id_hash	Unique prison spell ID
-------------------------	-----------------------	------------------------

### *Prisoner Movement Table*

Provides insight into all prisoner movements in and out of prison – for example, the date of prison reception, releases on temporary licence and final release dates. It can be linked to the prisoner flatfile using "offender\_book\_id\_hash", a unique ID for a prison spell. It consists of one row per movement.

<b>Key variable(s):</b>	offender_book_id_hash	Unique record ID
	movement_date	Date of movement in/out of prison
	movement_type_desc	Description of movement

**Magistrates' Flatfile**

Contains one row per defendant-case, listing a variety of information on key events occurring in court and the most serious offence heard in a case.

<b>Key variable(s):</b>	defendant_in_case_id	Person ID within magistrates' datasets
	case_id_hash	Case ID within magistrates' datasets

**Magistrates' All Offence Table**

Provides a breakdown of all offences heard within a single case in terms of their dates, type and resulting disposal. It can be linked to the magistrates' flatfile using defendant\_in\_case\_id and case\_id\_hash.

<b>Key variable(s):</b>	defendant_in_case_id	Person ID within magistrates' dataset
	case_id_hash	Case ID within magistrates' dataset
	offence_date	Offence date
	final_offence_ho_code	Home Office offence category
	finding_desc	The finding (decision) of the court given in the verdict relating to the offence
	jsas_result_group_desc	A broader grouping of disposals

**Crown Flatfile**

Contains one row per defendant-case and is the Crown court equivalent to the magistrates' flatfile. Only a small number of cases are transferred from the magistrates' to the Crown court, so this dataset is much smaller. It is possible to link these data at the person-case level, as opposed to the person-level, using a separate lookup table provided by Data First.

<b>Key variable(s):</b>	defendant_in_case_id	Person ID within Crown datasets
-------------------------	----------------------	---------------------------------

---

case_id_hash	Case ID within Crown datasets
--------------	-------------------------------

### ***Crown All Offence Table***

Provides a breakdown of all offences heard within a single case in the Crown court in terms of their dates, type and resulting disposal. It can be linked to the Crown court flatfile using defendant\_in\_case\_id and case\_id\_hash.

<b>Key variable(s):</b>	defendant_in_case_id	Person ID within Crown datasets
	case_id_hash	Case ID within Crown datasets
	offence_start_date	Offence date
	final_offence_ho_code	Home Office offence category
	convicted_rank	Identifier for convictions of trial cases
	case_type	Sentencing or trial cases

---

## Limitations in not combining datasets

Given the size of these datasets, researchers may be tempted to try calculating reoffending outcomes using just one or two of them. Below, we outline two such approaches and explain why each is likely insufficient.

A first approach would be to use the prisoner movement table to count the number of days between a prison release and (if applicable) the next prison admission. This is certainly do-able, and often an interesting outcome, but it is clearly not the same as measuring “time until re-offence”, which is typically of greater interest. Moreover, this approach cannot shed light on the volume of re-offences committed following a release.

A second approach would be to focus on the magistrates’ dataset and measure time between the date of a custodial sentence and a future offence. A drawback of this approach is that without observing release dates, you cannot easily adjust for the imprisonment period. This carries the risk of conflating rehabilitation with incapacitation. Adjusting for sentence length, which is observable in the court data, is unlikely to be a sufficient fix, since offenders are released at different points in their sentence (e.g. the halfway mark, following a parole review or at the sentence end) and may also be recalled if probation officers deem them to pose too high a risk to the public. As such, this approach is likely to generate substantial non-random measurement error in the outcome of interest.

In short, only by combining all datasets listed in the preceding section can we observe all the information needed to accurately measure reoffending among prisoners: release dates, re-offence dates, and case outcomes (i.e. whether a conviction occurred).

## 3-step approach to constructing re-offending measures

### ***Step 1: Creating an “exprisoner” lookup table***

We start by loading the `cross_justice_person` lookup and keeping only rows corresponding to a prison, magistrates’ or Crown court record. Since we are interested in calculating reoffending outcomes for ex-prisoners, we remove all individuals in this dataset with no prison record.

Next, use `row_id_hash` to (left) merge to the prison flatfile. If you are interested in a subgroup of prisoners – for example, prisoners sentenced to short custodial sentences - and not the full population, use information within the flatfile to filter for them.

After filtering, you can drop all variables merged in from the flatfile apart from `offender_book_id_hash`. Use this prison-spell identifier to merge in your release date of interest from the movement table. Note that prisoners tend to have several release dates even for a single sentence – for example, a release on bail or re-release following a recall. Consider which type of release you are interested in measuring reoffending against.

For example, if you were interested in re-offending while on temporary licence, only these release dates should be merged in. The type of release we chose for our project was an offender's first post-custody release, identified by removing bail releases from the movement table and then selecting the earliest remaining release date.

Again, it is important to note that many individuals have multiple prison spells and thus multiple releases. We found it useful to create a variable "*release\_number*" which orders an offender's prison spells chronologically. It is also key in how we calculate reoffending rates in Step 3.

We now have assigned the relevant release date to each prison spell. Remember that our dataset still contains rows on magistrates' and Crown court records. We next re-use *row\_id\_hash* but this time to (left) merge to the magistrates' court flatfile. The only variables you need to pull over from the flatfile are *defendant\_in\_case\_id\_hash* and *case\_id\_hash* - these are the person and case identifiers that are later used to link with the *all\_offence\_table*.

Repeat this procedure for the Crown court rows again using *row\_id\_hash* as matching variable.

Finally, save this dataset as "exprisoners\_lookup". It will serve as the "spine" dataset that we will link our offending outcomes to.

### **Step 2: Merging offending outcomes to the "exprisoner" lookup table**

Our next objective is to create a dataset that we can use to generate re-offending outcomes in step 3. To do so, we need to merge the magistrates' and Crown all\_offence information into the magistrate and Crown rows in our "exprisoners\_lookup" dataset. Put differently, we want to turn the magistrates' and Crown court rows in "exprisoners\_lookup" into unique offender-case-offence rows.

To enable the merge, first keep only the magistrate rows in "exprisoners\_lookup". Each row is now uniquely identified by the variables *defendant\_in\_case\_id\_hash* and *case\_id\_hash*. Execute the (left) merge to the *magistrates\_all\_offence* table and keep only the variables relating to offence dates and types, disposals and verdicts. These variables will enable us to identify re-offence dates and types, cases transferred to the Crown court, and re-convictions.

Next, it is important to delete magistrate rows where the case was transferred to the Crown court for a trial or sentence.<sup>1</sup> Since we are about to merge in the Crown court data separately, we don't want to double count offences heard in both a magistrates' and Crown court. You might wonder why we need to merge in the Crown court data separately at all given that virtually all cases start in the magistrates'. The answer is that merging in the Crown court data enables us to determine if an offence ultimately resulted in a re-conviction.<sup>2</sup>

---

<sup>1</sup> You can use the variable *jsas\_result\_group\_desc* from the magistrates' all offence table to do this.

<sup>2</sup> For cases transferred to the Crown court for sentencing, as opposed to trial, having the Crown court dataset merged in to our "exprisoners\_lookup" dataset enables one to observe final disposals. This may be of interest if one wishes to calculate outcomes such as the number of future custodial disposals.

Save this dataset as “mag\_offence\_rows” and repeat this procedure for the Crown court rows in the “exprisoners\_lookup” data. Make sure when you create “crown\_offence\_rows” that you rename the variables to be consistent with those in “mag\_offence\_rows” (e.g. rename, *offence\_start\_date* to *offence\_date*).

Finally, re-load the original “exprisoners\_lookup” and now keep the prison rows only. Vertically append in “mag\_offence\_rows” and “crown\_offence\_rows” so that now each row represents a unique prison spell or offender-case-offence. You can think of “mag\_offence\_rows” of an expanded version of the magistrates’ rows that were initially in the “exprisoners\_lookup” dataset.

Each magistrates’ or Crown row is now treated as a possible re-offence for a given prison row in the data. For magistrates’ court rows, use the variable *finding\_desc* to drop any offences that the defendant was not found guilty for. Crown court rows represent offences within cases either on trial or for sentencing; drop any rows where the trial did not result in a conviction, using variables *case\_type* and *convicted\_rank*. Finally, flag magistrates’ and Crown court rows by creating a *re\_offence* variable that equals 1 for these rows and 0 for prison rows. Save this dataset as “exprisoner\_lookup\_with\_offending\_outcomes”

### **Step 3: Calculating re-offending outcomes**

The dataset created in step 2, “exprisoner\_lookup\_with\_offending\_outcomes”, contains all the information needed to create re-offending outcomes. For each prisoner we have their relevant release dates in the prison rows. In the dataset we also have rows for magistrates’ and Crown court convictions (including the offence date that led to the court case) of each prisoner. We now need to determine the timeline between prison release and offence-date (if applicable).

The challenge here comes from the fact that a particular offender (as identified by *estimated\_xjs\_id*) may have multiple prison spells and hence multiple releases with which to measure reoffending against. For example, a particular re-offence might constitute a re-offence within, say, the 10<sup>th</sup> week since release A but the 3<sup>rd</sup> week since release B. This means we can’t create a single “*days\_between\_release\_and\_reoffence*” variable. A more sophisticated approach is required.

Our solution was to work with one release at a time, looping over each *release\_number* created in Step 1. In other words, we first calculated reoffending outcomes with respect to all first-time releases, then second-time releases and so on, up to the maximum number of releases an offender had in our sample. This was achieved through the following steps:

For each release number  $i$  (where  $i$  ranges from 1 to  $N$ ):

- (1) Import “exprisoner\_lookup\_with\_offending\_outcomes”
- (2) Populate all court rows with each prisoner’s  $i^{\text{th}}$  release date.

- (3) Drop historical offences if the following applies: *offence\_date* < *release\_date*, that is drop if the offence date pre-dates an offence and thus cannot be a re-offence
- (4) Keep comparing *offence\_date* and *release\_date*: Generate a binary variable "*reoffend\_week1*" that equals one if a re-offence occurs within 7 days of release. Use a for loop to repeat this for *reoffend\_week2*, *reoffend\_week3*, ..., *reoffend\_week52* (for weekly reoffending indicators up to one year).
- (5) Collapse/aggregate the dataset to the offender level, keeping a record of the release number of the current iteration and summing all *reoffend\_week\** variables to provide counts of the number of re-offences occurring within each week since release. Replace missing values of the summed *reoffend\_week\** with 0s as these are cases where no offence was found, suggesting that no re-offending took place. These variables are our *re-offence volume outcomes*.
- (6) Generate a binary variable "*ever\_reoffend\_week1*" that equals one if *reoffend\_week1* > 0. Use a loop to repeat this for *reoffend\_week1*, *reoffend\_week2*, *week3*, ..., *reoffend\_week52* (for reoffending indicators up to one year). These are our *time until re-offence outcomes*.
- (7) Save this *i*<sup>th</sup> dataset as "*reoffending\_outcomes\_with\_respect\_to\_release\_i*". There will be one dataset created for each *i*<sup>th</sup> iteration.

Note, it is straightforward to adjust this logic to construct reoffending outcomes for certain types of offences. For example, theft re-offending outcomes can be generated by writing an additional line of code after (3) that drops all non-theft offence rows (use *final\_offence\_ho\_code* to do so).

Once the loop is completed, append vertically all "*reoffending\_outcomes\_with\_respect\_to\_release\_i*" datasets together. The dataset should have one row per offender-release and have the two types of reoffending outcomes for each week since release: *reoffend\_week1*, *ever\_reoffend\_week1*, *reoffend\_week2*, *ever\_reoffend\_week2*, and so on..

The final step is to 1:1 merge this dataset to the prison rows in "*exprisoners\_lookup*" using *estimated\_xjs\_id* and *release\_number* to re-obtain *row\_id\_hash*. This will enable you to merge in other variable that are of use to your project.

Note that the re-offending outcomes constructed here define re-offending as an offense that results in a re-conviction. This is different from the MoJ's definition of "proven re-offending" which draws on the Police National Computer and includes cautions in its measure of recidivism.

## **Outlook and recommendations to data owners and users**

Recidivism and re-offending tend to be key outcomes in criminal justice related research. Yet there is little publicly available information on how to use the novel and uniquely rich Data First: Cross-Justice System datasets to construct these crucial outcomes. We hope that this Data Explained report at least in part remedies this issue. Researchers who follow our 3-step approach outlined above, should be able to construct re-offending outcomes for prisoners (and various prisoner subgroups) released from custodial sentences. As such this essay should be useful to both current and future data users.

It should be noted that an important advantage of our focus on ex-prisoners is that we were able to use the prison files as a launch pad for our linkage. This option may not always be available. For example, research on re-offending of offenders sentenced to, for example, community sentences, would have to take a different starting point. With that being said, we outlined a series of principles and data processing approaches that data users can leverage to pull together information from the very rich Data First: Cross-Justice System datasets.

## Disclaimer

This work was produced using administrative data accessed through the ONS Secure Research Service. The use of the data in this work does not imply the endorsement of the ONS data owners (e.g., HM Courts and Tribunals Service, HM Prison and Probation Service, and the Ministry of Justice) in relation to the interpretation or analysis. This work uses research datasets which may not exactly reproduce Accredited Official Statistics aggregates. Accredited Official 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/Z503289/1]

## Contact

Name: Dr Markus Gehrsitz

Email: [markus.gehrsitz@strath.ac.uk](mailto:markus.gehrsitz@strath.ac.uk)

