Ministry of Justice

Justice Data Lab

A Peer Review of existing methodology:

Investigations and conclusion

October 2017

Contents

1. Summary

The peer review of the methodology used by Justice Data Lab (JDL) for comparison group analysis to assess the impact of rehabilitation interventions on recidivism was reported on in March 2016[1]. Since then, a programme of investigations have been undertaken to identify possible enhancements to the process. Whilst no major improvement has been identified, it has been a valuable exercise to explore a number of possibilities as the JDL team always seek to improve and develop the analyses produced. Whilst this peer review process has concluded, the JDL team will continue to listen to feedback about aspects that could advance the service offered.

The key findings and conclusions of the investigations following the methodology review are detailed below:

- The peer review process has been beneficial to establish that there is no major change that would enhance the JDL methodology.
- Investigating **different kernel matching** approaches did not lend itself to one kernel consistently performing higher than the current approach. Utilising **coarsened exact matching** would require a substantial reduction in the number of observed variables to retain as much of the treatment group as possible. **Regression discontinuity designs** may be suitable for a small number of JDL requests but is not an approach that can be universally applied (sections 4.1-4.3).
- There is no notable effect on analysis outcomes when looking at **refining the JDL model,** and so the original full model will continue to be used as a starting point (section 5).
- Using alias information for **placebo testing** generally supports the assumption that the current process is not significantly impacted by the effect of unobserved variables (section 6.1).
- Further research suggests that assessing the **impact of dummy interventions** is not suitable for the established JDL approach. Additional **sensitivity analysis** was investigated above that promised within the response paper, with the aim of being developed in the future (sections 6.2 and 6.3).
- Current data retention practices will be reviewed in line with data protection developments as they become established and consulted on with stakeholders when suitable (section 7).

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/506327/methodology-review-response.pdf

2. Peer review background

Following the launch of the JDL in 2013[2] to enhance the evidence base for rehabilitation programmes, using aggregate re-offending analyses provided by the JDL service, a peer review of the original methodology was conducted during 2015/16. A range of interested parties took part in the review, from the Justice sector, across Government and from academia, to ensure that the methodology in place is appropriate and robust, and to identify any areas for improvement. The JDL team considered the feedback in conjunction with internal methodological experts and published the response paper published in March 2016.

This paper covers the key points outlined for investigation in the review response paper and how the JDL team addressed these aspects, identifying the impact and applicability to the JDL service.

3. Responses clarification

In the response paper, the feedback given by those involved were summarised within each question/section. For clarification, all responses were included in such summaries at a high level and not verbatim. Please use the contact details at the end of this report to request full unedited responses.

4. Test cases and detailed results

Across the various investigations that the JDL team undertook, it was important to test the results on a range of data. Test cases were selected from existing published analyses where suitable data remained available and aimed to cover a mix of sizes (whether the treatment group is deemed to be large or small), settings (whether the intervention was prison or community-based) and focus (for example, addressing accommodation or education needs). This paper is intended to provide a summary of the investigations and findings that is accessible the majority of its potential audience, rather than go into full technical detail. Further details of any analyses referenced in this paper and full results can be made available on request.

5. Alternative statistical approaches

Section 4.1 of the methodology response paper raised the question of whether one-to-many radius propensity score matching (PSM) was the most suitable approach for assessing the impact of a treatment/intervention, and what other methods should be considered. One-to-one matching without replacement and kernel-based PSM were initially proposed, with a consensus from the panel that kernel-based matching should take priority. Separate to the methodology review, coarsened exact matching and regression discontinuity designs had been noted as alternative techniques to explore.

[2] With a 2 year pilot phase, becoming a permanent service in April 2015.

5.1 Kernel-based propensity score matching (PSM)

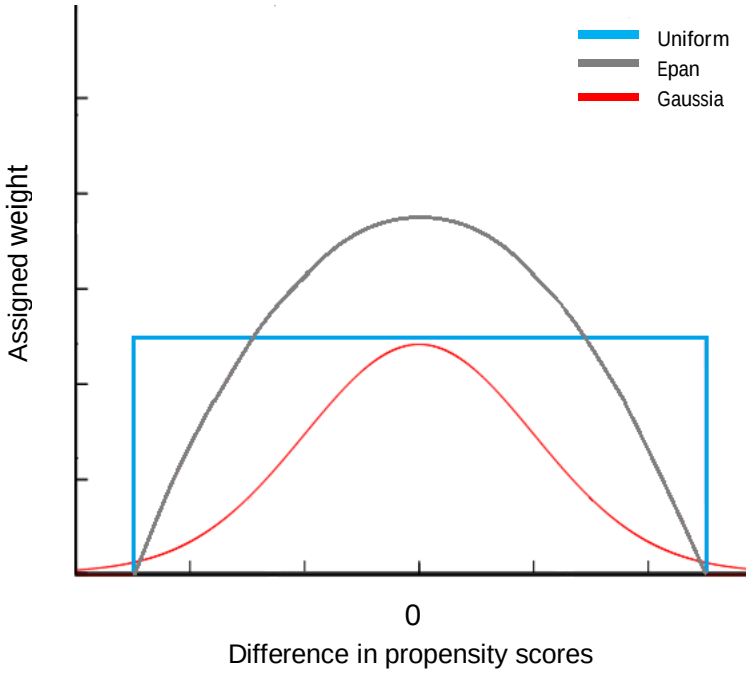
In PSM, the kernel refers to the model used to assign weights to a comparison group member based on the difference in propensity score to the treatment group member they are matched to. The standard radius matching method for PSM used by the JDL uses a 'Uniform' kernel when matching the treatment and comparison groups. The Uniform kernel assigns equal weighting to matches that fall within the agreed caliper[3]/radius, regardless of how close the two propensity scores are to one another. The investigations tested two other kernels and compared the findings with the original analysis (i.e. that with a Uniform kernel) across several test cases.

The Gaussian (also referred to as 'normal') and Epanechnikov (Epan) kernels were selected as alternative kernels to investigate, as they are both widely used in statistical analysis.

The Epan kernel weighs different comparison and treatment group members according to how closely their propensity scores match within the agreed matching range (i.e. bandwidth), with more similar propensity scores receiving a greater weighting than those less similar. For the purpose of this investigation, the caliper/bandwidth was kept constant across the different kernels. The Gaussian kernel differs in that, while the closer the propensity scores, the greater the weighting, there is no cut-off point, and so all matches are included regardless of how far apart the propensity scores. Those with large differences in propensity scores are given very small weighting.

Figure 1 illustrates the weight and matching methods of the three kernels used in the current analysis, with the horizontal axis illustrating the distance from the propensity scores and the vertical axis showing the assigned weight.

Figure 1: *Illustrative guide to the Uniform, Epan and Gaussian kernels. Actual bandwidths may vary*



[3] A caliper determines the maximum difference in propensity scores between a treatment and comparison group member that will be accepted as a match.

Investigations: The overall methodology remained unchanged from the existing JDL approach and the matching uses the Uniform kernel by default. Equivalent matching ranges were used for the Epan and Uniform kernels. The only difference in these tests was that the alternative kernels were specified in each run and the established results (i.e. the Uniform kernel) were used as a baseline.

Findings:
Matching quality: Looking at whether the different comparison groups are more closely matched than previously, the results are mixed. The impact on the standardised mean differences[4] were varied across the test cases, with three cases showing the smallest average of the standardised mean differences across the variables for the Epan kernel, one for the Gaussian kernel and one for the Uniform kernel. A larger sample of test cases may determine if the kernel has a consistent impact on standardised differences.

Matching rates: The match rate for the Epan and Uniform kernels remained the same across all analyses, as the equivalent agreed matching range was used across both analyses. When using the Gaussian Kernel, all of the matched treatment groups matched to a comparison group member. This led to an increase in the match rate for two analyses where less than 100% of the treatment group had matched in the original analysis.

Comparing to original results: In all but one of the analyses, the significance level remained the same across all three kernels. More importantly, there was no difference between whether a finding was significant or not significant when using the different kernels.

Recommendation: In most cases, the Gaussian kernel can allow for all linked treatment group members to match to a comparison group member as there was no cut-off point imposed. While the ability to give a better representation of the full sample can add greater statistical power, it is worth considering whether those members with such dissimilar propensity scores should be matched. The matching quality for the two analyses where the treatment group match rate increased using the Gaussian kernel, was inconsistent. In general, the attrition rate at the PSM stage is very low so the match rate is not a great concern. The investigation has shown the Uniform kernel to be a sufficient matching method.

5.2 Coarsened exact matching (CEM)

CEM works by matching on coarsened variables, i.e. it applies exact matching on each binary variable or within bands for each continuous variable. For example, we may wish to band (or 'coarsen') the age when the offender first entered the criminal justice system into meaningful age groups (under 18, 18-29 etc.) and then only match treatment and comparison group members within this band. We also wish to only match males to other males. This can be done for multiple variables so that a young white male from the UK would only ever match to other young white males from the UK, or fall out of the treatment group if such a specific match is not possible. In contrast, PSM matches individuals on their overall propensity to have received treatment based on a range of characteristics, which can mean that two matched people can potentially vary considerably on some variables but match if overall they are similar when looking across a wide range of characteristics.

[4] Standardised mean difference is a measure of distance between two group means and is used to measure balance between observed characteristics. Smaller standardised mean differences indicate the treatment and comparison groups are better matched on the variables.

Investigations: Initial attempts to match using CEM were made by only using a small number of continuous variables which were coarsened both manually and automatically via the available CEM software, with a large sample test case. Once all continuous variables were included, binary variables were added. Examples of binary variables used include gender and nationality as well as Offender Assessment System (OASys) variables[5]. The number of matches found for each combination of variables included and coarsening levels were recorded to determine the final coarsening levels to complete the analysis. A range of test cases were used and the results compared to their published findings using PSM. The same variables used in the original analyses were used, with the exception of the squared versions of certain variables as these were not considered useful for CEM[6] .

Findings:

Matching Quality: When comparing the matching quality between the treatment and comparison group, categorical outcomes (such as ethnicity and OASys variables) had smaller standardised mean differences between the groups. However continuous variables had larger differences. This is not to say that they matched less well on an individual level, but that the two groups overall were less well matched on these variables.

Matching Rates: Using a large test case, when only two variables were included in the model (age at index date[7] and age when first entered the criminal justice system), over 99% of the treatment group were able to match with a comparison group member. However, JDL analyses usually include over 100 variables in the final model, so including two variables is not sufficient to account for differences between the two groups. As variables were introduced to the CEM model, the number of matches reduced. Match rates varied depending on which combinations of coarsening were employed, however the optimum rate was found to be coarsening manually on age at index date, age first entering the criminal justice system and the Copas rate[8] and to coarsen all other variables automatically.

When excluding OASys variables, 54% of the linked treatment group matched to a comparison group member. This fell to 12% when OASys information was included. When OASys variables are included, this leads to over 70 additional binary variables giving a complex profile of a treatment group member that is difficult to replicate exactly in the comparison group. Substantially lower match rates were common across the test cases as illustrated in figure 2.

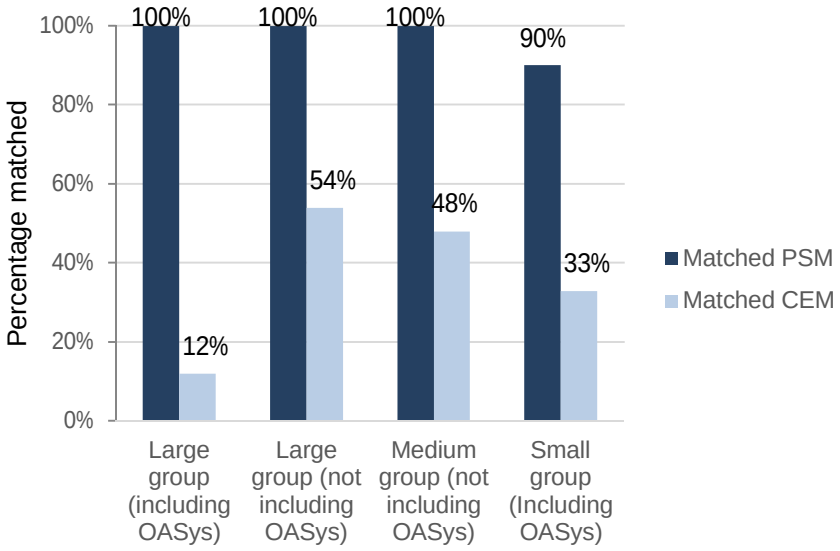
5 OASys data measures a wide range of information about the criminogenic needs and issues of offenders, such as their drug usage, accommodation needs and relationship history.

6 Squared variables are used in PSM as in some instances, they can provide extra information about their relationship to the propensity to reoffend or to be part of a treatment group. For matching using CEM, there would be no additional information gained form matching on the squared variable.

7 The index date refers either to the date an offender was released from custody, or the start of a community-based sentence.

8 The Copas rate indicates the rate at which an offender has built up convictions throughout their criminal career. The higher the rate, the more convictions an offender has in a given amount of time.

Figure 2: *Summary of the percentage of the matched treatment group matched when using PSM and CEM.*



Comparing results to original analysis: When comparing the reoffending rate of both the treatment and comparison groups to that of the original JDL analyses, in every test case the reoffending rate was lower in the CEM analysis than the PSM analysis, with a difference of up to 14 percentage points. This suggests that reoffenders are disproportionately excluded in the matching procedure in CEM, possibly demonstrating that reoffenders have a more complex profile that is not straightforward to mirror in a more directly matched comparison group than those who do not reoffend. This raises a possible concern of the reoffending rate being misrepresented in the analysis.

The results showed similar differences between the treatment and comparison groups as the original analyses looking across the three main reoffending measures[9], however in the large sample group where the difference had been statistically significant, it was not statistically significant with CEM. This was also the case with the small sample group for the frequency of reoffending. This is likely to be due to the lower match rate giving a smaller treatment group size. The medium test case changed from a non-significant difference in rate of reoffending and time to first reoffence, to showing a significant reduction in rate of reoffending and time taken to first reoffence.

Recommendation: Looking across academic research, a similar comparison of CEM and PSM was conducted by Thompson (2014)[10] who found that it was not feasible to complete the CEM analysis with the full set of base variables and used a reduced data set. They also concluded that their study did not provide evidence of better matching with CEM over PSM (however, it is noted their findings are limited to a single dataset).

[10] Thompson, D.,(2014) 'Comparing regression, propensity matching and coarsened exact matching in healthcare observational studies using SAS®: An example from the Medical Expenditure Panel Survey (MEPS)', Blue Cross Blue Shield of IL, MT, NM, OK & TX, Chicago, IL

However, King and Nelson (2016)[11] argued that PSM can lead to greater imbalance and proposed CEM as an alternative method to reduce imbalance. In the current JDL PSM analysis, standardised mean differences are used as an indicator of balance and the caliper is set at the optimal level to maximise matching quality.

On consideration of such research and the findings outlined above, CEM is not thought to be an appropriate alternative to PSM for the JDL due to the number of variables used in the matching procedure. For small sample sizes, it is likely the match rates would not provide enough matches meet the minimum threshold of 30 people required for robust analysis and fewer organisations would be able to use the JDL service. For CEM to be employed, the number of variables used in the matching procedure would need to be substantially reduced, including the removal of OASys variables. However, the inclusion of OASys information was a welcomed development of the JDL in 2015 and removing this data source is not a viable option.

5.3 Regression discontinuity design (RDD)

An RDD can be used when an intervention is given to individuals for whom a measured characteristic or variable lies on one side of a cut-off point, for example, those scoring above a certain test score receive the intervention and those scoring below the cut-off do not. RDD is based on the assumption that those just above and below the determined cut-off point are similar enough to one another that any difference in outcome can be attributed to the intervention. The key benefit of this model is that this could potentially account for unobservable as well as observable differences seen in those selected for an intervention and those who are not.

Investigations: Looking across the 20 most recent JDL requests, only two stated that a selection criterion involving a fixed cut-off point was used. In both cases, offenders with an Offender Group Reconviction Scale (OGRS) score[12] above a certain cut-off were accepted. It is worth noting that the selection criteria section of the JDL request form is a free text box and there could potentially be other requests that use such criteria but who did not disclose this when submitting data to the JDL.

In such cases, it may be possible to employ an RDD approach. In this scenario, the reoffending rate of those scoring just above the required OGRS score would be compared to those scoring just below the threshold. A significant difference would imply any reoffending impact can be attributed to the intervention itself, rather than other differences. This would be based on the assumption that those with a score just below the cut-off are fairly similar to those with a score just above the cut-off, and that the difference in outcome is due to the intervention, not the criterion itself. To complete such analyses, the assumption would be that this criterion is adhered to strictly i.e. that all those above the cut-off are given the intervention, and all those below are not. If this assumption is not met, a 'fuzzy RDD' can be used instead, as it does not require the cut-off point to be strictly adhered to as long as the probability of receiving treatment is different either side of the cut-off.

[11] King,G., Nelson,R., (2016) 'Why Propensity Scores Should Not Be Used for Matching'
[12] The Offender Group Reconviction Scale (OGRS) is a predictor of re-offending based on risk factors known to be associated with the likelihood of re-offending

Most JDL requests use other criteria that would not lend themselves to RDD analysis. For example, including offenders having particular employment needs, gender of the offender and absence of drug misuse or particular sentence types. As these are not continuous variables, RDD would not be appropriate for these analyses.

A further limitation to this method is that while it can provide robust estimates of causal impacts, the method would only really compare the reoffending rate on either side of the cut-off point as opposed to the impact on the whole cohort. The repercussions of this are that the results will not reflect those for the whole treatment group and a bigger reoffending impact will need to be obtained to obtain a statistically significant conclusion.

Recommendation: While RDD may enable a more robust estimate of the reoffending impact of interventions for a very small number of JDL requests, this would entail using different methodology for different requests and would have resource implications. Furthermore, it would not allow us to measure the impact across all participants, with the focus around the specific cut-off point for eligibility to a programme. Further investigation could involve running RDD alongside PSM for analyses which employ particular cut-off points that are observable in the administrative datasets employed by the JDL or available elsewhere across the MoJ. Such exploratory work is dependent on available resource.

5.4 Refining t-test approach

Originally the JDL process compared the extremes of the two 1-sample t-test confidence intervals for the treatment and comparison groups to establish the range of difference between their re-offending rates. This was for communication purposes for a mainly non-technical audience to avoid the use of p-values. However, it was readdressed through both the methodology review responses and through the JDL team revisiting the original method, concluding that this needed to be amended to using a two-sample t-test, which has been in place since October 2015.

Switching the approach, differences would be most likely to be observed when the treatment and comparison groups are similar in size. However, the JDL treatment groups are generally small in relation to their comparison group. As such, it would be highly unlikely there would be visible differences to the effect size and the switch would not have changed any headline measures published prior to implementing the change.

The previous method would have overestimated the range slightly and this overestimation can be assessed approximately by considering the output for each analysis of this formula (CI = confidence interval, N is the size of the group referred to):

$$CI\_old \sim= CI\_new + CI\_new*sqrt(N\_treat/N\_control)$$

6. Refining the JDL model

A concern from the JDL methodology peer review was that the logistic regression model used by the JDL to match the treatment group to a comparison group of similar people would generally be over-fitted (due to the large number of variables used) and therefore could be deemed as being unreliable. The JDL agreed to look into how/if the variables used in the JDL analyses could be refined (section 4.6 in the review response), bearing in mind that much of the academic community focus on matching diagnostics rather than model parsimony, as does the JDL.

Three approaches for reducing the number of variables in the full JDL model13 were investigated using several test cases. An overview of the three approaches and the outcome of applying them to the test cases is outlined below.

Investigations: In the first approach, ten JDL analyses were reviewed to assess which variables were frequently removed during the backwards stepwise regression procedure (i.e. because they were predictive of neither treatment nor reoffending behaviour at the p=0.2 level[14]). Of the variables that were removed in three or more of the analyses, 89% were the squared counterparts of previous offence variables (for example, the number of previous robbery offences, squared). All such variables, where they appeared in the final model of the test cases, were removed in this approach (eighteen variables in total). Excluding these variables from the full model removes the chance that they could be included in the final model in error (i.e. where they may appear to have some predictive power but are actually reflective of statistical noise).

The second approach reviewed correlation matrices of several JDL analyses to identify any variables that are closely related, and whose variance may already be accounted for by other variables in the model (**Note:** if any variables in the model are too closely correlated in JDL analyses, the regression model fails to run and the appropriate variables are removed anyway). Some of these variables are subsets of other variables in the model (for example, the number of previous offences is the sum of the number of previous violent offences, theft offences, criminal damage offences etc.) in which case the variable containing the overall total is selected for removal. Eight variables were identified for removal: the number of previous offences (and the squared counterpart), the number of previous conviction events (and the squared counterpart), Copas rate, the squared age at index offence, the flag for employment in the month prior to conviction, and flag the for out-of-work benefits in the year prior to conviction. Removing these variables from the model is expected to improve the stability of the model.

[13] The current JDL methodology starts with a 'full' regression model consisting of all relevant variables held for the analysis, which is reduced to a 'final' model using a backwards stepwise procedure to remove unnecessary variables. These investigations looked into excluding certain variables from the 'full' model.

[14] These are the p-values for each variable in the model at each stage of the backwards stepwise regression procedure, for predicting treatment and for predicting reoffending. If the p-value of a variable is below 0.2 in either of the models, then it will remain in the model for the next stage of the backwards stepwise procedure.

The third approach investigated the effect of excluding variables related to the severity of offences. This is as the JDL moves towards analysing reoffences by court outcomes (either summary, triable, or indictable offences) rather than offence severity tiers15 (tiers 1 – 3) to align with other statistics published by the Ministry of Justice that no longer provide the JDL with updated severity tier data. Nine variables were affected in this approach: one for each tier to flag the severity of the index offence, the counts of previous offences within each severity tier and the squared counterparts of these. Removing these variables from the model ensures that the comparison group is not selected based on variables that may become redundant.

A number of diagnostics were used to assess the effect of these approaches, including changes to the estimated treatment effect and the associated p value, changes to the standardised differences for each of the variables (which quantify how well matched the treatment and comparison groups are) and the size of the treatment and comparison groups.

Findings: None of the approaches resulted in a notable change in the magnitude of the treatment effect, or direction of the results (i.e. if the reoffending rate of the treatment group was lower than that of the comparison group in the original analysis, then it remained to be so when additional variables were removed from the model). Despite a slight increase in the size of the comparison groups in each of the approaches, the size of the confidence intervals and the statistical significance of the results remained unchanged (i.e. if the estimated treatment effect was not statistically significant in the original analysis, then it was not significant in any of the investigations either). The standardised differences for the variables remaining in the final models were not noticeably smaller, indicating that the quality of the matching had not improved.

Recommendation: Given that there was no notable effect of refining the JDL model on the analysis outcomes, the full model will continue to be used as the starting point for the backwards stepwise regression in future analyses. As long as a suitable comparison group can be selected, whether it reflects some characteristics of the treatment group that are not predictive of treatment or reoffending is of secondary importance relative to the need to retain additional variables that may otherwise bias the estimated reoffending differences that would be attributed to the intervention (see Section 5). For example, if the age of a particular cohort was not predictive of treatment and reoffending behaviour (but appeared to be so, due to noise in a small sample), the investigations indicate that there is no harm in matching to a comparison group of a similar age, regardless, as long as this does not limit the suitability of the comparison group on other factors that do influence allocation to treatment or reoffending behaviour. However, if age was not included but is actually predictive of treatment and reoffending behaviour then the estimated impacts will be biased.

[15] Offences are classified into three tiers of severity, with tier 1 being the most severe. For a list of offence types in tiers 1 and 2 (serious acquisitive offences are tier 2) see pg 22-26 of: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/472535/proven-reoffending-definitions-measurement-Oct15.pdf

7. Testing robustness of JDL methodology

A proposal from the JDL methodology peer review was to investigate the likelihood of bias in the estimated outputs of JDL analyses due to the presence of unobserved variables. The JDL agreed to investigate this further under the current methodology (section 4.11 in the review response).

It is important to note that there are almost certainly factors that are not accounted for in the JDL methodology when matching treatment and comparison groups. However, these variables would need to vary systematically between treatment and comparison groups **and** (substantially) affect reoffending behaviour to bias the JDL results. As noted in section 4 of this paper, the JDL incorporates an extensive list of available variables in the matching process; any variables that we have access to and may be in some way related to treatment allocation or reoffending behaviour are included in the propensity score matching regression model. In doing so, the aim is to increase the likelihood that any important confounding variables are accounted for. The JDL aimed to test this assumption using a number of falsification exercises, including placebo testing, dummy interventions and a sensitivity analysis, as outlined below.

7.1 Exploring the effect of treatment on placebo outcomes

Placebo testing involves identifying an outcome variable that could not feasibly have been affected by treatment (ideally something that occurs *before* treatment) and is in some way similar to the outcome of interest (in this case, reoffending) so that any unobserved variables that affect the measured outcome may similarly affect the placebo. By substituting this placebo outcome for the usual reoffending outcome variable, one is able to test the relationship between treatment and the placebo using the established JDL procedure. As the placebo outcome could not feasibly have been affected by the treatment, one would expect that no significant relationship between treatment and the placebo would be identified. If the test does show a statistically significant relationship, this could indicate that there are unobserved variables not accounted for in the PSM model that affect both allocation to treatment and the placebo outcome (and, by extension, reoffending behaviour). See figures 1-3 below.

Figure 3. *JDL assumption 1: treatment may influence reoffending behaviour*

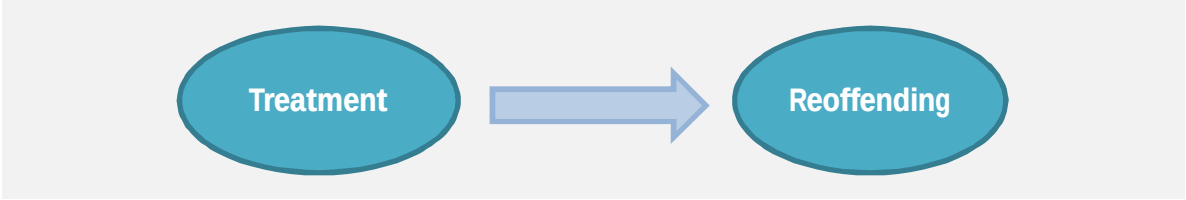


Figure 4. *JDL assumption 2: there may be variables that influence allocation to treatment and reoffending behaviour, so these variables are included within the PSM model to remove bias*

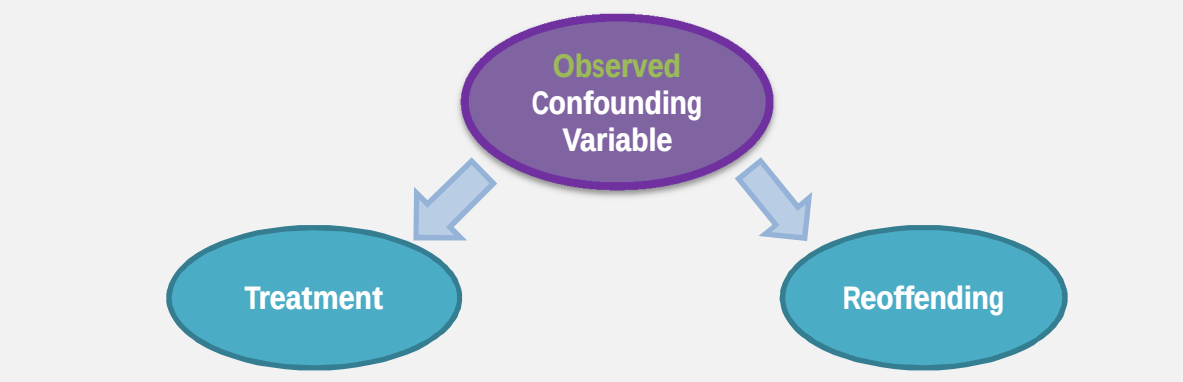
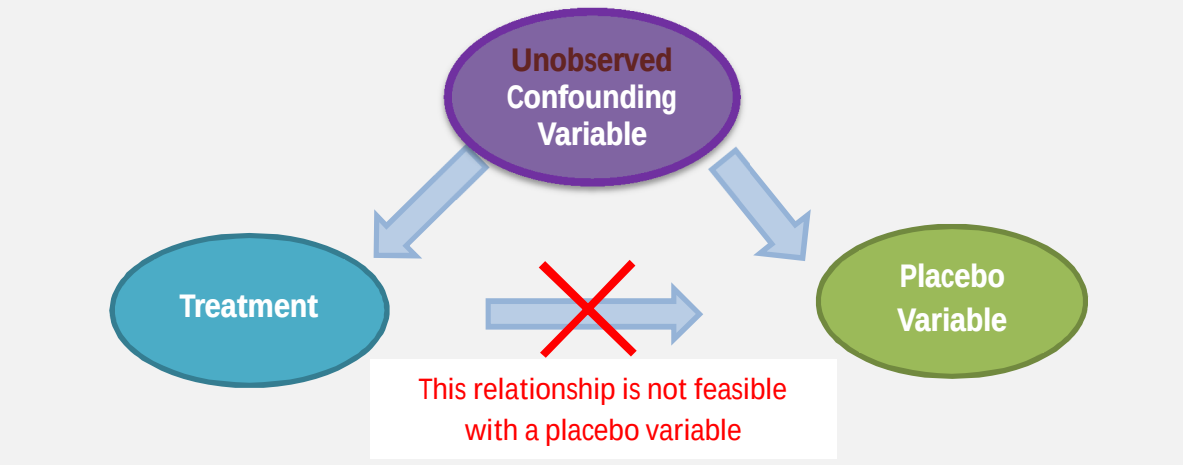


Figure 5. *Placebo testing: if there is a relationship between treatment and the placebo variable after controlling for variables in the PSM model, this may indicate the presence of an unobserved confounding variable i.e. the JDL assumption for unbiased use of PSM does not hold*



Investigations: Given that so many variables available to the JDL are accounted for when matching the treatment group to a similar comparison group, identifying a suitable placebo variable from the remaining selection is not without limitations. Firstly, the direction of causality between variables is not always clear-cut. When using placebo testing in medical experiments, identifying the direction of causality can be much simpler due to clear biological constraints (for example, it would be implausible for a child's eating habits to affect the birth weight of a baby, given that this occurs after the birth weight has been measured). With regards to JDL analyses, it is more difficult to identify a variable that could not plausibly have a) been affected by the treatment or b) affected allocation to treatment. The second limiting factor is identifying a variable that is similar enough to reoffending behaviour that one might expect it to be affected by the same unobserved variables.

With these limitations to consider, the chosen placebo variable was a binary outcome of whether the offender had ever used an alias (either name or date of birth). Alias use was considered to be similar to reoffending behaviour in the sense that it is an indicator of undesirable behaviour, supported by investigations showing greater alias use in reoffenders compared with non-reoffenders across multiple test cases (75-88% of reoffenders had used an alias compared with 47-61% of non-reoffenders; a correlation of r=0.3 in each test case). While the data available did not specify a date that the alias was used (i.e. to determine whether it was measured before or after treatment) it is unlikely that alias use would occur for the first time after receiving treatment designed to reduce reoffending (and therefore feasibly influenced by treatment).

Findings: When substituting the reoffending indicator for the placebo variable, one of the three test cases showed a significant reduction in alias use (for name, date of birth and overall alias use) for those in the treatment group. This could indicate that treatment did in fact influence alias use for this group (given that it was not possible to restrict the placebo variable to alias usage prior to treatment). Alternatively, it could indicate that an unobserved variable influenced allocation to treatment and also alias use. The reduction in alias use in the treatment group was small (1.8 percentage points for the overall test) but due to the size of the treatment and comparison groups, this carried enough statistical power to be considered a significant difference. All other test cases returned no significant effect of treatment on alias use.

Recommendation: While these investigations have limitations, the outcomes across the test cases[16] are broadly supportive of the assumption that the current methodology is not strongly impacted by the effect of unobserved variables. The evidence provided by the test case which might suggest otherwise is limited, especially when considered alongside the other test cases with large cohorts that do not show a relationship between treatment and alias use (and without the capability to identify the direction of causality between alias use and treatment allocation).

7.2 Testing dummy interventions

As part of the methodology review, dummy intervention testing was considered in order to assess the validity of the overall JDL approach. This involves estimating the impact of dummy interventions by randomly selecting N people from the database to be 'treated', implementing the standard JDL methodology to construct a matched comparison group, estimating the impact of this hypothetical treatment and calculating the treatment effect confidence interval, then repeating this at least one hundred times. If the confidence intervals span zero in 95 per cent of the cases, this would suggest that the overall JDL approach is robust.

Investigations uncovered academic evidence which suggests that such a randomisation task would not be suitable for PSM, as it works against the non-random aspect of the PSM approach. King & Neilsen[17] state that "researchers should be aware that PSM can help the most in data where valid causal inferences are least likely (i.e. with high levels of imbalance) and may do the most damage in data that are well suited to making causal inferences (i.e.

[16] See section 3 for further information on the test cases used
[17] King,G., Nelson,R., (2016) 'Why Propensity Scores Should Not Be Used for Matching'

with low levels of imbalance)". The latter would apply to the data that would be used if offenders were randomly allocated to a dummy treatment.

Given that applying PSM after random assignment of a dummy treatment is unlikely to help establish the validity of the JDL approach, this approach was not investigated as part of the robustness tests. In addition, the resource required to run a JDL analysis over one hundred times would incur a largely disproportionate cost.

7.3 Testing sensitivity of the results to unobserved factors

While it was not a specific suggestion from the peer review, the JDL also investigated the plausibility of conducting a sensitivity analysis, in order to calculate how stable the estimated results are to the presence of an unobserved variable. Such tests do not provide evidence about whether unobserved variables exist; instead, they indicate how influential an unobserved variable needs to be, for example, to change a result from being statistically significant to non-significant.

Preliminary investigations showed that it would not be possible to run such analyses within current IT resources due to the size of JDL comparison groups. A sensitivity analysis was run on a condensed version of the dataset as part of the preliminary investigations. Bearing in mind the limitations of using a reduced dataset, the results showed that for the difference in reoffending rates between treatment and comparison groups to become not statistically significant, there would need to be bias present that is the equivalent of gamma=1.26. This corresponds to there being an unobserved variable (or a combination of unobserved variables) that, after controlling for all the observed variables in the PSM model, doubles both the odds of treatment and the odds of reducing reoffending. The JDL aims to investigate sensitivity analyses in the future.

8. Data retention practices

In the response report, it was noted that data retention would be useful to enable exploratory analyses to help further understand what works in reducing reoffending. Currently the status is the individual-level data that is provided for JDL analyses is deleted after publication, with several exceptions who have specifically asked for their data to be retained for future analyses that either increase the cohort size or to expand on the original JDL analyses into more bespoke investigations. The JDL team now explicitly ask the customer at the end of their analysis whether they would be happy for their data to be retained for such purposes and review this practice regularly.

In light of future developments in the world of data protection, in particular the General Data Protection Regulation coming into place from May 2018 and the Data Protection Bill as announced in the 2017 Queen's Speech, the JDL will monitor any changes and adapt accordingly whilst seeking to maximise the full potential of data available to the JDL.

9. JDL Conclusion

The investigation period following the methodology review has been a useful and explorative time for the JDL, in terms of gaining knowledge of other techniques and approaches that could be applied to the established JDL process to identify any potential enhancements.

This paper has summarised all investigations promised in the original response paper, and whilst some aspects have been highlighted as possible improvements, the balance between the approach that provides a suitable fit alongside available resource (in terms of manpower, IT capabilities and demand) and the need to keep reports clear to users emphasizes that the methodology as it currently stands handles this appropriately.

There have been a number of improvements to the JDL process since the initial pilot, including more recently improved consistency in the coding used within the team (enabling easier use of aggregate results for future projects), streamlining of such code to help improve turnaround time and free up time better suited to understand the nuances of each intervention, and using the expanding world of data science to facilitate better dissemination of JDL results to date.

An exciting work plan is underway within the JDL, with a number of developments and enhancements identified to drive the JDL offer forward and to be as useful as possible to a wide range of organisations. This plan includes developing employment and benefit outcomes (helping to answer the question as to whether those with gainful employment are less likely to reoffend than those without), developing bespoke reoffending periods (both shorter term stints for newer organisations who do not have data suitable to use the existing one year follow up period, and longer term reoffending metrics to help account for more serious offences that take more than a year to get through the criminal justice system).

The JDL team remain open to discussions with existing customers as well as anyone considering using the JDL in the future, and just as importantly those who do not wish to use the JDL, to understand what is most useful to our stakeholders as the JDL offer expands over time.

Contact

Press enquiries should be directed to the Ministry of Justice press office:

Tel: 020 3334 3536
Email: newsdesk@justice.gsi.gov.uk

Other enquiries about this report should be directed to:

Sarah French, Justice Data Lab statistician
Ministry of Justice, 7th Floor, 102 Petty France, London, SW1H 9AJ
Tel: 07967 592428
Email: justice.datalab@justice.gsi.gov.uk

Produced by the Ministry of Justice

Alternative formats are available on request from justice.datalab@justice.gsi.gov.uk