A review of the matching process for the impact analysis of the HMP Peterborough Social Impact Bond

Jake Anders and Richard Dorsett

Sunday, 11 June 2017

Executive Summary

In 2010, the world's first Social Impact Bond (SIB) was launched at Peterborough Prison. It was used to fund an intervention – 'The One Service' – aimed at reducing the reoffending among prisoners discharged after serving a sentence of less than 12 months. Under the terms of the SIB, investors are paid according to how successful the One Service is in reducing reconvictions. If a minimum threshold of a 7.5% reduction in reconviction events is reached across the pilot, payment is triggered. Additionally, there is an option to trigger an early payment if a 10% reduction is noted in the number of reconviction events in individual cohorts.

A propensity score matching (PSM) approach was used to estimate the impact. For cohort 1, the impact was estimated, by a previous team of independent assessors, to be a reduction in reconviction events of 8.4% (Jolliffe and Hedderman, 2014).

This report reviews the PSM approach, prompted in part by the desire to understand the reasons behind the differences in reconviction rates between prisoners discharged from HMP Peterborough and prisoners discharged from other prisons. Its primary aim is to identify whether there is a need to revise the approach taken before it is applied to cohort 2 (and the final cohort - the weighted mean of cohort 1 and cohort 2). To do this, various amendments to the methodology were explored.

It was not possible to replicate the results of Jolliffe and Hedderman (2014). This should perhaps be no surprise given the difficulties often encountered with replication attempts. However, the difference between the replication result and the Jolliffe and Hedderman (2014) result was not statistically significant.

Since this review did not identify any clear improvement in the matching process, the conclusion is that the cohort 1 approach be maintained. We also recommend an adjustment to the sample selection in cohort 2. It is important to note that this recommendation is based on the analysis of cohort 1 data and is not informed by cohort 2 reoffending data.

In more detail, the review began with a replication of Jolliffe and Hedderman (2014) and then explored whether various changes might give better estimates:

- changing the set of variables included in the propensity score model;

- changing how propensity score matching was implemented;
- allowing for unobserved prison- and/or area-specific differences in outcomes.

The analysis used data for cohort 1 plus a number of years prior to the introduction of the One Service. The results suggest:

- it may be possible to alter the set of variables in the propensity score model in order to achieve a somewhat better fit, but this did not lead to any overall improvement in the methodology;
- changing the implementation of matching did not achieve any overall improvements in the model;
- controlling for historic differences in outcomes between prisons is problematic since the mean number of reconviction events at HMP Peterborough has followed a different trend from other prisons over the 2006-2009 pre-treatment period.

In view of these findings, the recommendations for cohort 2 are as follows:

- **Use the same matching variables as for cohort 1**. While some advantages to altering the variable set were found, these were not sufficient to justify a change.
- **Use the same matching approach**. The cohort 1 approach of (up to) 10:1 matching without replacement, within a 0.05 caliper performed as well as alternative approaches.
- **Estimate effects only for men aged 21 or above**. This reflects the fact that, in cohort 1, there were almost no under-21s at HMP Peterborough. If under-21s are similarly absent at HMP Peterborough in cohort 2, the recommendation is to exclude them from comparator prisons too. This will help make the composition of individuals from other prisons more similar to that of HMP Peterborough.
- **Report standard errors of impact estimates**. This is suggested as a means of providing some sense of the statistical significance of the estimated impacts.

The protocol for cohort 2 and the final cohort is included as an appendix.

Authors

Jake Anders is Senior Research Associate in the Department of Learning and Leadership at UCL Institute of Education, University College London. He was formerly Research Fellow at the National Institute of Economic and Social Research (NIESR).

Richard Dorsett is Professor of Economic Evaluation at the University of Westminster, Visiting Professor at UCL Institute of Education and NIESR Fellow. He was formerly Principal Research Fellow and Director of Policy Evaluation at NIESR.

1 Introduction

In 2010, the world's first Social Impact Bond (SIB) was launched at Peterborough Prison. The Ministry of Justice (MoJ) and the Big Lottery Fund agreed to pay for the successful outcomes of the project if it reduced reoffending by 7.5% overall. There was also an opportunity to receive early payments if individual cohorts reduced reoffending by 10%. A SIB is a form of 'payment by results' (PbR) where funding is raised from private, non-government investors and used to pay for interventions to improve social outcomes. If these interventions are effective, this could result in savings to Government and wider benefits to society. As part of a SIB the Government agrees to pay a proportion of these savings back to the investors as a return on their investment. If the outcomes do not improve, investors lose their investment.

The Peterborough pilot uses a SIB to fund interventions to reduce reoffending among male offenders released from HMP Peterborough having served short prison sentences (less than 12 months). It is coordinated by Social Finance, a not-for-profit financial intermediary, who obtained investment funding from private individuals, trusts and foundations to finance the pilot. This investment is used to fund an intervention called the 'One Service'. This is a voluntary scheme offering through the gate support to reduce reoffending, meaning that contact is made with prisoners before release and continued in the community. It is delivered by a mix of paid caseworkers and volunteers. It takes a pragmatic and client-led approach, in which the mix of activities for each offender is determined by caseworkers according to individual need.

The Peterborough SIB pilot was originally intended to operate until 2017, funding the delivery of the One Service to three cohorts of around 1,000 prisoners released from the prison. Support from the One Service was available to cohort members for a period of up to 12 months post-release, and engagement was on a voluntary basis. While the pilot operated on a payment by results basis under the SIB model for the first two cohorts of released prisoners, a third cohort received One Service support under a 'fee-for-service' arrangement, rather than under the original SIB-funded PbR model. This change to the model was due to the roll-out of 'Transforming Rehabilitation' reforms to probation, which introduced mandatory statutory supervision for short-sentenced offenders – the target group for the Peterborough pilot – and also included a PbR funding mechanism to incentivise providers to reduce reoffending. This meant that while the pilot was concluded early in order to avoid any

duplication in services to the same population, the alternative fee-for-service funding arrangement for the third cohort enabled the pilot to continue operating until the new Community Rehabilitation Company (CRC) providers implemented their approach to rehabilitation.

Under the terms of the SIB, the MoJ, supported by the Big Lottery Fund, will repay investors their capital and a return on their investment according to how successful the One Service is in reducing reconvictions. Specifically, payment requires a 7.5% reduction in the number of reconviction events in the 12 months following discharge across the whole pilot. This measurement is called the 'final cohort' – the weighted mean of cohort 1 and cohort 2. If reconviction events are reduced by 10% in either of the first two cohorts of prisoners, a payment will also be made.[1]

The impact of the intervention was estimated using a propensity score matching (PSM) approach. Cave et al. (2012) describe the development of the PSM approach used for cohort 1.[2] The impact for that cohort was estimated (by a previous team of independent assessors) to be a reduction in reoffending of 8.4% (Jolliffe and Hedderman, 2014).[3]

Following publication of the cohort 1 results, MoJ commissioned a review of the methodological approach. This was prompted in part by the desire to understand the reasons behind the differences in reconviction rates between prisoners discharged from HMP Peterborough and prisoners discharged from other prisons.[4]

This report presents the results of that review. As such it is one of three related reports:

- **Methodology Review** – (this report) assessing the cohort 1 approach and recommending an approach for cohort 2 and the final cohort
- **Cohort 2 Report** – presents impact estimates for cohort 2 and the final cohort (Anders and Dorsett, 2017)
- **Learning Exercise** – explores the sensitivity of the results to the recommendation in the Methodology Review to alter the sample definition (Dorsett, 2017).

This document begins by attempting to replicate the Jolliffe and Hedderman (2014) cohort 1 impact estimates. The aim of the review is then to:

[1] A reconviction is defined as an offence committed in the 12 months following release from prison, and resulting in conviction at court either in those 12 months or in a further 6 month period (allowing time for cases to progress through the courts). If an offender is reconvicted of multiple offences on one sentencing occasion, this counts as one reconviction event.

[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/217392/peterborough-social-impact-bond-assessment.pdf

[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/341684/peterborough-social-impact-bond-report.pdf

[4] The MoJ announced its intention to review the cohort 1 methodology in https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/341682/pbr-pilots-cohort-1-results.pdf (Annex B).

- examine the approach used for cohort 1 to assess its adequacy:
  - whether changing the set of variables included in the propensity score model can give better estimates
  - whether changes to how PSM was implemented might be appropriate
- consider the possibility of allowing for unobserved prison- and/or area-specific differences in outcomes
- make a recommendation on what changes, if any, should be made to the methodology.

The remainder of the document follows a structure that reflects these points. It begins with a description of the data in Chapter 2. Chapter 3 provides an account of the Jolliffe and Hedderman (2014) replication attempt. Chapter 4 assesses the adequacy of the cohort 1 approach. It is made up of two main sections: Section 4.1 considers variable selection and Section 4.2 considers the matching approach, as well as the possibility of including unobserved prison- and/or area-specific differences in outcomes using a matched difference-in-differences approach. Recommendations are given in Chapter 5, which also concludes. For reference, an overview of the theory behind matching is given in Appendix A.

2 Data

2.1 Offending data

The data supplied by MoJ for the analysis in this document were taken from the Police National Computer (PNC). They covered men released from prison who had been serving sentences of less than one year. The number of reconviction events in the 12 months following release for each individual was recorded and used as the outcome variable for which impacts are estimated. The data recorded details of the nature of the prisoner's offence, his offending history and some background characteristics.

Summaries of the offending history were also reported. This included individuals' Copas scores, which provided a measure of the rate at which they had built up convictions throughout their criminal career (Copas and Marshall, 1998). It also included whether an individual was a chronic offender, defined as being among those individuals who accounted for half of the total number of previous offences in a cohort.

Data were supplied for several cohorts of prisoners. In addition to cohort 1 (who left between September 2010 and June 2012), we also had data for men released in 2005, 2006, 2007, 2008 and 2009. This allowed pre-SIB tests to be carried out.

In addition to Peterborough, the analysis included men released from the following prisons:[1] Altcourse; Bedford; Belmarsh; Birmingham; Blakenhurst;[2] Bristol; Brixton; Bullingdon; Cardiff; Chelmsford; Dorchester; Durham; Elmley; Exeter; Forest Bank; Gloucester; High Down; Holme House; Hull; Leeds; Leicester; Lewes; Lincoln; Liverpool; Manchester; Norwich; Nottingham; Parc; Preston; Swansea; Thameside; Wandsworth; Winchester; Woodhill; and Wormwood Scrubs.

An initial inspection of the data revealed two possibly important points. First, there were very few individuals below the age of 21 in Peterborough prison, whereas they were considerably more common in other prisons (Table 2.1). It seems plausible that reoffending probabilities vary markedly across 18 to 21 year olds and, if so, matching a 21 year-old to an 18 year-old could introduce substantial

[1] An obvious exclusion from this list is HMP Doncaster, which was removed as it was also conducting a payment by results trial during the same period.

[2] Although included in this list, HMP Blakenhurst is removed from the replication exercise for consistency with Jolliffe and Hedderman (2014).

Table 2.1: Age distribution in Peterborough and other prisons

Age	Other Prisons	Peterborough	Overall
18	0.012	0.001	0.012
19	0.018	0.000	0.018
20	0.020	0.000	0.020
21	0.051	0.043	0.051
22	0.052	0.059	0.052
23	0.051	0.069	0.052
24	0.048	0.046	0.048
25-34	0.393	0.410	0.393
35-44	0.233	0.241	0.233
45-54	0.094	0.093	0.094
55-64	0.022	0.031	0.023
65+	0.004	0.006	0.005
Total	1.000	1.000	1.000
N	31362	936	32298

Notes: Reporting column proportions of the grouped ages of those released from HMP Peterborough and the comparator prisons.

unobserved differences that would bias impact estimates. In view of this, when comparing matching approaches, we excluded individuals under the age of 21.[3]

Secondly, inspection of the 2005 cohort highlights that Peterborough Prison was not opened until April, since there were no release dates before this point. To use this cohort in a comparable way would require also restricting the other pre-SIB cohorts to April onwards (or May, given the very small number of releases during April itself) in order for them to be comparable. Instead, we simply excluded the 2005 cohort.

2.2 Geographic data

Geographic variations are potentially important. The data provided information on where the offence was committed. This was used (or, where this is not available, the police station where they are first processed) to link in information on area characteristics that may influence the probability of reoffending.

Specifically, we linked in the Employment Deprivation Domain from the English Index of Multiple Deprivation (IMD) 2010 (for English areas) (McLennan et al., 2011) or from the Welsh IMD 2011 (for Welsh areas) (Statistics for Wales, 2011)[4] for the Lower Layer Super Output Area (LLSOA) in which the offence was committed.

[3]Note that our replication exercise in Chapter 3 did include those aged under 21 in order to stay as close as possible to the approach of (Jolliffe and Hedderman, 2014). There is further discussion of the difference excluding individuals under 21 makes at the end of that chapter.

[4]The time points for these estimates vary between England (during 2008) and Wales (during 2009/10), but this does not appear to be avoidable.

Furthermore, the geographical distance between the offence location and the prison may influence reoffending. One plausible reason for this is that individuals incarcerated further from their homes could be more likely to suffer from degraded social networks, making it harder for them to return to a normal life after release and, hence, increasing the probability of reoffending. To allow for this, we calculated this distance and included it as a possible variable in the propensity score estimation.

3 Replicating Jolliffe and Hedderman (2014)

In this Chapter, the results of attempting to replicate the published cohort 1 impact estimate are presented. The main purpose of doing this is to provide reassurance that we can implement the cohort 1 methodology in an identical way to Jolliffe and Hedderman (2014). Since later Chapters compare impact estimates resulting from changes in the methodological approach against the cohort 1 'baseline', it is important that this baseline is estimated using the original approach. If this is not the case, such comparisons will provide ambiguous results.

We start by reporting the pre-matching cohort 1 sample descriptive statistics in Table 3.1, based on the relevant sample within the data provided. These match the results reported by Jolliffe and Hedderman (2014) closely. We note that there are statistically significant differences between the HMP Peterborough sample and the sample from other prisons in a number of regards.

Next, we estimated the same propensity score model reported by the previous evaluators. The results are reported in Table 3.2. As would be anticipated by the close replication of the sample descriptives, we estimated very similar conditional associations between the sample characteristics and being a former inmate of HMP Peterborough.

The PSM approach taken was to match up to 10 comparison individuals with each treated individual, based on nearest-neighbour matching without replacement and with a caliper of 0.05 applied. We attempted to replicate this approach, but came across two potential sources of variation in what will be selected as our matched sample:

1. Jolliffe and Hedderman (2014) report being able to achieve a full 10 matches for each treated individual in all cases; in our replication, however, this was not found to be the case. Some individuals had a smaller number of matches within the caliper distance.[1] As such, our matched sample was a little smaller than that achieved in the previous evaluation report.

2. Even if the estimated propensity scores were identical, the matched sample selected could be affected by, for example, the sort order of the data if there were multiple individuals with identical estimated propensity scores (our analysis suggests this was the case). In that case, it is unlikely

[1] One treated individual had no matches within the caliper at all and so was dropped. This is in contrast to Jolliffe and Hedderman (2014), who reported the caliper to be non-binding.

Table 3.1: Cohort 1 (replication) average characteristics pre-matching

Characteristic	Other Prisons	Peterborough	Std. Diff.	p
Age at Release	32.7	33.5	0.08	0.01***
Previous Offences	38.4	32.2	-0.16	0.00***
Previous Conviction Occasions	17.7	14.8	-0.18	0.00***
Previous Custodial Sentences	4.9	4.1	-0.12	0.00***
Sentence Length	125.3	125.9	0.01	0.82
Time Served	51.1	50.5	-0.02	0.59
Previous T1 T2 Convictions	1.6	1.2	-0.17	0.00***
White - British (%)	77.1	64.0	-0.29	0.00***
White - Foreign (%)	9.0	21.5	0.35	0.00***
Black - British (%)	5.7	6.1	0.02	0.57
Black - Foreign (%)	1.8	1.7	-0.00	0.91
Asian/Middle Eastern - British (%)	3.6	3.6	-0.00	0.98
Asian/Middle Eastern - Foreign (%)	1.9	1.7	-0.02	0.62
Age of First Offence	18.3	20.4	0.26	0.00***
Copas Score * 100	-54.8	-66.4	-0.15	0.00***
Severe Offence (%)	1.1	0.6	-0.05	0.17
Chronic Offender (%)	17.8	13.2	-0.13	0.00***
Offence: Absconding (%)	1.7	2.0	0.02	0.51
Offence: Breach CO (%)	9.2	6.3	-0.11	0.00***
Offence: Breach SSO (%)	11.6	12.2	0.02	0.61
Offence: Criminal/Malicious Damage (%)	1.9	2.1	0.01	0.65
Offence: Domestic Burglary (%)	2.0	2.9	0.06	0.05**
Offence: Drink Driving (%)	2.8	4.1	0.07	0.02**
Offence: Drug Import/Export/Production (%)	0.8	0.3	-0.06	0.11
Offence: Drug Possession/Small-Scale Supply (%)	1.7	2.8	0.07	0.02
Offence: Fraud/Forgery (%)	2.6	3.2	0.04	0.22
Offence: Handling (%)	1.4	1.3	-0.01	0.81
Offence: Other (%)	3.3	2.9	-0.02	0.52
Offence: Other Burglary (%)	3.8	2.8	-0.06	0.12
Offence: Other Motoring Offences (%)	5.3	7.3	0.08	0.01***
Offence: Public Order (%)	4.5	2.8	-0.09	0.01***
Offence: Robbery (%)	0.2	0.0	-0.07	0.14
Offence: Serious Violence (%)	0.6	0.3	-0.04	0.28
Offence: Sexual (%)	1.2	0.5	-0.07	0.07*
Offence: Child Sexual (%)	0.6	0.9	0.03	0.40
Offence: Soliciting/Prostitution (%)	0.0	0.0	-0.03	0.51
Offence: Taking and Driving Away (%)	1.4	2.0	0.05	0.11
Offence: Theft (%)	18.4	18.5	0.00	0.93
Offence: Theft from a Vehicle (%)	1.4	1.8	0.04	0.24
Offence: Violence (%)	23.7	23.1	-0.01	0.67
N	30,068	936		

Notes: Reporting means (or percentages for binary variables) by whether prisoner was released from Peterborough or another eligible prison. Sample restricted to be as comparable as possible to that used by Jolliffe and Hedderman (2014). Std. diff = Difference between the characteristic in Peterborough and in other prisons in units of the standard deviation of the variable in the sample i.e. translated into a standardised difference which is comparable across variables. p = p-value from a test of the null hypothesis of no mean difference between Peterborough and other prisons with regard to the variable in question (i.e. the level of statistical significance of the observed difference). Stars also indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.2: Logistic regression model to estimate propensity score

	Matching	
Age at release	0.0344	(1.41)
Age at release squared /1000	-0.191	(-0.61)
White - British	-0.158	(-1.51)
White - Foreign	0.712	(5.64)***
Asian/Middle Eastern - Foreign	-0.251	(-0.92)
Previous offences	0.0108	(1.87)*
Previous offences squared /1000	-0.0302	(-1.34)
Previous conviction occasions	-0.0549	(-4.60)***
Previous conviction occasions squared /1000	0.221	(3.08)***
Previous custodial sentences	0.0234	(1.56)
Previous custodial sentences squared /1000	-0.385	(-1.61)
Age at first offence	0.0231	(1.04)
Age at first offence-squared /1000	-0.397	(-1.13)
Copas score	0.377	(3.70)***
Previous T1 T2 convictions	-0.0659	(-1.79)*
Previous T1 T2 convictions squared /1000	2.069	(0.71)
Severe offence	-0.297	(-0.70)
Chronic offender	-0.0802	(-0.45)
Offence: domestic burglary	0.374	(1.85)*
Offence: other burglary	-0.212	(-1.04)
Offence: drink driving	0.0725	(0.42)
Offence: Drug import/export/production	-0.810	(-1.39)
Offence: Drug possession/small-scale supply	0.555	(2.69)***
Offence: motoring	0.204	(1.54)
Offence: public order	-0.418	(-2.06)**
Offence: Sexual	-0.849	(-1.78)*
Offence: taking and driving away	0.486	(2.02)**
Offence: breach of community order	-0.316	(-2.27)**
Observations	31004	
Log likelihood	-4086.003	
Chi-squared	223.872	

Notes: Reporting raw coefficients from logistic regression model of whether individuals are incarcerated in HM Peterborough, rather than any other prison. Sample restricted to be as comparable as possible to that used by Jolliffe and Hedderman (2014). *t* statistics in parentheses. Stars indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

that the comparators chosen in our replication would be the same as those selected by Jolliffe and Hedderman (2014).

Nevertheless, we found that matching was highly successful in reducing observable differences in characteristics between the former inmates or Peterborough and other prisons. In our post-matching sample of 9,981 (which compares to a sample size of 10,296 for Jolliffe and Hedderman) we found only two statistically significant differences between the two groups at the conventional 95% significance level ("White - foreign" and "Offence: robbery"). With 40 characteristics considered, we would expect two to register as significant purely by chance (Table 3.3).

Finally, we estimated the impact of the intervention using this matched sample. The results, estimated using regression on the matched sample, are reported in Table 3.4. The row labelled 'Peterborough' gives the average number of reconviction events per person for those released from HMP Peterborough, the row labelled 'Other Prisons' gives the estimated counterfactual[2] based on the matched sample. The row labelled 'Difference' is the estimated impact. The results suggest the intervention reduced reconviction events per person from 1.493 to 1.455; this is a reduction of 0.039 reconviction events per person, with a 95% confidence interval that goes from -0.194 to 0.117.

Jolliffe and Hedderman (2014) estimated a slightly bigger reduction, from 1.55 to 1.42, which translates into a reduction of 8.4% for cohort 1. We estimated a reduction of 2.6% (with a confidence interval that goes from a reduction of 12.9% to an increase of 7.8%). As discussed above, there are possible reasons for differences between the original analysis and this replication, some of which are inherent to the matching process. Furthermore, it indicates that the estimated impact was not significantly different from that of Jolliffe and Hedderman (2014).

It is relevant to highlight that the difference between these two estimates is not statistically significant. This is apparent by considering their confidence intervals. As noted, the estimated reduction of 2.6% is consistent with a reduction of up to 12.9%. At the conventional 95% significance level then, it is comfortably consistent with the Jolliffe and Hedderman (2014) estimate of a reduction of 8.4%.

In fact, the results in Jolliffe and Hedderman (2014) are reported without any indication of their statistical significance. The results shown in Table 3.4 include 95% confidence intervals based on standard errors estimated using classical inference. This shows the impact not to be statistically significant as conventionally interpreted.

We discussed in Section 2 the argument for removing offenders under the age of 21 from the analysis, since there were almost no such individuals in Peterborough in cohort 1. We repeated this replication exercise after removing individuals under the age of 21. The estimated impact on average reconviction

[2]The counterfactual is what would have happened in the treatment group (i.e. for prisoners released from HM Peterborough) in the absence of the treatment occurring.

Table 3.3: Cohort 1 (replication) average characteristics post-matching

Characteristic	Other Prisons	Peterborough	Std. Diff.	p
Age at Release	33.4	33.4	0.00	0.91
Previous Offences	32.1	32.3	0.00	0.90
Previous Conviction Occasions	14.9	14.9	-0.00	0.95
Previous Custodial Sentences	4.0	4.1	0.01	0.69
Sentence Length	124.4	125.8	0.02	0.57
Time Served	50.4	50.5	0.00	0.95
Previous T1 T2 Convictions	1.2	1.2	-0.02	0.57
White - British (%)	66.1	64.1	-0.04	0.21
White - Foreign (%)	18.7	21.4	0.07	0.05**
Black - British (%)	6.2	6.1	-0.00	0.90
Black - Foreign (%)	2.0	1.7	-0.02	0.49
Asian/Middle Eastern - British (%)	3.9	3.6	-0.01	0.68
Asian/Middle Eastern - Foreign (%)	2.0	1.7	-0.02	0.53
Age of First Offence	20.1	20.4	0.04	0.30
Copas Score * 100	-65.8	-66.3	-0.01	0.83
Severe Offence (%)	1.1	0.6	-0.05	0.19
Chronic Offender (%)	13.0	13.3	0.01	0.85
Offence: Absconding (%)	1.7	2.0	0.02	0.50
Offence: Breach CO (%)	6.4	6.3	-0.00	0.92
Offence: Breach SSO (%)	11.7	12.2	0.01	0.67
Offence: Criminal/Malicious Damage (%)	1.9	2.1	0.02	0.63
Offence: Domestic Burglary (%)	2.5	2.9	0.03	0.43
Offence: Drink Driving (%)	3.8	4.1	0.01	0.69
Offence: Drug Import/Export/Production (%)	0.3	0.3	0.01	0.76
Offence: Drug Possession/Small-Scale Supply (%)	2.3	2.8	0.03	0.34
Offence: Fraud/Forgery (%)	3.0	3.2	0.01	0.69
Offence: Handling (%)	1.2	1.3	0.01	0.81
Offence: Other (%)	3.2	2.9	-0.02	0.61
Offence: Other Burglary (%)	2.5	2.8	0.02	0.65
Offence: Other Motoring Offences (%)	7.1	7.3	0.01	0.87
Offence: Public Order (%)	2.6	2.8	0.01	0.77
Offence: Robbery (%)	0.4	0.0	-0.09	0.05**
Offence: Serious Violence (%)	0.7	0.3	-0.05	0.20
Offence: Sexual (%)	0.7	0.5	-0.02	0.62
Offence: Child Sexual (%)	0.8	0.9	0.01	0.79
Offence: Soliciting/Prostitution (%)	0.1	0.0	-0.04	0.33
Offence: Taking and Driving Away (%)	1.8	1.9	0.01	0.79
Offence: Theft (%)	20.1	18.5	-0.04	0.24
Offence: Theft from a Vehicle (%)	1.3	1.8	0.04	0.18
Offence: Violence (%)	23.9	23.1	-0.02	0.57
N	9,046	935		

Notes: Reporting means (or percentages for binary variables) by whether prisoner was released from Peterborough or another eligible prison. Sample restricted to be as comparable as possible to that used by Jolliffe and Hedderman (2014) and restricted to matched sample based on 10:1 nearest neighbour matching without replacement. Std. diff = Difference between the characteristic in Peterborough and in other prisons in units of the standard deviation of the variable in the sample i.e. translated into a standardised difference which is comparable across variables. p = p-value from a test of the null hypothesis of no mean difference between Peterborough and other prisons with regard to the variable in question (i.e. the level of statistical significance of the observed difference). Stars also indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

event rates was little changed by removing these individuals (see Table 3.5); a reduction of 0.043 reconviction events per person (2.9% reduction; confidence interval from -13.1% to 7.4%).

Table 3.4: Estimated impact of Peterborough Prison pilot programme on average reconviction event rates

	Cohort 1		
	Estimate	95% Confidence Interval	
Peterborough	1.455	[1.307	1.602]
Other prisons	1.493	[1.445	1.542]
Difference	-0.039	[-0.194	0.117]
Observations	9,981		

Notes: Reporting coefficients from linear regression model of reconviction event rate. 95% confidence intervals based on classical standard errors reported in brackets. Sample restricted to be as comparable as possible to that used by Jolliffe and Hedderman (2014) and restricted to matched sample based on 10:1 nearest neighbour matching without replacement.

Table 3.5: Estimated impact of Peterborough Prison pilot programme on average reconviction event rates - Age 21 and over sample

	Cohort 1		
	Estimate	95% Confidence Interval	
Peterborough	1.458	[1.310	1.606]
Other prisons	1.500	[1.453	1.547]
Difference	-0.043	[-0.198	0.113]
Observations	9,962		

Notes: Reporting coefficients from linear regression model of reconviction event rate. 95% confidence intervals based on classical standard errors reported in brackets. Sample restricted to be as comparable as possible to that used by Jolliffe and Hedderman (2014), except for the additional restriction to individuals aged 21 and over, and restricted to matched sample based on 10:1 nearest neighbour matching without replacement.

4 Review of evaluation method

This chapter considers possible changes to the approach taken by Jolliffe and Hedderman (2014) in their evaluation of the first cohort. We consider changes to the variables included in the propensity score model in Section 4.1 and the approach taken to the matching itself in Section 4.2 in order to assess whether such changes would provide a substantial improvement in the overall method.

4.1 Variable selection for the propensity score model

The first step with propensity score matching is to estimate the propensity score. The key consideration is around which variables to include. To recap, the aim in this application is to control for all important differences between Peterborough prison and other prisons. In this section we revisit the variable selection approach taken by the previous evaluators and assess it using statistical learning techniques. These techniques focus on achieving high levels of out-of-sample predictive accuracy, which is important in this setting where we derive our propensity score model on one dataset before applying it to another.

4.1.1 Variable selection

Jolliffe and Hedderman (2014) took the following approach to variable selection. They first estimated a model including all available regressors. They then removed variables which, in this first model, had p-values of greater than 0.20. In addition, they removed dichotomous variables where one of the groups had fewer than 5 observations and state that they may also have removed such variables when one of the groups had between 5 and 9 observations.

Table 4.1 lists the possible variables. The column labelled ‘Jolliffe & Hedderman’ shows which variables were retained in their propensity score model. As will be explained below, we used an alternative approach to inform the choice of variables in the propensity score. The resulting set is identified in the column labelled ‘RF Peterborough’. The column labelled ‘RF Reoffend’ identifies variables that are correlated with reoffence. The final column (‘Preferred’) is made up of just those variables that are in both the previous two columns.

Our reason for suggesting an alternative approach is prompted by two concerns. First, our preferred approach protects against ‘overfitting’. This problem arises when model specifications reflect in part

the random variations of a particular sample. The approach we set out below avoids this and is therefore better able to identify a model that is more generally valid rather than being tailored to a specific sample. Second, Jolliffe and Hedderman appeared not to include interaction terms. These are potentially important and allow the propensity score specification to capture influences in a more flexible way. This flexibility arises by allowing for the possibility that there are some combinations of characteristics that are particularly important.

4.1.2 Random forests and variable importance

We used Random Forests (RFs) with data pooled across all prisons and pre-treatment periods[1] to identify variables that were important both for predicting whether an individual was in Peterborough prison (rather than a different prison) and for predicting whether an individual was reconvicted. We also used the approach to identify potentially important interactions. This is a far more flexible process than that previously applied.

RFs are a generalisation of decision trees, themselves a flexible method of classification. However, decisions trees suffer from the problem of overfitting, discussed above. RFs get around this by growing numerous decision trees on random sub-samples of the available dataset and then combining these into a single classification model that averages out the overfitting found within each tree.

Having fit an RF, it is possible to estimate the importance of a variable in the model by considering the average increase in classification error associated with randomly changing (permuting) that variable (Breiman and Cutler, 2004). Our approach made use of this importance measure to indicate which variables should be selected as part of the matching model. A variable's importance measure should be interpreted as the change in the classification error we would expect as a result of randomly changing the value of the variable in the RF.

The importance of interactions between pairs of variables can also be explored. This approach compares the additive importance of each pair (i.e. simply adding together each one's individual importance) with their joint importance. A large difference suggests that the interaction is also important or, in other words, that it is important to pay attention to that combination of variables. As this is a computationally intensive process, we applied this approach to all variables with individual importance to classification error of greater than 0.0001, a conservative threshold, in the sense that it errs on the side of including variables.

[1] As indicated in Chapter 2, 2005 data were not used in any part of the analysis.

Table 4.1: Variables selected by Jolliffe and Hedderman, important variables from Random Forest (RF) models, and preferred variables selected based on RF results

All possible variables	**Jolliffe & Hedderman**	**RF Peterborough**	**RF Reoffend**	**Preferred**
Age at release	✓	✓	✓	✓
Age at first offence	✓	✓	✓	✓
Number of previous offences	✓	✓	✓	✓
Number of previous conviction occasions	✓	✓	✓	✓
Number of previous custodial sentences	✓	✓	✓	✓
Number of previous T1/T2 convictions	✓	✓	✓	✓
Copas Score	✓	✓	✓	✓
T1/T2 serious current offence	✓	✓	✓	✓
Chronic offender	✓	✓	✓	✓
Length of sentence		✓	✓	✓
Time served		✓	✓	✓
Distance from offence to prison		✓	✓	✓
Employment domain from IMD		✓	✓	✓
White - British	✓		✓	
White - Foreign	✓		✓	
Black - British				
Black - Foreign				
Asian - British			✓	
Asian - Foreign	✓			
S E Asian - British				
S E Asian - Foreign	✓			
Unknown - British				
Unknown - Foreign				
Index offence: Absconding bail				
Index offence: Malicious damage			✓	
Index offence: Domestic burglary	✓			
Index offence: Other burglary	✓		✓	
Index offence: Drink driving	✓		✓	
Index offence: Drug imp./exp./prod.	✓			
Index offence: Drug possession	✓			
Index offence: Fraud/forgery			✓	
Index offence: Handling				
Index offence: Other				
Index offence: Other motoring offences	✓		✓	
Index offence: Public order	✓			
Index offence: Robbery				
Index offence: Sexual	✓			
Index offence: Sexual (child)			✓	
Index offence: Soliciting/prostitution				
Index offence: Taking and driving away	✓			
Index offence: Theft		✓	✓	✓
Index offence: Theft from a vehicle			✓	
Index offence: Violence			✓	
Index offence: Serious violence				
Index offence: Breach of SSO			✓	
Index offence: Breach of CO	✓	✓	✓	✓
Interactions		✓	✓	✓

Notes: 'Jolliffe & Hedderman' indicates variables used in the analysis by Jolliffe and Hedderman (2014). 'RF Peterborough' indicates variables with absolute importance ≥ 0.0001 in RF of being incarcerated at Peterborough prison. 'RF Reoffend' indicates variables with absolute importance ≥ 0.0001 in RF of being reconvicted. 'Preferred' indicates final variables selected through approach described in this document (excluding interactions).

4.1.3 Predicting incarceration at Peterborough Prison and reconviction events

Table 4.2 ranks variables in order of their 'permutation importance measure'; a measure of how strongly predictive of HMP Peterborough a variable was. In other words, it identifies those characteristics of the Peterborough inmate population that differed most from other prisons. Table 4.3 provides similar measures but this time predicting reconviction events rather than the prison in which they had been incarcerated. Since the aim was to control for characteristics predictive of offending that differ substantially between Peterborough and other prisons, it was the intersection of these two sets of variables that it was important to control for through matching. In both tables, among the most important variables were an offender's Copas score, number of previous convictions, previous offences, previous custodial sentences and age.

We selected the variables with greater than or equal to 0.0001 importance in both the RF of incarceration in Peterborough and the RF of reoffending. Interactions between these selected variables that had an estimated importance of greater than or equal to 0.001 in both RFs were also included (details of the importance of these interactions are available on request).[2]

Having identified our preferred variables, we followed Jolliffe and Hedderman (2014) and included quadratic terms for continuous variables. In addition, we accounted for the possibility of a step change associated with going from zero to one in many of these continuous variables. For example, we might well think there is a change in probability associated with having one previous conviction rather than having zero that is quite different from the change associated with, for example, having four previous convictions rather than three.

[2] This level of importance was chosen to ensure that not too many interactions were added, which could again result in overfitting.

Table 4.2: Variable importance - RF model predicting prison as HMP Peterborough

	Importance
Number of previous conviction events	0.023
Copas Score	0.021
Number of previous offences	0.019
Number of previous custodial sentences	0.012
Distance: offence location to prison (miles)	0.008
Age at first offence	0.006
Sentence length	0.005
Age at release	0.005
Time served	0.005
Number of previous T1 T2 conv. events	0.004
Employment deprivation of offence location	0.004
Chronic offender	0.003
Offence: Theft	0.0002
Offence: Breach of CO	0.0002
Severe offence	0.0001
Asian - British	0.0001
Black - Foreign	0.0001
Offence: vehicle theft	0.0001
Asian - Foreign	0.00005
Offence: motoring	0.00005
Offence: absconding	0.00004
Offence: taking and driving away	0.00004
Offence: Child sexual	-0.00004
Offence: Public order	-0.00004
Offence: drink driving	-0.00003
Offence: Breach of SSO	0.00003
Offence: violence	-0.00002
Offence: malicious damage	-0.00002
Offence: fraud	0.00002
Offence: sexual	0.00002
Unknown - Foreign	0.00001
Unknown - British	0.00001
Offence: domestic burglary	-0.00001
SE Asian - Foreign	0.00001
Black - British	0.00001
Offence: serious violence	-0.00001
Middle Eastern - Foreign	-0.00001
White - Foreign	0.00001
SE Asian - British	-0.00001
Offence: prostitution	-0.00000
Offence: robbery	0.00000
Offence: handling	0.00000
Offence: other burglary	-0.00000
Offence: import/export/production	0.00000
Middle Eastern - British	-0.00000
Offence: drug possession	-0.00000

Notes: Reporting permutation importance of all variables available for matching, ranked according to their absolute importance for classification error. Conducted using the RF-SRC package for R (Ishwaran and Kogalur, 2015). RF includes 200 trees. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009.

Table 4.3: Variable importance - RF model predicting reconviction events

	Importance
Copas score	0.062
Previous conviction occasions	0.032
Previous offences	0.028
Previous custodial sentences	0.020
Age at release	0.014
Offence: theft	0.009
Age at first offence	0.008
Sentence length	0.008
Chronic offender	0.007
Time Served	0.006
Previous T1 T2 conv. occasions	0.005
Offence: drink driving	0.002
Offence: motoring	0.002
Distance: offence location to prison (miles)	0.002
Employment deprivation of offence location	0.001
Offence: breach CO	0.001
Offence: other burglary	0.001
Offence: violence	0.0004
Offence: breach SSO	0.0004
Offence: fraud	0.0004
Offence: vehicle theft	0.0004
Black - British	0.0002
White - Foreign	0.0002
Offence: child sexual	0.0002
Asian - British	0.0002
SE Asian - Foreign	-0.0001
Offence: malicious damage	0.0001
Severe offence	0.0001
Offence: serious violence	0.0001
Offence: drug import/export/production	0.0001
Offence: drug possession	-0.0001
Black - Foreign	0.0001
Asian - Foreign	0.00004
Unknown - Foreign	0.00003
Offence: domestic burglary	0.00003
Offence: taking and driving away	-0.00002
Offence: sexual	0.00002
Middle Eastern - Foreign	-0.00002
Offence: public order	-0.00001
Offence: robbery	-0.00001
Unknown - British	0.00001
Offence: prostitution	0.00001
Offence: absconding	-0.00001
Offence: handling	-0.00000
SE Asian - British	0.00000
Middle Eastern - British	0.00000

Notes: Reporting permutation importance of all variables available for matching. Conducted using the RF-SRC package for R (Ishwaran and Kogalur, 2015). RF includes 200 trees. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009.

4.1.4 Model fit: a comparison with the Jolliffe and Hedderman specification

Propensity scores were estimated using a logistic regression. In view of this, it is appropriate to consider the model fit of the propensity score model using the variables identified through prior estimation of random forests and to compare that with the fit when using the variable set chosen by the previous evaluators.

The fit statistics for logistic regression models of incarceration in Peterborough, rather than another prison, are reported in the upper panel of Table 4.4 for each cohort. We assessed the models on the basis of the information criteria: AIC and BIC. With both of these measures, lower is better. The difference between the two is that the BIC penalises the number of variables to a greater degree than the AIC.

The AIC results suggest superior performance of the model using RF results to guide specification rather than the Jolliffe and Hedderman (2014) approach. The BIC results suggest the opposite, indicating that this improvement comes at the cost of an excessive increase in the number of variables.

Table 4.4: Model fit statistics of logistic regression models of incarceration at HM Peterborough and poisson regression models of number of reconviction events

	2006		2007		2008		2009	
	Old	New	Old	New	Old	New	Old	New
Peterborough Incarceration								
AIC	3831.3	3548.9	4674.5	4426.6	5648.9	5456.9	5149.8	5068.8
BIC	4075.2	4249.0	4918.6	5127.4	5897.2	6169.7	5397.0	5778.5
Observations	19277	19277	19414	19414	22244	22244	21464	21464
	2006		2007		2008		2009	
Reconviction Events								
AIC	62573.8	61964.5	66216.2	65412.6	77305.4	76342.7	74884.8	74176.0
BIC	62817.6	62664.6	66460.3	66113.4	77553.7	77055.6	75132.0	74885.7
Observations	19277	19277	19414	19414	22244	22244	21464	21464

Notes: Reporting AIC and BIC model fit statistics from logistic regression models of incarceration at HM Peterborough rather than another prison and poisson regression models of incarceration of number of reconviction events. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009.

With the number of reconviction events we fitted a poisson regression model and again assessed model performance using the AIC and BIC. The lower panel of Table 4.4 generally shows smaller (i.e. better) fit statistics for the model using the new set of variables compared to the model using the old set.

On balance, there was some evidence that it may be preferable to base model specification on the random forest results, especially for outcomes, but this was not wholly conclusive. In considering the substantive differences between the two sets of variables (reported in Tables B.1 and B.2 in Appendix

B), we note the importance of the geographic variables, both the deprivation measures and the distance from the prison of the offence location. While not shown in the table, we also note that several of the interaction terms were important.

However, a relevant point to consider is how much difference switching to this approach would make to the impact estimates. The impact estimates resulting from the two different sets of matching variables are reported in Table 4.5 for each of the pre-treatment years. At this point, and for all subsequent impact estimates, we revert to using linear regression models on the matched samples. This approach is common in matching analysis and provides simple differences in means for the treated group relative to the baseline: these are more easily interpretable than coefficients from Poisson regression and are comparable with the approach of Jolliffe and Hedderman (2014).

Although there were some differences, the standard errors indicate that these were not statistically significant. This is unsurprising given that these are essentially placebo tests that we would expect to show impacts close to zero. However, neither is it the case that these placebo tests were consistently closer to zero with the 'new' variable set than they were with the 'old' variable set. On this basis, there does not appear to be a strong case for changing to the 'new' set of matching variables.

Table 4.5: Impact estimates from matched datasets using different sets of matching variables

	2006		2007		2008		2009	
	Old	New	Old	New	Old	New	Old	New
Other Prisons	1.295***	1.245***	1.424***	1.451***	1.531***	1.521***	1.548***	1.495***
	(40.55)	(39.18)	(46.11)	(43.02)	(52.43)	(42.36)	(50.86)	(44.94)
Peterborough Diff.	0.0481	0.0940	-0.0248	-0.0956	0.0400	-0.0232	0.191	0.233*
	(0.46)	(0.90)	(-0.26)	(-1.05)	(0.38)	(-0.23)	(1.61)	(1.95)
Peterborough N	391	383	504	498	630	619	564	562
Others N	3,830	3,511	4,929	4,591	6,077	5,864	5,432	5,432

Notes: Reporting coefficients from linear regression model of reconviction events based on matched samples selected using propensity scores either estimated using the Jolliffe and Hedderman (2014) variable set, or the variable set selected through the approach described above. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. Stars indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.2 Matching approach

Jolliffe and Hedderman (2014) used 10:1 matching without replacement, applying a caliper of 0.05. That is, the comparison group for Peterborough was constructed by taking, for each individual at Peterborough, those 10 men from other prisons with the most similar propensity scores. Since matching was without replacement, each prisoner outside Peterborough could only be included once in the comparison group. In this Chapter, we report the results of experimenting with the matching approach. In particular, we consider whether we can achieve a comparison group that more closely resembles the Peterborough group by modifying the matching approach. Throughout this section the propensity score was estimated using the regressors selected through the approach outlined in Chapter 4.1.

The performance of different matching approaches can be assessed by comparing the descriptive statistics of the matched datasets they yielded and by examining the estimated impact in the pre-treatment periods, when we know the true effect to be zero.

The sections that follow each report the results of varying one aspect of the baseline (Jolliffe and Hedderman, 2014) approach:

- covariate balancing – rather than logistic regression – to estimate the propensity score;
- varying the caliper;
- post-matching regression adjustment;
- matching with replacement rather than without replacement;
- kernel matching;
- matching combined with difference-in-differences.

4.2.1 Covariate balancing propensity score

The covariate balancing propensity score (CBPS) approach (Imai and Ratkovic, 2014) is designed to achieve a good balance of covariates in the matched treatment group, not just a good prediction of being in HMP Peterborough. As such, it aims to automate the usual iterative approach of estimating the propensity score, examining covariate balance and then re-estimating the propensity with a changed specification intended to improve balance.

The performance of this approach relative to the baseline (Jolliffe and Hedderman, 2014) approach is summarised in Table 4.6. This follows the same general format as many others in this Chapter. It shows, for each of the pre-SIB years, the mean number of reconviction events among the comparison group identified through matching (labelled 'Other Prisons') and the estimated Peterborough effect (labelled 'Peterborough Diff.'). As noted previously, these impacts should ideally not be significantly different from zero. Statistical significance is indicated by asterisks.

In the bottom panel of Table 4.6, the key row is labelled 'Mean Abs. Diff.'. This is a standardised measure of how different the comparison group identified through matching was from the Peterborough group. It is an average and summarises the degree of difference across all those variables included in the propensity score. A lower value indicates greater similarity.

On the basis of Table 4.6, there is no strong reason to prefer the CBPS results over the baseline results (labelled PS, for 'propensity score'). The pre-SIB impact estimates follow a similar pattern in both cases and neither approach out-performs the other with regard to the similarity of the resulting comparison group. In view of this, we do not consider CBPS further.

Table 4.6: Impact estimates from matched datasets using logistic regression-estimated and CBPS-estimated propensity scores

	2006		2007		2008		2009	
	PS	CBPS	PS	CBPS	PS	CBPS	PS	CBPS
Other Prisons	1.295***	1.300***	1.424***	1.398***	1.531***	1.535***	1.548***	1.546***
	(40.55)	(37.83)	(46.11)	(47.46)	(52.43)	(45.73)	(50.86)	(48.44)
Peterborough Diff.	0.0481	0.0496	-0.0248	0.00473	0.0400	0.0390	0.191	0.188
	(0.46)	(0.47)	(-0.26)	(0.05)	(0.38)	(0.37)	(1.61)	(1.57)
Mean Abs. Diff	0.008	0.014	0.014	0.017	0.015	0.012	0.014	0.015
Peterborough N	391	389	504	502	630	627	564	563
Others N	3,830	3,791	4,929	4,844	6,077	6,001	5,432	5,397

Notes: Reporting coefficients from linear regression model of reconviction events based on matched samples selected using either a logistic regression estimated propensity score, or a CBPS-estimated propensity score. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. t-statistics reported in parentheses. 10:1 matching without replacement, using a 0.05 caliper. Stars denote statistical significance (based on classical inference) as follows: ***, **, * correspond to 99%, 95% and 90% significance.

4.2.2 Varying the caliper width

Setting a caliper width imposes a kind of quality control on the matches that may be selected. It is not necessarily clear at what level of tolerance the caliper should be set, however (Bryson et al., 2002, p.27).

As previously noted, the baseline results used a caliper of 0.05. This choice represented "an attempt to balance the closeness of match between those from HMP Peterborough to those from comparison prisons while also trying to increase the probability of being able to obtain ten matches." (Cave et al., 2012, pp.21-22). To some degree, this is rather arbitrary so it is worth considering how sensitive the results are to this choice.

Table 4.7 presents mixed evidence. With all caliper widths considered – and we varied these from 0.025 (half the baseline case) to 0.10 (twice the baseline case) – the pattern of pre-SIB impacts is broadly similar. No significant effect was found in 2006, 2007 or 2008 but a marginally significant positive effect was found in 2009 when using the narrowest caliper. It was not the case that a wider caliper was associated with consistently smaller or consistently larger pre-SIB impact estimates, nor was there a consistent finding for the narrower caliper.[3] With regard to covariate balance, the 'Mean Abs. Diff.' row does not provide evidence that the caliper of 0.05 performed consistently worse than the alternatives. In any event, the differences were not substantial. Overall, there was little to recommend changing from the caliper of 0.05.

[3]None of this within-year variation was statistically significant.

Table 4.7: Impact estimates from matched datasets using different caliper widths

	2006			2007			2008			2009		
	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10
Other Prisons	1.291***	1.295***	1.317***	1.429***	1.424***	1.449***	1.497***	1.531***	1.494***	1.534***	1.548***	1.565***
	(41.32)	(40.55)	(42.55)	(47.41)	(46.11)	(49.47)	(50.76)	(52.43)	(51.56)	(48.85)	(50.86)	(51.52)
Peterborough Diff.	0.0514	0.0481	0.0254	-0.0300	-0.0248	-0.0500	0.0741	0.0400	0.0771	0.199*	0.191	0.174
	(0.50)	(0.46)	(0.25)	(-0.32)	(-0.26)	(-0.53)	(0.71)	(0.38)	(0.74)	(1.67)	(1.61)	(1.46)
Mean Abs. Diff	0.015	0.008	0.010	0.017	0.014	0.012	0.023	0.015	0.013	0.013	0.014	0.015
Peterborough N	391	391	391	504	504	504	630	630	630	563	564	564
Others N	3,797	3,830	3,854	4,866	4,929	4,957	5,997	6,077	6,104	5,350	5,432	5,478

Notes: Reporting coefficients from linear regression model of reconviction events based on matched samples selected using caliper widths of 0.025, 0.05, or 0.10. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. 10:1 matching without replacement. Stars denote statistical significance (based on classical inference) as follows: ***, **, * correspond to 99%, 95% and 90% significance.

4.2.3 Regression adjustment

Table 4.8 demonstrates the effect of post-matching regression-adjustment. This approach essentially treats matching as a 'pre-processing' stage, intended to get the non-Peterborough sample resembling the Peterborough sample. By then regressing the outcome (reconviction events) on the matching variables – including the Peterborough indicator – a more precise impact estimate may be possible.

In fact, the results show regression-adjustment to have had little effect. Consequently, we do not recommend it be adopted as the preferred approach. One factor influencing this is that regression does not take into account that the propensity score is itself an estimate. This can affect the standard errors of the impact estimates, something we consider further below.

Table 4.8: Impact estimates from matched datasets using different regression adjustment approaches

	2006		2007		2008		2009	
	None	Adjust	None	Adjust	None	Adjust	None	Adjust
Other Prisons	1.295***	1.305***	1.424***	1.443***	1.531***	1.547***	1.548***	1.537***
	(30.44)	(35.20)	(35.59)	(42.87)	(37.25)	(45.46)	(34.32)	(39.68)
Peterborough Diff.	0.0481	0.0295	-0.0248	-0.0368	0.0400	-0.0169	0.191***	0.213***
	(0.80)	(0.56)	(-0.44)	(-0.77)	(0.69)	(-0.35)	(2.99)	(3.89)
N	4,221	4,221	5,433	5,433	6,707	6,707	5,996	5,996
R^2	0.00	0.25	0.00	0.30	0.00	0.33	0.00	0.28

Notes: Reporting coefficients from linear regression models of reconviction events on a matched sample (10:1 nearest neighbour matching without replacement) either including only the treatment dummy variable, or additionally including the matching variables. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. t-statistics reported in parentheses. 10:1 matching without replacement, using a 0.05 caliper. Stars denote statistical significance (based on classical inference) as follows: ***, **, * correspond to 99%, 95% and 90% significance.

4.2.4 Matching with replacement

Matching without replacement generally increases bias, since it means that individuals may not be matched to the treatment group members to whom they are most similar if they have already been matched to a different treatment group member. However, in a dataset where the number of individuals available to be matched is large relative to the number of individuals in the treatment group, it may be that this is not particularly binding and so makes little difference.

The results in Table 4.9 confirm that whether matching was with or without replacement tended not to make much difference. Neither the impact estimates nor the similarity of the comparison group was substantially affected. On these grounds, the only real reason to consider matching with replacement is that the ability to derive appropriate standard errors is strongest in this case. Specifically, it becomes possible with standard software to take account of the additional imprecision arising from the fact that the propensity score is itself estimated. However, in the current application, this has to be balanced against the value of adhering to the original evaluation design. We return to this in Chapter 5.

Table 4.9: Impact estimates from matched datasets using matching with and without replacement

	2006		2007		2008		2009	
	No Rep.	Rep.	No Rep.	Rep.	No Rep.	Rep.	No Rep.	Rep.
Other Prisons	1.295***	1.299***	1.424***	1.482***	1.531***	1.482***	1.548***	1.533***
	(40.55)	(38.00)	(46.11)	(41.52)	(52.43)	(44.08)	(50.86)	(44.06)
Peterborough Diff.	0.0481	0.0436	-0.0248	-0.0831	0.0400	0.0892	0.191	0.207*
	(0.46)	(0.42)	(-0.26)	(-0.87)	(0.38)	(0.85)	(1.61)	(1.72)
Mean Abs. Diff	0.008	0.013	0.014	0.020	0.015	0.020	0.014	0.011
Peterborough N	391	391	504	504	630	630	564	564
Others N	3,830	3,410	4,929	4,291	6,077	5,163	5,432	4,683

Notes: Reporting coefficients from linear regression models of reconviction events based on matched samples (10:1 nearest neighbour matching) selected either with or without replacement. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. t-statistics reported in parentheses. 10:1 matching, using a 0.05 caliper. Stars denote statistical significance (based on classical inference) as follows: ***, **, * correspond to 99%, 95% and 90% significance.

The possibility of implementing optimal matching (Rosenbaum, 1989) was also investigated, but found to be infeasible. In any case, matching with and without replacement provides bounds on the resulting matched sample; the small size of differences between these two cases in our analysis suggests optimal matching would not have made a large difference either.

4.2.5 Kernel matching

Kernel matching, see for instance Caliendo and Kopeinig (2008), builds a comparison group by creating, for each treated individual, a weighted average of all potential comparator individuals. Weights reflect similarity of propensity scores; comparators that are more similar receive a larger weight. The two most important choices for these estimators are the bandwidth (loosely, the degree to which weights decline as similarity reduces) and the kernel type. The choice of kernel type also affects the resulting weights. A common choice, used in the results that follows, is the Epanechnikov kernel. This has the property of giving weights of zero when the difference is sufficiently large. In this respect, it operates like a caliper.

Table 4.2.5 compares the performance of kernel matching estimators with results from nearest neighbour matching. Bandwidths of 0.03, 0.06 and 0.12 were tried. The comparison groups identified using kernel matching looked substantially less similar to the Peterborough group than the comparison group identified using the 10:1 caliper matching without replacement. This held for all years

and, predictably, the poorer performance was greater the larger was the bandwidth. This reduces the confidence we can have in the kernel estimates and so we do not recommend their use in this case.

Table 4.10: Impact estimates from matched datasets using kernel matching with different bandwidths

	2006				2007				2008				2009			
	NN	0.03	0.06	0.12	NN	0.03	0.06	0.12	NN	0.03	0.06	0.12	NN	0.03	0.06	0.12
Other Prisons	1.295***	1.448***	1.493***	1.501***	1.424***	1.558***	1.621***	1.633***	1.531***	1.569***	1.616***	1.651***	1.548***	1.585***	1.591***	1.627***
	(40.55)	(97.94)	(98.11)	(97.64)	(46.11)	(95.99)	(97.44)	(96.58)	(52.43)	(83.42)	(97.74)	(101.67)	(50.86)	(86.50)	(96.57)	(97.01)
Peterborough Diff.	0.0481	-0.105	-0.151	-0.159	-0.0248	-0.160*	-0.222**	-0.234***	0.0400	0.00284	-0.0444	-0.0793	0.191	0.151	0.145	0.123
	(0.46)	(-1.05)	(-1.51)	(-1.59)	(-0.26)	(-1.77)	(-2.46)	(-2.59)	(0.38)	(0.03)	(-0.44)	(-0.78)	(1.61)	(1.30)	(1.25)	(1.05)
Mean Abs. Diff	0.008	0.083	0.106	0.107	0.014	0.056	0.085	0.087	0.015	0.039	0.076	0.108	0.014	0.036	0.076	0.102
Peterborough N	391	391	391	391	504	504	504	504	630	630	630	630	564	565	565	566
Others N	3,830	17,988	17,989	17,989	4,929	17,766	17,766	17,766	6,077	20,242	20,243	20,243	5,432	19,709	19,710	19,710

Notes: Reporting coefficients from linear regression models of reconviction events based on matched samples selected using either a nearest neighbour approach (10:1 nearest neighbour matching without replacement) or a kernel matching approach. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. t-statistics reported in parentheses. Results labelled NN use 10:1 matching with replacement and a 0.05 caliper. Other results are Epanechnikov kernel matching estimates, with bandwidth as indicated. Stars denote statistical significance (based on classical inference) as follows: ***, **, * correspond to 99%, 95% and 90% significance.

4.2.6 Matched difference in differences

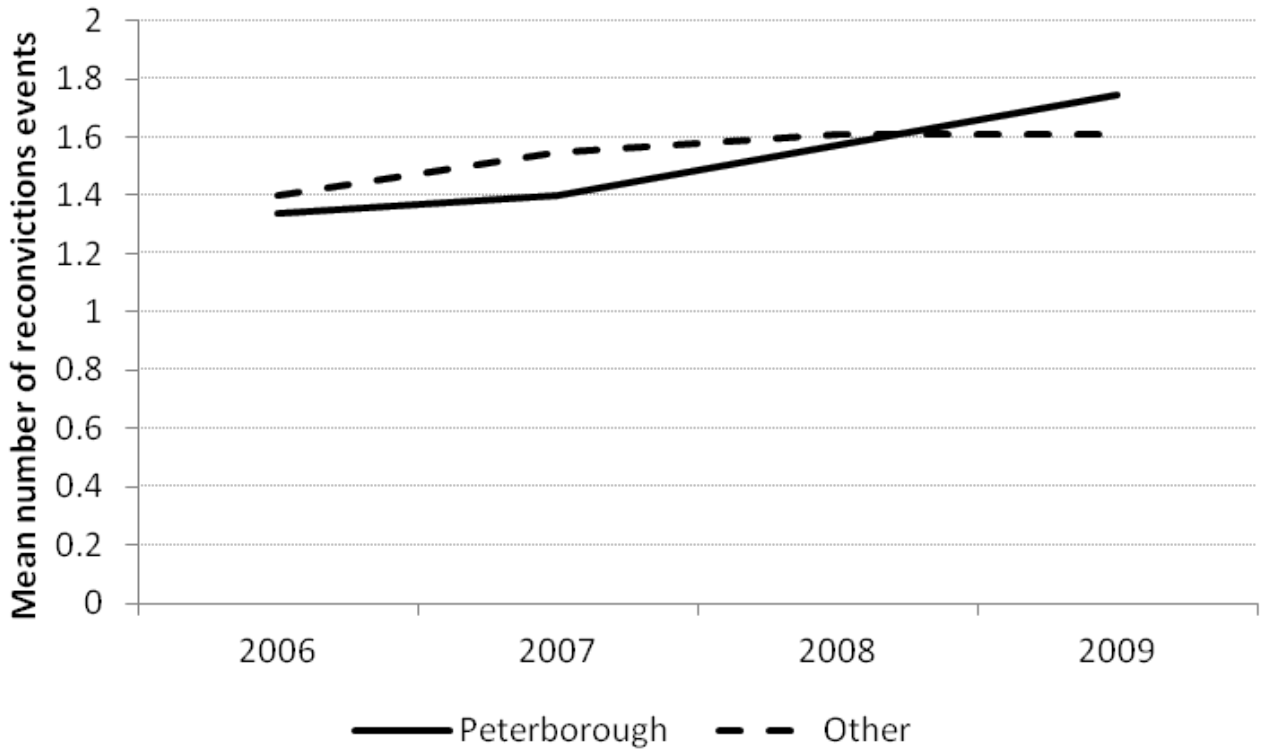
In the last section of this Chapter, we consider matched difference-in-differences (MDID) estimates. This can be an appropriate strategy when matching alone is unlikely to control for all differences between treatment and comparison groups but there are grounds for believing that the bias from not controlling for these remains stable over time.

A first impression as to whether there is likely to be such stability is possible by graphing changes over time in the mean number of reconviction events per prisoner in HMP Peterborough and in other prisons. This is shown in Figure 4.1. There is little to support the suggestion that pre-SIB trends in Peterborough and other prisons ran in parallel. Instead, the mean number of reconviction events among those released from HMP Peterborough went from being below that in other prisons to, in 2009, being higher. This provides an immediate indication that MDID is unlikely to be appropriate.

Table 4.11 confirms the impression from Figure 4.1 that the MDID estimator is not well-suited to the data at hand. The first row gives the result of estimating the impact in 2009, controlling for the change since 2008. Since 2009 is pre-SIB, we would hope the estimated impact to be zero. Instead, it was 0.14. The final two columns in Table 4.11 are the 95% level bootstrapped confidence intervals. These show the estimated impact was far from being statistically significant.

The second row gives the result of estimating the impact in 2009, controlling for the change over the years 2006-2008. This shows an impact of 0.18, which also was not statistically significant at the 95% level. In view of these results, we do not recommend combining matching with difference-in-differences.

Figure 4.1: Trends in mean number of reconviction events



Notes: Reports mean number of reconviction events by year of release and by whether individuals had been incarcerated in HM Peterborough or another prison. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009.

Table 4.11: Impact estimates using matched difference-in-differences

			Confidence interval	
Post	Pre	Impact	2.50%	97.50%
2009	2008	0.14	-0.17	0.39
2009	2006-2008	0.18	-0.06	0.46

Notes: Reporting estimated differences in means between the “post” year of 2009 and pre-years of either 2006-2008 or 2008 alone. Matched sample selected based on Epanechnikov kernel matching, with a bandwidth of 0.06. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. 95% confidence intervals around the point estimates of impact are reported in the columns 2.50% and 97.50%, based on 200 bootstrapped replications.

5 Conclusions and Recommendations

This document has presented a number of results. In this Chapter, we look across these results to draw out recommendations for cohort 2. We also make some more general points.

5.1 Replication

It was not possible to replicate the result of Jolliffe and Hedderman (2014). This should perhaps be no surprise given the difficulties often encountered with replication attempts.[1] However, the difference between the replication result and the Jolliffe and Hedderman (2014) result was not statistically significant.

5.2 Variable selection

The approach we used to select variables may be an improvement on that used by Jolliffe and Hedderman (2014). There are three aspects to this. First, we used an approach to variable selection that is highly flexible and so avoids the restrictions of the usual logistic regression. Second, we were able to include additional variables that were predictive both of being in Peterborough prison and of reoffending. Third, using pooled multiple pre-SIB years to estimate the random forest models meant the results did not relate to just one year but instead held more generally. By contrast, the approach used by Jolliffe and Hedderman (2014) is susceptible to variables exhibiting random variation around the statistical thresholds adopted for stepwise selection.

Despite these advantages, we recommend sticking with the Jolliffe and Hedderman (2014) set of matching variables for cohort 2. The reason for this is that, despite the new variable set identified in this report offering some advantages, these are not so great as to warrant deviating from the already-agreed approach. Most tellingly, the matching estimators obtained using the new variable set do not perform any better with regard to the pre-SIB placebo tests.

[1] Duvendack et al. (2015) looked across 162 replication studies in economics journals and found that two out of three were unable to confirm the original findings. Chang and Li (2015) in a smaller study were able to themselves replicate the main result in one third of cases where they had access to the original data and code. Assistance from the authors of the studies increased this to about one half. Note that we did not we have access to the code used by Jolliffe and Hedderman (2014).

5.3 Matching approach

The conclusion from our consideration of variants to the cohort 1 matching algorithm (rather than the propensity score estimation approach) is that there is no need for a major change to the methodology. We found no evidence that alternative approaches are likely to outperform the baseline 10:1 caliper matching approach without replacement. We therefore recommend sticking with that approach.

5.4 Standard errors

While the aim of the evaluation is to produce point estimates of the impact of the One Service, it is relevant to consider also the precision of these estimates. This is usually summarised as a standard error of the estimate. Obtaining correct standard errors is complicated under matching. In theory, account should be taken of the additional error arising from the fact that the propensity score itself is estimated. We are not aware of software that has incorporated the required adjustment in the case of many-to-one matching without replacement. However, even unadjusted standard errors are useful in conveying a sense of the uncertainty surrounding the estimated impacts. We recommend that these be reported for cohort 2.

5.5 Discussion

Having access to data from a number of pre-SIB years has been very helpful in assessing the performance of different approaches. It has also revealed the degree to which the estimated impacts jump about year-on-year. Such variation is less alarming when we take into account the fact that each impact is estimated with a confidence interval. In fact, the variation year-on-year is unlikely to be statistically significant. The point has already been made that standard errors or, equivalently, confidence intervals, can help convey a sense of how much the 'headline' result is subject to random variation. In fact, basing payments on the estimated impacts for cohort 1, cohort 2 and the final cohort increases robustness since it is likely to average out year-specific variations.

A final observation is that a comparison of the number of reconviction events over the period 2006-2009 among men leaving HMP Peterborough and men leaving other prisons reveals a marked difference. Men released from Peterborough in 2006 had fewer reconviction events on average than men released from other prisons but, by 2009, this situation had reversed. On the one hand, this highlights the challenge facing the evaluation, as described above. On the other hand, it does raise an issue around how we might view efforts to identify a meaningful counterfactual.

Specifically, if there is some unobserved factor responsible for the mean number of reconviction events increasing in Peterborough more than in other prisons, a failure to control for this would bias down-

wards the estimated impact. In other words, the true impact of the One Service would be larger than the estimated impact would suggest. However, this has to remain a point of speculation; while Figure 4.1 suggests a steady upward trend among men leaving HMP Peterborough, the short time-period over which this can be observed cautions against assuming this would have continued had the intervention not been introduced.

5.6 Recommendations

Based on the results in this report, our recommendations for cohort 2 are as follows:

- Use the same matching variables as Jolliffe and Hedderman (2014);
- Use the same matching approach: (up to) 10:1 matching without replacement, within a 0.05 caliper;
- Estimate effects only for men aged 21 or above if under-21s are not represented at HMP Peterborough in cohort 2;
- Report standard errors of impact estimates.

The protocol for cohort 2 and the final cohort is included as appendix C.

Bibliography

Anders, J. and Dorsett, R. (2017). HMP Peterborough Social Impact Bond - cohort 2 and final cohort impact estimates. MoJ Research Report.

Breiman, L. and Cutler, A. (2004). Random Forests. Manual.

Bryson, A., Dorsett, R., and Purdon, S. (2002). The use of propensity score matching in the evaluation of active labour market policies. DWP Working Paper 4, Department for Work and Pensions.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Cave, S., Williams, T., Jolliffe, D., and Hedderman, C. (2012). Peterborough Social Impact Bond: an independent assessment. Development of the PSM methodology. Ministry of Justice Research Series 8/12, Ministry of Justice.

Chang, A. C. and Li, P. (2015). Is economics research replicable? sixty published papers from thirteen journals say “usually not”. Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

Copas, J. and Marshall, P. (1998). The offender group reconviction scale: a statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47:159–171.

Dorsett, R. (2017). HMP Peterborough Social Impact Bond - learning exercise. MoJ Research Report.

Duvendack, M., Palmer-Jones, R. W., and Reed, W. R. (2015). Replications in economics: A progress report. *Scholarly Comments on Academic Economics*, 12(2):164–191.

Heckman, J. J. and Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American statistical Association*, 84(408):862–874.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Ishwaran, H. and Kogalur, U. B. (2015). Random Forests for Survival, Regression and Classification (RF-SRC). R package version 1.6.1.

Jolliffe, D. and Hedderman, C. (2014). Peterborough Social Impact Bond: Final Report on Cohort 1 Analysis. Report prepared for the Ministry of Justice, University of Leicester.

McLennan, D., Barnes, H., Noble, M., Davies, J., and Garratt, E. (2011). The English Indices of Deprivation 2010. Dclg technical report, Department of Communities and Local Government.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Statistics for Wales (2011). Welsh Index of Multiple Deprivation 2011. Technical report, Welsh Government.

A Appendix: An overview of matching estimators

Matching is a statistical approach for estimating the impact of a 'treatment'. To truly know the impact requires observing what outcomes would have been had the treatment not been introduced. Since we have no way of knowing this, it must be estimated. Matching operates by identifying a comparison group that is similar to the treatment group with regard to those characteristics thought to influence outcomes. Observed outcomes for this comparison group are then regarded as providing an estimate of the outcomes that would have been realised for the treatment group had there instead been no treatment. This is the 'counterfactual'.

Propensity score matching involves matching on the estimated probability of being in the treatment group. It draws on a result by Rosenbaum and Rubin (1983) showing how this can achieve covariate balance across treatment and comparison groups. Its advantage relative to matching on covariates directly is that it reduces the demands placed on the data by allowing a match on a single index rather than multiple individual characteristics, combinations of which may not be present in sufficient numbers in the available data.

As with matching more generally, propensity score matching relies on an assumption that all important differences between treatment and comparison groups have been captured. This is the Conditional Independence Assumption (CIA). If satisfied, it is valid to regard mean differences in outcomes between the treatment and matched comparison groups as estimates of the average causal impact of the treatment. The credibility of matching estimators depends critically on whether this assumption is satisfied, so careful consideration must be paid to this. The CIA is most credible when rich data are available and when the specification of the propensity score model reflects an understanding of the factors likely to influence outcomes.

A formal exposition of propensity score matching uses the notion of potential outcomes. An individual in the treatment group has, conceptually, an observed outcome associated with treatment but also an unrealised (and therefore unobserved) potential outcome associated with non-treatment. We write the potential treatment and non-treatment outcomes for individual i as Y_i^1 and Y_i^0, respectively. If we denote the characteristics of individual i by X_i and membership of the treatment group by a variable $T_i = \{0, 1\}$, the CIA can be expressed:

$$Y_i^0 \perp T_i | X_i \tag{A.1}$$

where $\perp$ symbolises independence. The relevance of this assumption can be seen from the definition of the average effect of treatment on the treated (ATT):

$$ATT = E\left\{E\left[Y_i^1 | X_i, T_i = 1\right] - E\left[Y_i^0 | X_i, T_i = 1\right]\right\}. \tag{A.2}$$

The second term on the right hand side is not observed but the CIA implies $E\left[Y_i^0 | X_i, T_i = 1\right] = E\left[Y_i^0 | X_i, T_i = 0\right]$ so the ATT can be estimated as

$$\widehat{ATT} = E\left[Y_i^1 | X_i, T_i = 1\right] - E\left\{E\left[Y_i^0 | X_i, T_i = 0\right]\right\}. \tag{A.3}$$

The Rosenbaum and Rubin (1983) result allows the CIA to be expressed in terms of the estimated propensity score, $Pr(\widehat{T_i = 1} | X_i)$:

$$Y_i^0 \perp T_i | Pr(\widehat{T_i = 1} | X_i) \tag{A.4}$$

so the ATT can be estimated as

$$\widehat{ATT}_{PSM} = E\left[Y_i^1 | Pr(\widehat{T_i = 1} | X_i), T_i = 1\right] - E\left\{E\left[Y_i^0 | Pr(\widehat{T_i = 1} | X_i), T_i = 0\right]\right\}. \tag{A.5}$$

Implementation typically involves estimating a propensity score and then finding, for each treated individual, one or more non-treated individuals who have similar values of the propensity score. Comparing the mean outcome of the treatment group members against this matched comparison group provides our estimate of impact. In practice, the estimate requires a number of decisions, particularly with regard to specification of the propensity score and the approach taken to ensure that the matches are sufficiently close. This is discussed further in Chapters 4.1 and 4.2. However, the main issue is usually whether the underlying assumption is met; have all important differences been taken into account?

The answer to this will often be negative. In this case, cross-sectional matching of the type outlined above may be insufficient. Some insight is possible in the situation where data are available prior to the treatment being introduced. This provides the opportunity to carry out pre-programme tests (Heckman and Hotz, 1989); that is, tests of non-existent treatments. Should significant impacts be detected, one might be concerned that the CIA is not satisfied.

If this is the case, it may be appropriate to instead use a matched difference-in-differences approach. This relies on an alternative assumption that may, in certain circumstances, be more likely to hold. Specifically, if the result of the pre-programme test is viewed as representing the bias of a cross-sectional matching estimator, a matched difference-in-differences approach can provide causal impact estimates if that bias is assumed to remain constant over time. This will be the case if the change over time in the mean non-treated potential outcome is the same for those in the treatment and comparison groups.

$$\begin{aligned} &E\left[Y^0_{i,t_1}|Pr(\widehat{T_i=1}|X_i),T_i=1\right]-E\left[Y^0_{i,t_0}|Pr(\widehat{T_i=1}|X_i),T_i=1\right] \;= \\ &E\left[Y^0_{i,t_1}|Pr(\widehat{T_i=1}|X_i),T_i=0\right]-E\left[Y^0_{i,t_0}|Pr(\widehat{T_i=1}|X_i),T_i=0\right]. \end{aligned} \tag{A.6}$$

By similar reasoning to that described above, the ATT can now be estimated as

$$\begin{aligned} \widehat{ATT}_{MDID} \;=\; & E\left[Y^1_{i,t_1}|Pr(\widehat{T_i=1}|X_i),T_i=1\right]-E\Big\{E\left[Y^1_{i,t_0}|Pr(\widehat{T_i=1}|X_i),T_i=0\right] \\ +\; & E\left[Y^0_{i,t_1}|Pr(\widehat{T_i=1}|X_i),T_i=1\right]-E\left[Y^0_{i,t_0}|Pr(\widehat{T_i=1}|X_i),T_i=0\right]\Big\}. \end{aligned} \tag{A.7}$$

Estimation involves matching treatment group individuals after the treatment has been introduced (the 'post' period) three times: to treatment group individuals in the pre period; to control group individuals in the post period and to control group individuals in the pre period. This provides estimates of the second, third and fourth terms on the right hand side of equation (A.7). Since the first term is observed, the ATT can be estimated.

B Appendix: Full regression models of incarceration at Peterborough and reconviction events

Table B.1: Logistic models of being incarcerated in HM Peterborough rather than another prison

	2006		2007		2008		2009	
	Old	New	Old	New	Old	New	Old	New
Copas score	-0.104	0.655	0.0771	-0.281	0.572	1.194	0.480	0.391
	(-0.22)	(0.49)	(0.19)	(-0.24)	(1.66)*	(1.29)	(1.36)	(0.40)
Copas score-squared	-0.0946	-0.0443	-0.00939	-0.341	0.113	-0.232	0.0135	-0.267
	(-0.68)	(-0.10)	(-0.08)	(-0.89)	(1.06)	(-0.76)	(0.12)	(-0.85)
Any Previous Offences		0.613		-12.25		-12.24		-12.02
		(0.56)		(-0.02)		(-0.03)		(-0.02)
Previous Offences	0.00260	0.0254	-0.00506	-0.00629	0.00222	-0.0264	0.00780	0.00373
	(0.33)	(0.89)	(-0.80)	(-0.24)	(0.35)	(-1.19)	(1.15)	(0.15)
Previous Offences-squared	-0.00000780	-0.000135	0.0000232	-0.0000941	0.0000102	-0.00000724	-0.00000914	-0.000242
	(-0.25)	(-1.25)	(1.06)	(-0.91)	(0.46)	(-0.09)	(-0.38)	(-1.75)*
Any Previous Convictions		-0.352		11.79		11.94		11.84
		(-0.30)		(0.02)		(0.03)		(0.02)
Previous Convictions	-0.0371	-0.146	-0.0299	0.00415	-0.0443	-0.102	-0.0755	-0.150
	(-1.37)	(-1.20)	(-1.50)	(0.04)	(-2.24)**	(-1.17)	(-3.69)***	(-1.56)
Previous Convictions-squared	0.000179	-0.00102	0.0000735	-0.00172	0.000199	-0.000281	0.000294	-0.000274
	(1.14)	(-1.03)	(1.18)	(-1.98)**	(2.21)**	(-0.75)	(3.35)***	(-0.57)
Any Previous Custodial		0.395		-0.0154		-0.0173		-0.0220
		(1.97)**		(-0.09)		(-0.12)		(-0.14)
Previous Custodial	0.0267	0.0579	0.0176	0.0265	-0.00471	0.140	0.0283	0.0549
	(1.08)	(0.49)	(0.89)	(0.28)	(-0.22)	(1.75)*	(1.22)	(0.63)
Previous Custodial-squared	-0.000207	-0.000537	0.0000164	0.000136	-0.000353	0.00329	-0.000480	-0.000998
	(-0.51)	(-0.23)	(0.08)	(0.08)	(-0.70)	(1.92)*	(-0.88)	(-0.43)
Any prev. T12 convictions		0.167		-0.0754		0.130		0.110
		(0.91)		(-0.44)		(0.88)		(0.66)
Previous T12 convictions	-0.0367	-0.143	-0.0384	0.0870	-0.0268	-0.0221	0.0269	0.225
	(-0.65)	(-0.82)	(-0.94)	(0.59)	(-0.51)	(-0.16)	(0.48)	(1.47)
Prev. T12 convictions-squared	0.000802	0.0000191	0.00144	0.000128	-0.00184	0.00617	-0.00528	-0.00868
	(0.16)	(0.00)	(0.52)	(0.02)	(-0.34)	(0.92)	(-0.91)	(-0.92)
Age	-0.00101	-0.0144	0.0857	0.0544	0.0420	0.0226	0.0168	0.0452
	(-0.03)	(-0.16)	(2.21)**	(0.67)	(1.28)	(0.35)	(0.51)	(0.65)
Age-squared	0.000134	-0.000326	-0.000956	-0.00258	-0.000286	-0.00188	0.0000557	-0.00122
	(0.29)	(-0.45)	(-1.97)**	(-3.39)***	(-0.72)	(-3.04)***	(0.14)	(-2.05)**
Age of first offence	0.0795	0.143	0.0456	0.153	0.0394	0.130	0.0555	0.0407
	(2.16)**	(1.49)	(1.30)	(1.85)*	(1.32)	(1.91)*	(1.77)*	(0.56)
Age of first offence-squared	-0.00106	-0.00104	-0.000794	0.000117	-0.000746	-0.000146	-0.00107	-0.000129
	(-1.86)*	(-1.31)	(-1.36)	(0.15)	(-1.57)	(-0.22)	(-2.15)**	(-0.20)
Sentence length		-0.00284		0.0000413		-0.00707		0.0102
		(-0.45)		(0.01)		(-1.59)		(2.12)**
Time served		0.0102		-0.00302		0.0169		-0.0194
		(0.76)		(-0.28)		(1.73)*		(-1.92)*
Chronic offender	0.353	2.389	0.378	1.492	0.0221	4.713	0.0883	1.245
	(1.38)	(1.31)	(1.67)*	(0.86)	(0.10)	(3.48)***	(0.39)	(0.77)
Severe offence	0.139	0.257	0.705	1.844	-0.00302	0.163	0.119	0.589
	(0.46)	(0.26)	(3.02)***	(1.99)**	(-0.01)	(0.20)	(0.45)	(0.56)
IMD Employment Domain		3.152		0.725		-3.636		-12.58
		(0.65)		(0.22)		(-1.50)		(-4.75)***
IMD Employment Domain-squared		-8.517		-2.664		3.102		5.029
		(-1.15)		(-0.75)		(1.66)*		(1.99)**
IMD Match Missing		0.178		0.416		0.625		0.661
		(0.86)		(2.81)***		(5.09)***		(4.73)***
Distance from offence to prison (miles)		0.0569		0.0597		0.0361		0.0179
		(6.00)***		(8.58)***		(6.71)***		(2.97)***
Distance off. to prison-squared		-0.000364		-0.000447		-0.000273		-0.000224
		(-5.40)***		(-8.85)***		(-7.77)***		(-5.09)***
Distance Match Missing		1.456		0.523		-0.133		-0.274
		(6.69)***		(2.86)***		(-0.77)		(-1.43)
White - British	-0.0315		0.197		0.0771		0.207	
	(-0.21)		(1.40)		(0.59)		(1.44)	
White - Foreign	0.397		0.510		1.010		1.152	
	(1.86)*		(2.64)***		(6.33)***		(6.86)***	
Asian - Foreign	0.402		-0.895		-0.706		0.240	
	(1.19)		(-1.50)		(-1.36)		(0.66)	
S E Asian - Foreign	-0.753		0.190		0.123		-0.0666	
	(-1.03)		(0.39)		(0.26)		(-0.11)	
Offence: Theft		0.508		0.476		-0.0531		1.036
		(1.10)		(1.13)		(-0.17)		(2.91)***
Offence: Drug I/E/P	0.336		-0.755		-0.805		-0.755	
	(0.46)		(-0.75)		(-0.80)		(-0.75)	
Offence: Drug Possession/Supply	-0.0167		0.0820		0.0961		0.0405	
	(-0.04)		(0.23)		(0.31)		(0.11)	
Offence: Drink Driving	-0.0392		0.380		0.576		-0.0820	
	(-0.17)		(1.97)**		(3.40)***		(-0.35)	
Offence: Other Motoring Offences	0.300		-0.00741		-0.0117		0.0414	
	(2.07)**		(-0.05)		(-0.07)		(0.25)	
Offence: Public Order	0.441		0.266		0.211		0.352	
	(1.71)*		(1.07)		(1.00)		(1.81)*	
Offence: Domestic Burglary	0.134		-0.456		0.00677		-1.182	
	(0.25)		(-1.07)		(0.01)		(-1.85)*	
Offence: Other Burglary	0.0556		-0.157		-0.00873		-0.340	
	(0.18)		(-0.54)		(-0.04)		(-1.18)	
Offence: Sexual	0.368		-1.397		-0.253		-0.945	
	(0.91)		(-1.94)*		(-0.59)		(-1.60)	
Offence: Taking and Driving Away	-0.473		-2.376		0.441		-0.761	
	(-0.89)		(-2.34)**		(1.36)		(-1.44)	
Offence: Breach of Com. Order	0.214	-1.088	0.366	0.113	0.496	-0.597	-0.0745	-0.796
	(1.11)	(-1.24)	(2.34)**	(0.17)	(3.65)***	(-1.04)	(-0.44)	(-1.15)
Constant	-4.845	-6.844	-5.567	-7.350	-4.422	-4.234	-4.149	-2.791
	(-7.21)***	(-4.93)***	(-8.68)***	(-6.23)***	(-8.28)***	(-4.39)***	(-7.44)***	(-2.72)***
AIC	3831.3	3548.9	4674.5	4426.6	5648.9	5456.9	5149.8	5068.8
BIC	4075.2	4249.0	4918.6	5127.4	5897.2	6169.7	5397.0	5778.5
Observations	19277	19277	19414	19414	22244	22244	21464	21464

Notes: Reporting coefficients from logistic regression model of being incarcerated in HM Peterborough, rather than any other prison. Coefficients from interaction terms not reported. Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. *t* statistics in parentheses. Stars indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Poisson regression models of number of reconviction events

	2006		2007		2008		2009	
	Old	New	Old	New	Old	New	Old	New
Copas score	0.532	0.228	0.694	0.207	0.626	0.245	0.585	0.0906
	(10.41)***	(1.65)*	(15.68)***	(1.63)	(14.44)***	(2.05)**	(14.34)***	(0.78)
Copas score-squared	-0.198	-0.00567	-0.178	0.0301	-0.159	0.0729	-0.201	-0.0964
	(-9.94)***	(-0.11)	(-9.50)***	(0.63)	(-9.07)***	(1.60)	(-11.84)***	(-2.31)**
Any Previous Offences		1.344		0.722		0.707		0.893
		(4.53)***		(1.70)*		(1.93)*		(2.86)***
Previous Offences	-0.000694	-0.0106	-0.00454	0.00158	-0.00264	-0.00207	-0.000882	-0.00272
	(-1.01)	(-4.55)***	(-6.85)***	(0.73)	(-4.02)***	(-0.99)	(-1.36)	(-1.39)
Previous Offences-squared	-0.000000986	-0.0000171	0.0000123	0.0000274	0.00000413	0.0000189	0.00000221	0.00000535
	(-0.47)	(-2.46)**	(6.27)***	(4.69)***	(2.07)**	(3.23)***	(1.22)	(1.00)
Any Previous Convictions		-0.838		-0.205		-0.157		-0.746
		(-2.81)***		(-0.48)		(-0.43)		(-2.36)**
Previous Convictions	0.0174	0.0507	0.0133	0.0257	0.0192	0.0432	0.0173	0.0246
	(6.81)***	(5.68)***	(7.72)***	(3.32)***	(9.12)***	(5.31)***	(8.97)***	(3.04)***
Previous Convictions-squared	-0.0000420	-0.000336	-0.0000256	-0.000146	-0.0000463	-0.000242	-0.0000436	-0.0000849
	(-3.74)***	(-7.42)***	(-6.24)***	(-5.13)***	(-5.54)***	(-6.58)***	(-6.23)***	(-2.78)***
Any Previous Custodial		0.0209		0.0307		-0.0121		0.0421
		(0.78)		(1.27)		(-0.54)		(1.87)*
Previous Custodial	0.0109	0.0245	0.0156	0.0306	0.0141	0.0255	0.0153	0.0495
	(5.42)***	(3.08)***	(7.95)***	(4.32)***	(7.24)***	(3.58)***	(8.31)***	(7.69)***
Previous Custodial-squared	-0.000113	0.000348	-0.000225	0.000291	-0.000186	0.0000312	-0.000189	0.000392
	(-4.56)***	(3.99)***	(-8.70)***	(4.25)***	(-6.06)***	(0.32)	(-7.65)***	(4.63)***
Any prev. T12 convictions		-0.0177		-0.0289		0.0411		0.0154
		(-0.93)		(-1.59)		(2.35)**		(0.86)
Previous T12 convictions	-0.0190	0.0379	-0.0168	0.00492	-0.0252	0.00968	-0.0276	-0.00680
	(-4.10)***	(2.58)***	(-3.71)***	(0.37)	(-5.01)***	(0.75)	(-5.40)***	(-0.53)
Prev. T12 convictions-squared	0.000463	0.000820	0.0000720	0.000208	0.000774	0.00123	0.000438	0.000452
	(1.42)	(1.46)	(0.22)	(0.43)	(1.87)*	(2.22)**	(1.05)	(0.81)
Age	-0.0449	-0.0538	-0.0314	-0.0588	-0.0355	-0.0619	-0.0556	-0.0709
	(-9.83)***	(-5.95)***	(-7.59)***	(-7.28)***	(-8.92)***	(-7.37)***	(-14.12)***	(-8.71)***
Age-squared	0.000520	0.000670	0.000409	0.000834	0.000401	0.000939	0.000652	0.000861
	(9.39)***	(7.65)***	(7.96)***	(10.19)***	(8.26)***	(11.94)***	(13.67)***	(11.39)***
Age of first offence	0.0352	0.0617	0.0271	0.0712	0.0211	0.0580	0.0345	0.0642
	(7.83)***	(6.21)***	(6.06)***	(7.88)***	(5.12)***	(6.39)***	(8.75)***	(7.40)***
Age of first offence-squared	-0.000351	-0.000617	-0.000347	-0.000683	-0.000135	-0.000484	-0.000318	-0.000412
	(-4.56)***	(-5.95)***	(-4.25)***	(-6.44)***	(-1.89)*	(-5.01)***	(-4.79)***	(-4.62)***
Sentence length		0.00148		0.00304		0.00373		-0.000622
		(2.02)**		(4.66)***		(6.00)***		(-1.17)
Time served		0.00227		-0.00513		-0.00479		0.00524
		(1.38)		(-3.73)***		(-3.56)***		(4.47)***
Chronic offender	-0.0296	0.221	0.0243	-0.146	0.00526	-0.206	-0.00189	0.106
	(-1.26)	(1.47)	(1.07)	(-1.02)	(0.25)	(-1.58)	(-0.09)	(0.84)
Severe offence	-0.108	0.342	0.0332	0.165	-0.0214	0.184	0.0192	0.240
	(-3.11)***	(2.77)***	(1.01)	(1.51)	(-0.67)	(1.82)*	(0.60)	(2.37)**
IMD Employment Domain		-0.0896		-0.402		-0.141		-0.122
		(-0.29)		(-1.40)		(-0.56)		(-0.50)
IMD Employment Domain-squared		-0.139		-0.182		-0.0938		-0.123
		(-0.73)		(-0.97)		(-0.57)		(-0.72)
IMD Match Missing		0.0311		0.00477		-0.0169		0.00208
		(1.55)		(0.26)		(-0.97)		(0.12)
Distance from offence to prison (miles)		0.0000659		-0.000290		-0.000735		-0.000961
		(0.08)		(-0.45)		(-1.17)		(-1.65)*
Distance off. to prison-squared		-0.0000136		-0.000000136		-0.00000495		0.00000793
		(-2.76)***		(-0.04)		(-1.55)		(2.50)**
Distance Match Missing		-0.0518		-0.0211		0.0412		0.00906
		(-1.70)*		(-0.75)		(1.58)		(0.33)
White - British	0.0478		-0.00836		-0.00981		0.00293	
	(2.46)**		(-0.45)		(-0.57)		(0.17)	
White - Foreign	0.239		0.278		0.317		0.372	
	(6.96)***		(9.11)***		(11.52)***		(14.19)***	
Asian - Foreign	-0.0452		0.106		-0.0111		0.0439	
	(-0.61)		(1.49)		(-0.17)		(0.67)	
S E Asian - Foreign	0.104		0.132		-0.165		-0.251	
	(0.85)		(1.16)		(-1.39)		(-1.94)*	
Offence: Theft		0.310		0.305		0.422		0.406
		(8.86)***		(9.54)***		(13.95)***		(14.05)***
Offence: Drug I/E/P	-0.773		-0.277		-0.779		-0.935	
	(-4.82)***		(-2.36)**		(-4.98)***		(-5.03)***	
Offence: Drug Possession/Supply	-0.323		-0.224		-0.260		-0.354	
	(-6.40)***		(-4.52)***		(-5.83)***		(-6.58)***	
Offence: Drink Driving	-0.686		-0.731		-0.725		-0.757	
	(-15.22)***		(-14.46)***		(-14.32)***		(-13.87)***	
Offence: Other Motoring Offences	-0.475		-0.480		-0.553		-0.575	
	(-22.31)***		(-19.90)***		(-21.04)***		(-19.85)***	
Offence: Public Order	-0.132		-0.0745		-0.0525		-0.166	
	(-3.72)***		(-2.21)**		(-1.82)*		(-5.60)***	
Offence: Domestic Burglary	0.0901		-0.0966		-0.0135		0.0493	
	(1.40)		(-1.68)*		(-0.24)		(0.93)	
Offence: Other Burglary	-0.0204		0.0291		-0.0596		-0.0388	
	(-0.71)		(1.08)		(-2.32)**		(-1.45)	
Offence: Sexual	-0.156		0.0555		0.0568		0.0386	
	(-2.21)**		(0.95)		(1.00)		(0.67)	
Offence: Taking and Driving Away	0.0277		-0.107		-0.149		-0.181	
	(0.65)		(-2.30)**		(-3.24)***		(-3.76)***	
Offence: Breach of Com. Order	-0.383	0.158	-0.331	0.0283	-0.328	0.0545	-0.357	-0.326
	(-14.88)***	(1.44)	(-13.22)***	(0.27)	(-14.37)***	(0.58)	(-15.28)***	(-3.56)***
Constant	0.823	-0.291	0.933	-0.0728	0.928	-0.0817	1.061	0.523
	(10.56)***	(-1.62)	(12.51)***	(-0.41)	(13.43)***	(-0.51)	(15.60)***	(3.40)***
AIC	62573.8	61964.5	66216.2	65412.6	77305.4	76342.7	74884.8	74176.0
BIC	62817.6	62664.6	66460.3	66113.4	77553.7	77055.6	75132.0	74885.7
Observations	19277	19277	19414	19414	22244	22244	21464	21464

Notes: Reporting coefficients from poisson regression model of number of reconviction events. Coefficients from interaction terms not reported.Sample: Male prisoners released from a sentence of less than 12 months during the period 2006-2009. *t* statistics in parentheses. Stars indicate statistical significance as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C Appendix: A protocol for cohort 2 and the final cohort

C.1 Protocol for cohort 2

The analysis for cohort 2 will follow Jolliffe and Hedderman (2014), using the same matching approach (up to 10:1 matching without replacement, within a 0.05 caliper) and the same matching variables. It will do this only for men aged 21, unless men younger than this are more in evidence at HMP Peterborough in cohort 2. Impact estimates will be reported with standard errors.

The dataset used to carry out the matching will not contain the outcome variable. This is so that the matched sample can be constructed without knowledge of how it may influence the estimated impact. The need for this approach is to allow for the possibility that following the Jolliffe and Hedderman (2014) approach results in a comparison sample that does not resemble the HMP Peterborough sample. In this case, we may wish to exercise discretion as to how to proceed.

To describe the quality of the match, standardised differences will be reported for each variable and compared against those in Table 3.3. For each variable, the standardised difference is calculated as the mean difference between Peterborough and other prisons, divided by the square root of the mean variance in Peterborough and other prisons. The mean standardised difference will also be reported.

The assessment of whether the match quality is acceptable will be based on the distribution of p-values resulting from individual tests of differences between HMP Peterborough and other prisons across all matching variables (similar to those in the final column of Table 3.3). When carrying out multiple tests, some significant differences may arise by chance; 100 independent tests of null effects, for instance, would be expected to show 5 as statistically significant at the 5% level. In line with this, we will view match quality as *potentially* unacceptable if more than 3 of the 40 tests are found to be significant at the 5% level.

If match quality is potentially a concern, we will carry out further investigation. A particular focus is on the independence of the variables since, if not independent, the expected number of false positives changes. However, if after exploring this the match does not appear to be acceptable, we will explore options for improving balance. This may include introducing interaction terms, using a different type of matching approach or using regression adjustment.

This approach allows for the evaluators to have some discretion. The important point to bear in mind is that any explorations will based on data that exclude the outcome variable. This protects against the possibility of selecting an approach on the basis of how it influences the estimated impact.

In addition to these, largely methodological points, there is a change to the sample definition in cohort 2. With cohort 1, prisoners leaving HMP Peterborough at any point during the cohort period were regarded as being in the treatment group. Those leaving a different prison during the cohort 1 period and not later leaving HMP Peterborough within the same cohort period made up the pool of potential comparators. A feature of this definition is that it entails a systematic difference between the treatment group and the (resulting) comparison group. Individuals who would have potentially been in the comparison group were instead included in the treatment group if they had a subsequent short sentence, this time at HMP Peterborough. One reason this matters is that such individuals tend to have a higher reconviction rate. Furthermore, since the treatment group is much smaller than the comparison group, its mean number of reconvictions – the outcome of interest – is more susceptible to being inflated by their inclusion.

In view of this, a different sample definition will apply for cohort 2. The treatment group will include all those whose first discharge in the cohort 2 period is from HMP Peterborough. The comparison group will include all those whose first discharge in the cohort 2 period is from a non-Peterborough prison.

C.2 Calculating payment for cohort 2 and the final cohort

As with cohort 1, early payment for cohort 2 is contingent on there being an estimated reduction in the number of reconviction events of at least 10%. If this is achieved, the total payment (excluding VAT) will be calculated by multiplying the agreed unit outcome payment by the estimated reduction in the number of reconviction events. This reduction will be calculated as the estimated reduction (expressed as a percentage) multiplied by the estimated counterfactual number of reconvictions, as represented by the total number of reconvictions in the matched comparison group.

More succinctly, total payment for cohort i, P_i, is calculated by the following formula:

$$P_i = A \cdot R_i \left(1 - \frac{Q_i}{R_i}\right) \text{ if } 1 - \frac{Q_i}{R_i} \geq 0.1, \text{ 0 otherwise.} \tag{C.1}$$

where A is the (indexed) payment per reconviction reduction, R_i is the number of reconvictions in the matched comparison group for cohort i and Q_i is the number of reconvictions in the cohort i treatment group.

Cohorts with an estimated impact smaller than 10% do not qualify for early payment. However, If a

minimum threshold of a 7.5% reduction in reconviction events is reached across the pilot, payment is triggered.

Formally, the payment for cohort i on the basis of the final cohort result, E_i, is calculated by the following formula:

$$E_i = A \cdot R_i \left(1 - \frac{Q_i}{R_i}\right) \text{ if } 1 - \frac{Q_1 + Q_2}{R_1 + R_2} \geq 0.075 \text{ and } P_i = 0, \text{ 0 otherwise.} \tag{C.2}$$

We note that the estimated impact over both cohorts can be calculated as a weighted mean of the cohort specific impacts:

$$\begin{aligned} 1 - \frac{Q_1 + Q_2}{R_1 + R_2} &= \frac{R_1 + R_2 - (Q_1 + Q_2)}{R_1 + R_2} && \text{(C.3)} \\ &= \frac{R_1 - Q1}{R_1 + R_2} + \frac{R_2 - Q_2}{R_1 + R_2} && \text{(C.4)} \\ &= \frac{R_1 - Q1}{R_1} \frac{R_1}{R_1 + R_2} + \frac{R_2 - Q_2}{R_2} \frac{R_2}{R_1 + R_2} && \text{(C.5)} \\ &= \Delta_1 w_1 + \Delta_2 w_2 && \text{(C.6)} \end{aligned}$$

where Δ_i is the impact in cohort i and w_i is a weight representing the proportion of the total counterfactual reconviction events across both cohorts accounted for by the cohort i.

In calculating this quantity, it is debatable whether the Δ_1 should be represented by the Jolliffe and Hedderman (2014) estimate or the replication estimate reported in chapter 3. An argument in favour of using the replication result is that this would ensure a cross-cohort consistency of methodology in the calculation of the final cohort result. However, since the Jolliffe and Hedderman (2014) cohort 1 result is already established and agreed, we will combine that result with the new cohort 2 result in order to get the final cohort result, using weights as described above.