Investigating the effect of CFO Activity Hubs on Reoffending Risk Post Release from Custody

Heléna Jarvis, November 2023

Technical report, supplement to report "The effect of CFO Activity Hubs on Reoffending Risk Post Release from Custody". This report describes in detail the framework and analytical methodology used to investigate the effect of CFO Activity Hubs intervention on reoffending rates post release from custody. Results show that over a 6-month follow up period after enrolling onto the CFO Activity Hubs programme, **participants who received intervention were less likely** to reoffend compared to a 'similar' group of participants that did not receive intervention.

1 INTRODUCTION

CFO Activity Hubs is a voluntary programme coordinated by HMPPS Creating Future Opportunities[1] (CFO) and delivered by a range of providers. The programme aims to assist individuals on a licence or community order to abstain from criminal behaviour by supporting them to successfully reintegrate and contribute to their communities. The programme is primarily delivered within the community (with one Hub located in HM Prison Holme House – not included in this study) and focuses on offering support to those who face multiple barriers, have complex needs, and would have difficulty accessing existing mainstream services. Many offenders have skill deficits that make it difficult for them to succeed in the community, the CFO Activity Hubs seek to address this by offering a holistic range of meaningful activities where participants can learn new skills, spend time with others who understand their rehabilitation journey, and receive support to develop the necessary personal skills to ultimately desist from offending.

Reoffending accounts for 80% of all recorded crime, costing the taxpayer approximately £18 billion[2] annually, with adult offenders who have previously received a custodial sentence accounting for a large portion of the estimated costs at £6 billion [1], [2]. The "social costs" of crime to the victims and the wider community such as victim services or victims of violent crime being treated by the NHS due to physical injury.

This report is concerned with measuring the extent to which programme intervention reduced observed reoffending rates. To do this we stratified the level of programme intervention into three ordinal categories: None, Partial and Complete. Notionally, an intervention level of 'none' reflects those individuals who enrolled on the programme but received no intervention, our baseline control group; an intervention level of 'partial' reflects those who undertook some activity but where not all available activity was utilised; and an intervention level of 'complete' reflects those who undertook a considerable amount of activity and have or are close to exhausting the activities available to them. We elaborate on these definitions in section 1.3.

1.1 DATA SOURCES

The source of the data used in this study is primarily the CFO Case Assessment and Tracking System (CATS+) application used to administer CFO programme participants, this dataset was supplemented with further information from HMPS's p-NOMIS system and NPS's nDelius system, used to administer offenders in custody and the community, respectively. The baseline data is taken at the time the participant enrols onto the CFO Activity Hubs programme.

1.2 APPROACH

This analysis assessed the impact of the CFO Activity Hubs programme on reoffending rates. However, as it would be unethical to randomly assign programme intervention, we must instead rely on the results of an observational study, which often suffer from an inherent incomparability between the those that received the intervention and those who did not. Thus, in order to make valid causal inferences from the observational data we must adequately address confounding. A confounding variable is a factor that influences both the treatment and the outcome, which results in obscuring the true relationship as the treatment effects are mixed in with the effects of a third factor. It is important to adjust for confounding as part of the statistical analysis as it can greatly distort the association between the treatment and outcome, even changing the

[1] Previously known as Co-Financing Organisation

[2] Expressed in 2017/18 prices.

apparent direction of an effect. There are multiple methods that can be used to correct for confounding, the method we choose to focus on for this analysis is propensity score matching.

To present the results of the study, we divided them into two sections. Part 1 seeks to establish the causal pathways to gain an understanding of how participating in the CFO Activity Hubs programme influences reoffending risk whilst considering other factors, such as participants age or gender. Part 2 estimates the true effects of the CFO Activity Hubs programme on reoffending risk by correcting for confounding.

The study is further broken down into two subcategories for parts 1 and 2. The first, the treatment is an intervention level of 'complete'; second, the treatment variable is an intervention level of 'partial'. The definitions of 'complete' and 'partial' activity are expanded on in the subsequent section.

1.3 Study Cohort

The study is restricted to participants who have enrolled on the CFO Activity Hubs programme (in the community Hubs) between June 2021 and December 2022 and have spent a period of time in custody (including those remanded in custody or recalled to custody) prior to enrolling. The dataset comprises of 3,143 individuals (all participants where ethnicity was not stated have been removed from the dataset). For the purpose of this study, participants are stratified into three ordinal categories: None, Partial and Complete.

The three categories are defined as follows:

- The intervention level of **'none'** consists of those who have enrolled onto the programme but received no intervention i.e., completed the two compulsory activities (Hub induction and Initial Action Plan) and no more, as they have enrolled onto the programme they have consented to be a part of the research and have completed the same assessment upon enrolment as the rest of the study cohort. This group serves as our comparison group and will be referred to as the control group or untreated throughout this report.

- The intervention level of **'complete'** comprises of individuals who undertook a considerable amount of activity and have or are close to exhausting the activities available to them, more specifically, they have visited a Hub on eight or more occasions and are considered to be "actively engaged". We define actively engaged as attending a Hub no less than once per month on average and never exceeding more than three months between each visit. The timespan over which 'complete' intervention was achieved ranges from 16 days to 8 months, on average 'complete' intervention was achieved in 95 days.

- The intervention level of **'partial'** captures those who undertook some activity but where not all available activity was utilised, more specifically, they have visited a Hub less than eight times in total or have visited a Hub on more than eight occasions but have spent a period of time disengaged i.e., not "actively engaged". It is expected that some of the participants within the 'partial' level will go on to achieve an intervention level of 'complete' in time.

A total of 1,186 participants (38%) have an intervention level of 'none', 1,649 participants (52%) have an intervention level of 'partial' and 308 participants (10%) have an intervention level of 'complete'.

1.4 Study Population Characteristics

The proportion of male to female participants is similar across all intervention levels, with approximately 92.2%, 93.8% and 93.2% of participants being male for the 'none', 'partial' and 'complete' groups, respectively (note that trans participants have been categorised according to their legal gender). Figure 1-1 shows that the proportion of participants that identified themselves as belonging to an ethnic minority (in this report ethnic minority is defined as all other ethnic groups combined, excluding white minorities, with participants with unknown ethnicity or refused having been excluded from the dataset) group were similar across the 'none' and 'partial' groups, being 24% and 26%, respectively, in comparison to the 'complete' group in which only 16% identified themselves as belonging to an ethnic minority group. Approximately 48% of the 'none' group have enrolled on the CFO3 programme prior to enrolling or whilst enrolled on the CFO Activity Hubs programme, compared with 54% and 59% of the 'partial' and 'complete' groups, respectively. This suggests that those who enrol onto the CFO3 programme are more likely to engage with the Activity Hubs programme.

Offenders who are older, receive longer prison sentences or are serving their first prison sentence have a lower reoffending risk *[3]*. Those in the 'complete' group (mean age 43), or 'partial' group (mean age 40) are on average older than their 'none' group (mean age 38) counterparts and each additional year of age is associated with a two percent reduction in the odds of reoffending *[4]*. Figure 1-2 shows the distribution of age groups amongst the three levels of intervention. Figure 1-3 shows that longer sentences are more common amongst the 'complete' group than the 'partial' or 'none' groups, with the mean number of days in custody (for most recent offence) being 1370 days, 1193 days, and 921 days, respectively. Figure 1-4 shows that all groups are more likely to have been in custody on two to four occasions with the 'complete' group having been in custody approximately 2.6 times, compared with 2.8 times for the 'partial' group and 3.3 times for the 'none' group.

In contrast, offenders who have no academic qualifications, have accommodation problems, or have a higher offending intensity, measured by Copas rate (a score based on the number of previous sanctions and time elapsed between current and first sanction), have an increased reoffending risk [3], [4]. Almost one in five (19%) of the 'none' group have no education or qualifications, compared with just over one in ten (12%) of the 'complete' group, shown in Figure 1-5. Figure 1-6 shows that the 'none' group are less likely to be in secure housing (29%) when compared with the 'partial' and 'complete' groups (32%), evidence has shown that offenders in stable accommodation are 50% less likely to reoffend [5]. Figure 1-7 shows that approximately 28% of the 'none' group is considered to have a "high" offending intensity, compared with 22% and 15% of the 'partial' and 'complete' groups, respectively.

Over three in four (77%) of the 'none' group are managed under multi-agency public protection arrangements (MAPPA), compared with approximately two in three (66%) of the 'partial' and less than three-fifths (57%) of the 'none' group. Recent research has shown that reoffending rates for individuals managed under MAPPA are significantly lower than the national average [6].

Approximately 14% of the 'none' group were recalled to prison for their most recent sentence prior to enrolment, compared to 10% of the 'partial' and 'complete' groups. Offenders who suffer from mental health problems, have limited support and experience drug and alcohol addiction are more likely to breach their licence conditions. This can be a result of failing to keep appointments, failing to reside in approved accommodation or committing additional offences. This suggests that factors such as mental health problems may be more prevalent amongst the 'none' group.

Offence type is an important predictor of reoffending risk [3]. Offenders serving a sentence for an acquisitive offence, defined as an offence where the offender derives material gain from the crime, are the highest risk category for reoffending. Whereas offenders serving a sentence for a sexual offence have the lowest rate of reoffending [3], [7]. From Figure 1-8 it can be seen that sexual offences are considerably more common amongst the 'complete' group with almost one in three having committed a sexual offence.

Research suggests that those who desist from offending are more likely to have better coping skills and a positive perception of their lives and future prospects [8]. Approximately 72% of the 'none' group answered 'yes' when asked if they hold a positive attitude toward themselves compared with 67% of the 'complete' group. The 'none' group are also less likely to agree that it takes them a long time to get over setbacks in their life (51%) when compared with the 'complete' group (55%).

As pre-treatment characteristics differ across the three groups, we expect to observe differences in their reoffending rates. As such we need to find a way of isolating to what extent of any such differences can be attributed to the programme intervention, rather than the differences that exist between the groups.

Figure 1-1 Distribution of ethnicity

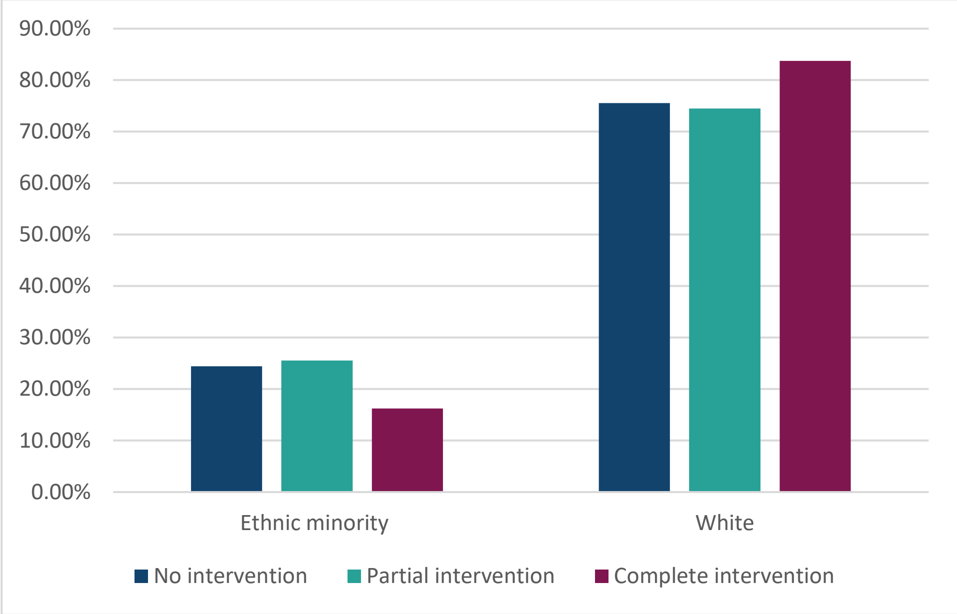


Figure 1-2 Distribution of age

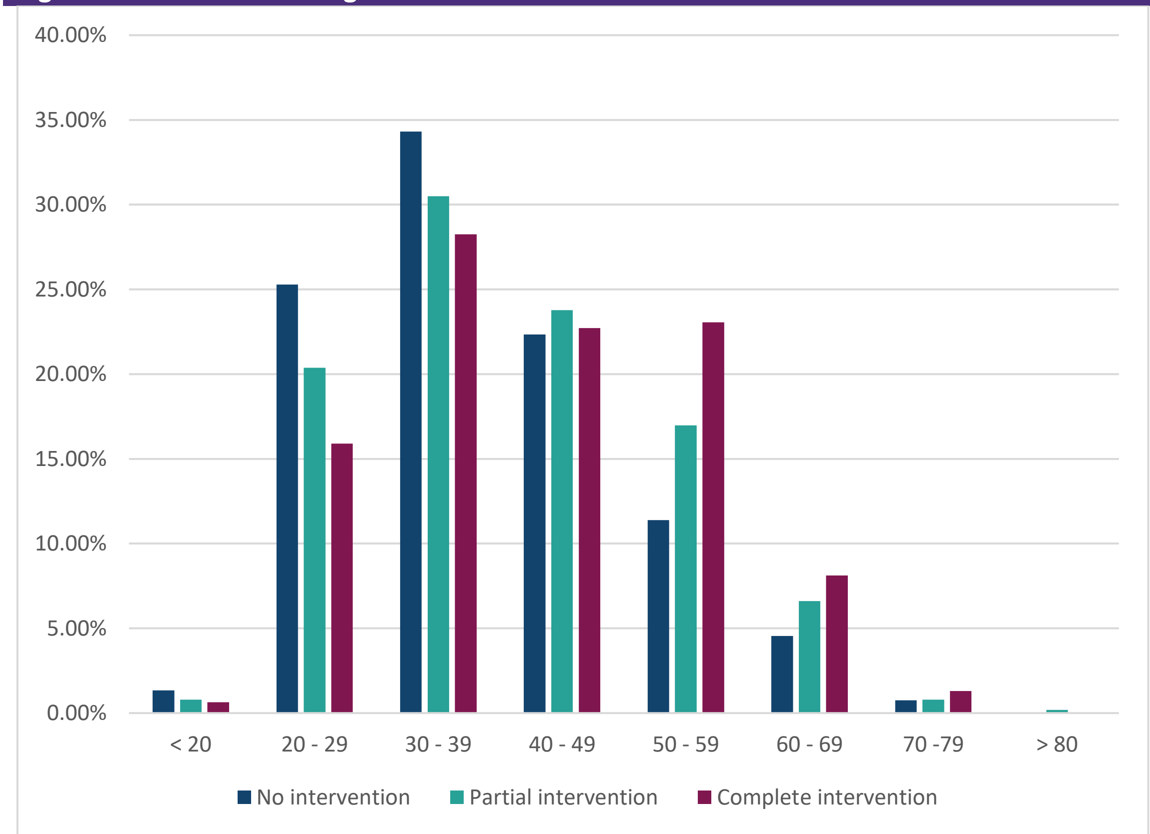


Figure 1-3 Sentence length (for most recent offence prior to enrolment)

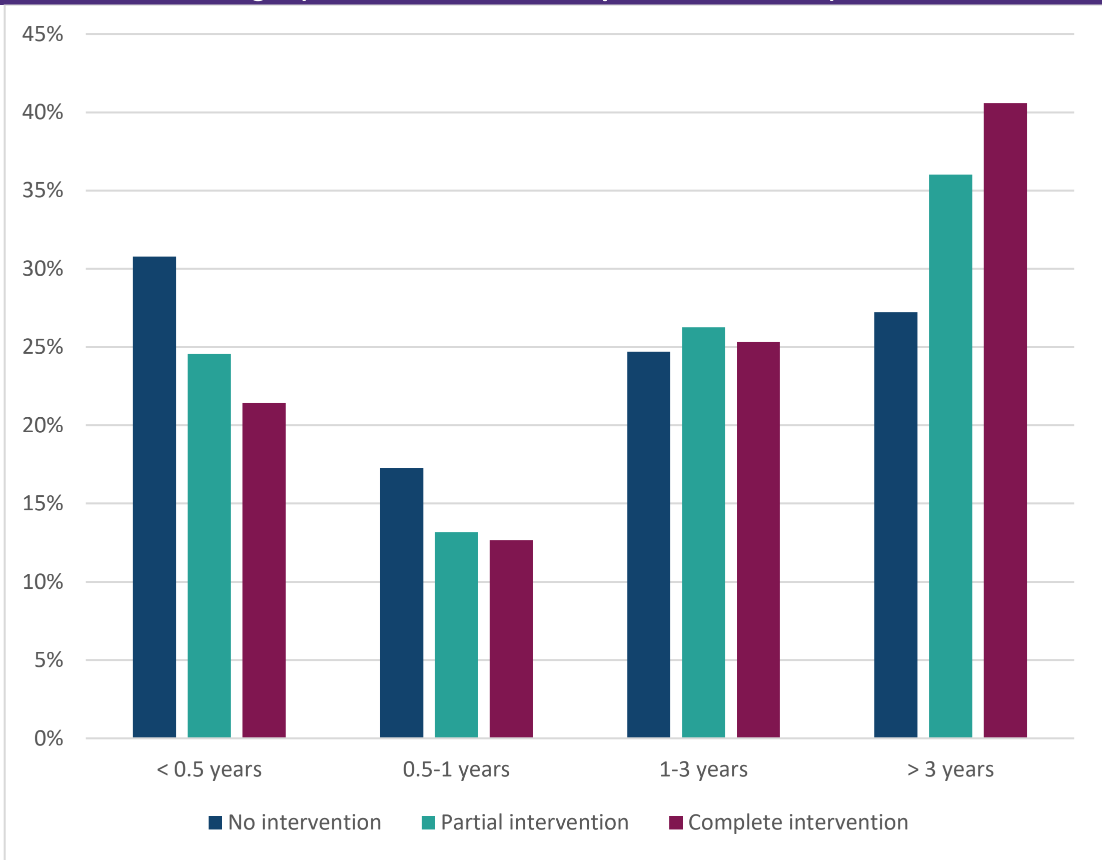


Figure 1-4 Number of times in custody prior to enrolment

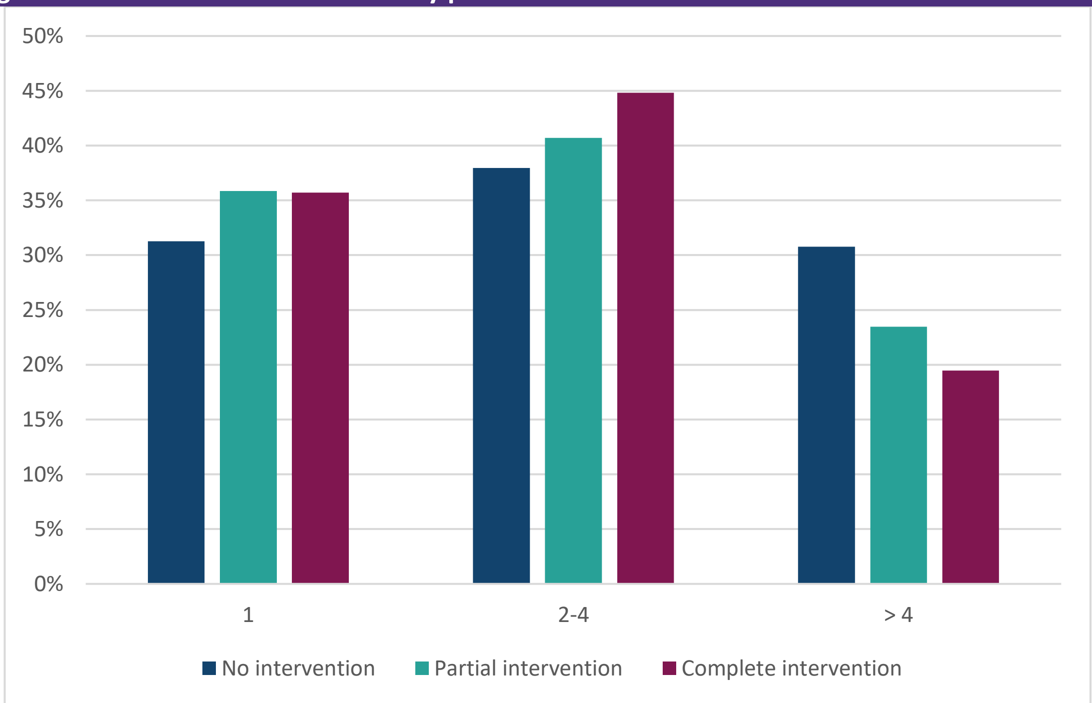


Figure 1-5 Highest level of education

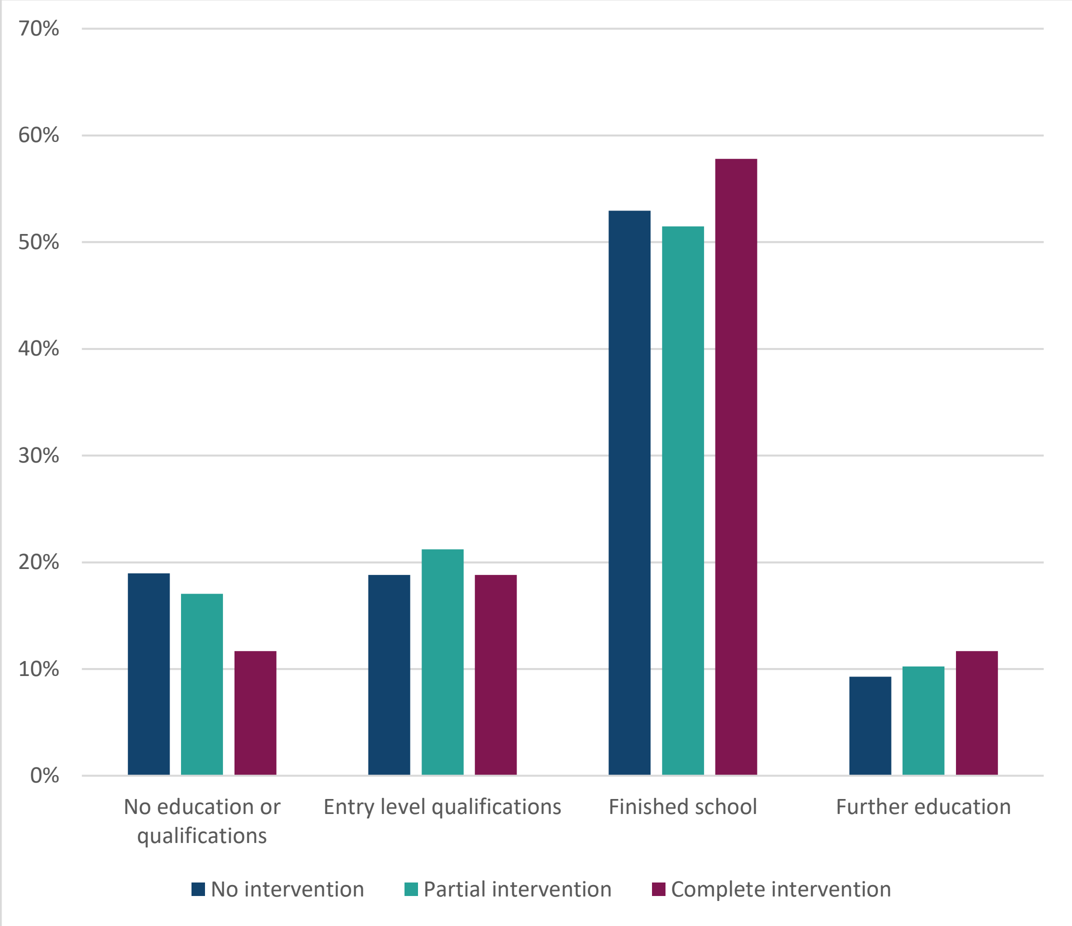


Figure 1-6 Housing status

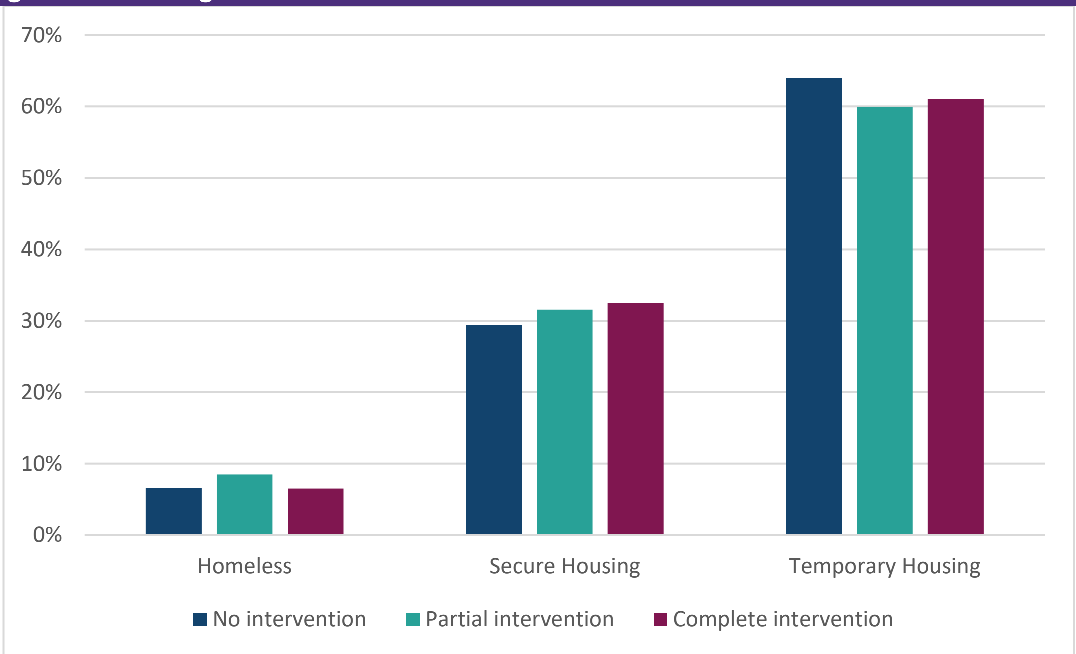


Figure 1-7 Offending intensity

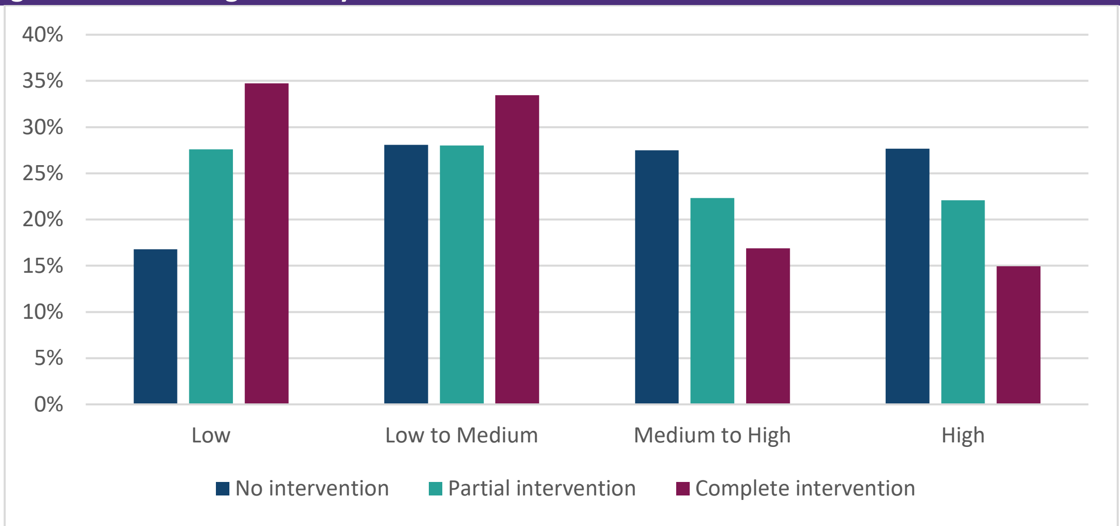
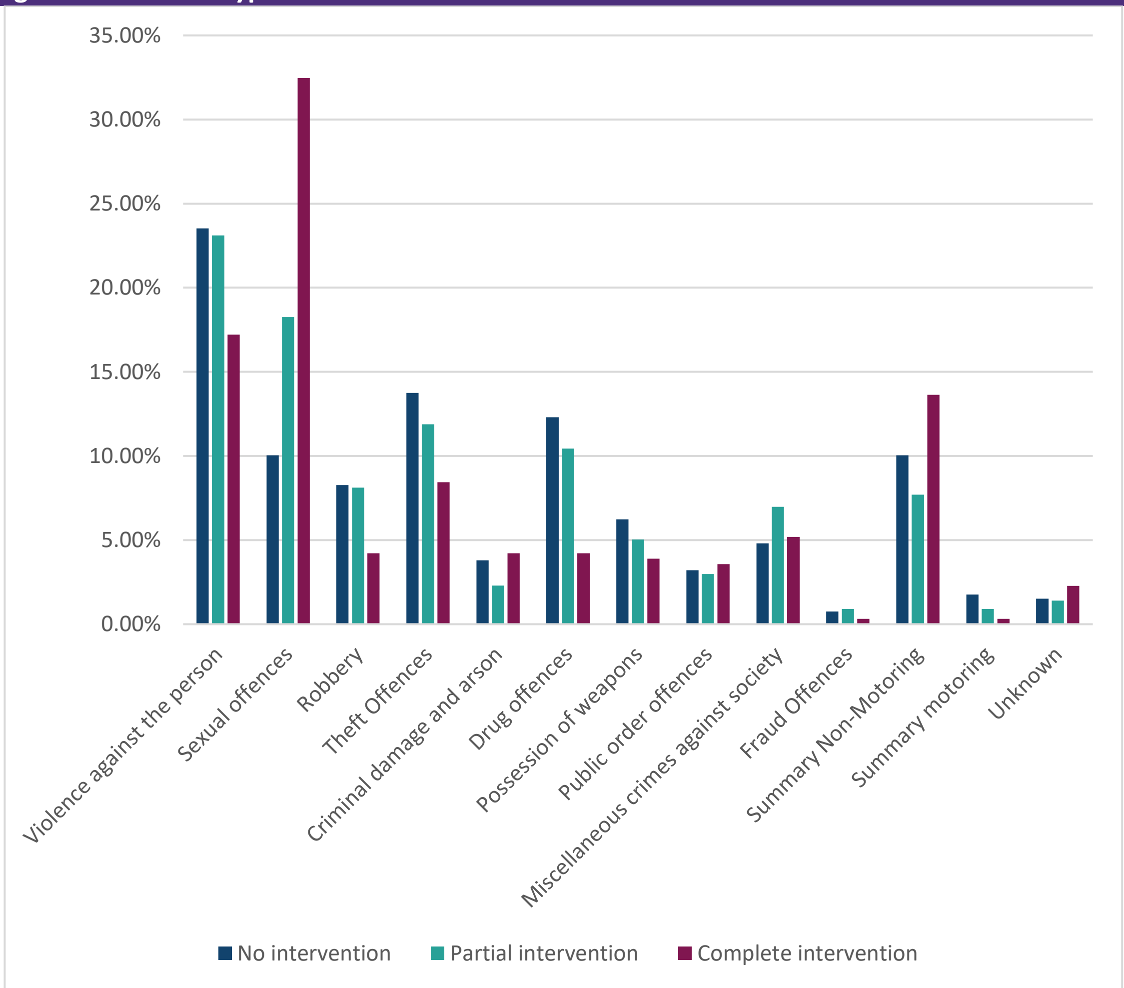
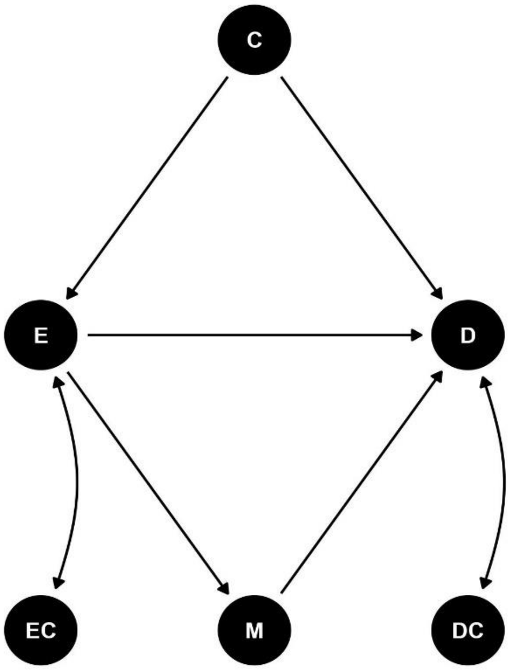


Figure 1-8 Offence type



2 PART 1: MAPPING THE CAUSAL PATHWAY

Figure 2-1 Directed acyclic graph showing the causal pathway



2.1 TREATMENT VARIABLE

A dichotomous (binary) variable, whose effect on a dependent variable is studied, represented by E in the causal diagram shown in Figure 2-1. In this analysis the treatment variable is an intervention level of ‘complete’, defined earlier in section 1.3 of this report. The analysis is then repeated following the same process with the treatment variable now as an intervention level of ‘partial’. In both cases the control group consists of those with an intervention level of ‘none’.

2.2 OUTCOME VARIABLE

A dichotomous (binary) variable which is expected to be influenced by the treatment variable, represented by D in the causal diagram shown in Figure 2-1. In this analysis the outcome of interest is a reoffence within 6 months of the index date, which in this case is the date the participant enrolled onto the CFO Activity Hubs programme. Specifically, a participant is considered to have reoffended if they received a new custodial or non-custodial sentence that resulted in a sentence length of greater than zero days, within a set time period; in the case of this study within 183 days of enrolling onto the CFO Activity Hubs programme.

2.3 CONFOUNDING AND MEDIATING VARIABLES

Confounders are defined as variables that obscure or accentuate the relationship between the treatment and the outcome, in contrast, mediators help to explain the relationship between the treatment and the outcome, represented in Figure 2-1 by C and M, respectively. Confounders are associated with both the treatment and outcome but are influenced by neither. Whereas mediators are part of the causal pathway from treatment to outcome. Hence, their distinction is an important one, as we do not wish to control for mediators since they are a part of how the treatment effects the outcome i.e., they lie on the causal pathway. However, there is no analytical way to make a distinction between a confounder and a mediator as they are identical statistically, the distinction therefore comes down to our conceptual view of the causal pathway [9]. In the context of this analysis, it was decided that it is not possible for any of the covariates to be mediators as all covariates were measured prior to the treatment and thus the treatment is unable to influence the covariates.

For example, age may confound the relationship between physical activity and heart disease incidence. Older individuals are likely to be less physically active than younger individuals, and older individuals have a greater risk of heart disease. If age is unequally distributed between the groups being compared, then this would overestimate the preventative effect physical activity has on heart disease. Since physical activity does not cause age, age does not lie on the causal pathway and therefore cannot be a mediator in this case. Income may have a mediating effect on the relationship between educational attainment and spending, higher educational attainment causes increased income, which in turn can lead to increased spending. In this example, income contextualises the effect educational attainment has on spending, unlike in the previous example where age obscures the relationship between physical activity and heart disease incidence.

2.4 Identifying confounding/mediating variables

In order to identify the potential confounding variables, we assess the magnitude of confounding by computing the 'change-in-estimate' (CIE). First, we compute the crude or unadjusted relative risk. We then use the Cochran–Mantel–Haenszel (CMH) formula, which measures the strength of association between the treatment and outcome after adjusting for any potential confounding variables. We compute this by stratifying the data into subgroups or strata of the potential confounding variable and create a series of two-by-two tables that show the association between the treatment and outcome. From this we then compute a weighted average of the relative risk across the strata [10]. We then compare the crude relative risk to the CMH estimate for the relative risk. Confounding is present if the effect estimates are considerably different, with a difference of more than 10% being the commonly cited cut off in the literature [11]. However, in this study we opt to be cautious and define the confounders as follows:

- CIE greater than 10%: Strong confounder
- CIE between 5% and 10%: Moderate confounder
- CIE between 1% and 5%: Mild confounder
- CIE less than 1%: Not a confounder

A limitation of this method is that it can only be used on the data that has been collected. This means that detecting confounders outside of the set of collected variables is not possible. This is unfortunately a fundamental problem of observational studies, and especially those that use matching as a control mechanism.

2.5 Moderating variables

A variable that changes the strength or direction of an effect between the treatment and outcome [12]. In other words, the treatment has a different effect among different subgroups. For example, it could be that the relationship between a medication and an illness is moderated by age. The medication could be more effective at treating younger patients as elderly patients are more likely to have comorbid medical conditions.

In order to identify the potential moderating variables, we use the Breslow-Day statistic, which tests if the odds ratios across the different strata are homogeneous. If the computed p-value is less than 0.05, then we assume the odds ratios are not homogeneous, in other words, the extent to which the odds ratios differ across the strata is the extent to which effect modification occurs. A limitation of this method is that in order for the Breslow-Day test to be valid, the sample size must be adequately large in each stratum [13].

2.6 Independent Covariates

Independent covariates can be associated with either the treatment or the outcome but not both, else it would be a confounder, represented in Figure 2-1 by EC and DC, respectively. In order to identify independent covariates, we perform two Pearson Chi-Square tests of independence, one to test if there is an association with the treatment and another to test if there is an association with the outcome. The Chi-Square statistic is a measure of the difference between the observed frequencies and frequencies we would expect if there were no association, with a significant difference implying there is an association. A limitation of the Chi-square test is that it cannot determine the direction of causality, it can only tell us if an association exists between the two variables [14].

2.7 Part 1 Causal Pathway Results – Intervention Level: Complete

Table 2-1 Confounding Factors – Complete intervention

Strong Confounders:	**CIE**	**Comments**
Offending intensity (Copas rate[3])	27.8%	The treatment group generally had a lower offending intensity, this is associated with lower reoffending risk
Number of offences	21.4%	Generally, the treatment group have committed fewer offences which is associated with lower reoffending risk
Length of time in custody	17.3%	The treatment group are more likely to have spent a longer period in custody, this is associated with lower reoffending risk
Number of times in custody	14.0%	The treatment group have typically been in custody fewer times, this is associated with lower reoffending risk

[3] The Copas Rate represents the rate at which an offender has built up convictions throughout their criminal career.

Provider	10.6%	The treatment group are more likely to attend a Hub in the North East (covered by provider Ingeus), this is associated with higher reoffending risk
Moderate Confounders:	**CIE**	**Comments**
Age group	9.4%	The treatment group is typically older, this is associated with lower reoffending risk
Offence	8.8%	The treatment group are more likely to have committed a sexual offence which is associated with lower reoffending risk
Does the participant have a stable relationship with family/friends	8.5%	The treatment group are less likely to have a stable relationship with family/friends, this is associated with higher reoffending risk
Who does the participant go to for support	6.1%	Generally, the treatment group is less likely to go to friends/family for support, this is associated with higher reoffending risk
How much support does the participant get from family	6.1%	The treatment group are less likely to have support from family, this is associated with higher reoffending risk
Mild Confounders:	**CIE**	**Comments**
Inmate status type	4.9%	The treatment group are less likely to be recalled, this is associated with lower reoffending risk
Does the participant say they are able to adapt to change	4.0%	Generally, the treatment group answer 'No' which is associated with higher reoffending risk
Does the participant hold a positive attitude toward themselves	3.2%	Generally, the treatment group answer 'No' which is associated with higher reoffending risk
Is the participant likely to engage with activities in the Hub	3.1%	The treatment group are more likely to be scored highly on this question, which is associated with lower reoffending risk
Highest level of education	3.1%	The treatment group are more likely to have higher levels of education, which is associated with lower reoffending risk
Participant has a disability	2.7%	The treatment group are less likely to have a learning disability or no disability, this is associated with higher reoffending risk
Ethnicity	2.4%	The treatment group are more likely to be white, this is associated with a higher reoffending risk

How did the participant get to the hub	2.3%	The treatment group are more likely to walk to the Hub, this is associated with a higher reoffending risk
How likely is the participant to return to criminal activity	2.3%	The treatment group generally have lower scores for this question, this is associated with lower reoffending risk
Years since first offence	2.0%	The treatment group are more likely to be within 9 years of their first offence, this is associated with lower reoffending risk
How likely is the participant to gain sustainable employment	2.0%	The treatment group are more likely to have a low score for this question, which is associated with higher reoffending risk
Participant is a carer	1.8%	The treatment group are less likely to be carers, this is associated with higher reoffending risk
Does the participant feel they have good personal qualities	1.6%	Generally, the treatment group answer 'No' which is associated with higher reoffending risk
Does the participant want support with employability	1.5%	Generally, the treatment group answer 'Yes' which is associated with higher reoffending risk

Table 2-2 Moderating Factors – Complete intervention

Moderators:	**p-value**	**Comments**
MAPPA registered	0.001	The preventative effect of 'complete' programme intervention on reoffending appears to have disappeared for those who are not MAPPA registered
Risk of Serious Harm (RoSH)	0.027	The preventative effect of 'complete' programme intervention on reoffending is stronger for those identified as RoSH
Gender	0.004	The preventative effect of 'complete' programme intervention on reoffending appears to have disappeared for female participants
Is the participant able to undertake daily living tasks	0.027	The preventative effect of 'complete' programme intervention on reoffending appears to have disappeared for those who scored 1 or 2 on the 1-5 scale, with the preventative effect being the strongest for those who scored 5

2.8 Part 1 Causal Pathway Results – Intervention Level: Partial

Table 2-3 Confounding Factors – Partial intervention

Strong Confounders:	**CIE**	**Comments**
Offending intensity (Copas rate[4])	16.7%	The treatment group generally had a lower offending intensity, this is associated with lower reoffending risk
Length of time in custody	13.4%	The treatment group are more likely to have spent a longer period in custody, this is associated with lower reoffending risk
Offence	10.5%	The treatment group are more likely to have committed a sexual offence which is associated with lower reoffending risk
Moderate Confounders:	**CIE**	**Comments**
Number of offences	9.9%	Generally, the treatment group have committed fewer offences which is associated with lower reoffending risk
Number of times in custody	9.1%	The treatment group have typically been in custody fewer times, this is associated with lower reoffending risk
MAPPA registered	5.7%	The treatment group are more likely to be MAPPA registered, this is associated with lower reoffending risk
Provider name	5.2%	The treatment group are more likely to attend a Hub in the North East (covered by provider Ingeus), this is associated with higher reoffending risk
Mild Confounders:	**CIE**	**Comments**
Age group	4.9%	The treatment group is typically older, this is associated with lower reoffending risk
Inmate status type	4.8%	The treatment group are more likely to IPP or Life/Other/Unknown which is associated with lower reoffending risk
Housing status	3.9%	The treatment group are more likely to be homeless which is associated with higher reoffending risk
Participant has a disability	3.9%	The treatment group are less likely to have Dyslexia, this is associated with lower reoffending risk

[4] The Copas Rate represents the rate at which an offender has built up convictions throughout their criminal career.

Does the participant have a stable relationship with family/friends	3.7%	The treatment group are less likely to have stable relationship with family/friends, this is associated with a higher reoffending risk
Is the participant able to undertake daily living tasks	3.3%	The treatment group are more likely to have a low score for this question, this is associated with a higher reoffending risk
How likely is the participant to gain sustainable employment	2.3%	The treatment group are more likely to have a low score for this question, which is associated with higher reoffending risk
How much support does the participant get from family	2.2%	The treatment group are more likely to have a low score for this question, which is associated with higher reoffending risk
Who does the participant go to for support	1.7%	Generally, the treatment group is less likely to go to friends/family for support, this is associated with higher reoffending risk
Does the participant say they are able to adapt to change	1.7%	Generally, the treatment group answer 'No' which is associated with higher reoffending risk
Time to enrolment	1.5%	The treatment group generally take longer to enrol onto the CFO Activity Hubs programme which is associated with lower reoffending risk
How did the participant get to the hub	1.4%	The treatment group are more likely to walk to the Hub, this is associated with higher reoffending risk
Is the participant satisfied with their life currently	1.1%	Generally, the treatment group answer 'No' which is associated with higher reoffending risk

Table 2-4 Moderating Factors – Partial intervention

Moderators:	**p-value**	**Comments**
Enrolled on the CFO3 programme[5]	0.006	The preventative effect of 'partial' programme intervention on reoffending appears to stronger for those who have also enrolled on the CFO3 programme
Participant has dependent children	0.029	The preventative effect of 'partial' programme intervention on reoffending appears to be stronger for those who have dependent children
Does the participant feel that they have good personal qualities	0.049	The preventative effect of 'partial' programme intervention on reoffending appears to be stronger for participants who feel they do not have good personal qualities

[5] Prior to, or whilst enrolled on the CFO Activity Hubs programme.

3 PART 2: ESTIMATING TREATMENT EFFECT

3.1 MEASURING EFFECT SIZES

We would like to know the difference between the participants outcomes with and without the treatment. However, a complication is that participating in the programme is not a random event. Naturally, those who more actively engage with the programme are systematically different from those who do not, in both observed and unobserved ways. This is known as covariate imbalance, and we must account for this in our study [15]. Outcomes will be influenced by the characteristics of the treatment group that are absent in the control group. For example, participants that attend a Hub more regularly are more likely to be older and the odds of recidivism decrease significantly with age [16]. Hence, we cannot conclude that the observed difference of outcomes is due to the treatment, as the effect could instead be attributed to the underlying differences that exist between the two groups prior to enrolling onto the programme. Therefore, we must adjust for confounding prior to estimating treatment effects, as otherwise this estimate will likely be biased.

One method often used to adjust for confounding is matching, which creates comparable treatment and comparison groups by ensuring that the distributions of observed covariates in both groups are similar, replicating what would have occurred had the treatment been randomly assigned. This removes confounding and allows us to confirm that any differences in the outcomes can be attributed to the treatment rather than any underlying differences between the two groups [17].

Propensity scores facilitate the construction of matched sets of treated and untreated cases by reducing the set of observed covariates, which represent many dimensions into a single one-dimensional measure known as the propensity score, introduced by Rosenbaum and Rubin (1983). The propensity score is defined to be the conditional probability of receiving the treatment of interest given the observed pre-treatment characteristics. We are then able to match cases based on this score to create comparable treatment and control groups, rather than requiring exact matches on all the different covariates. Although exact matching would be the most ideal scenario, exact matching can become infeasible when dealing with multiple covariates and finite numbers of potential matches as we are in this analysis [18], [19].

An analysis using propensity score matching can be broken down into 4 main steps. First, planning and estimating the propensity scores. Second, the data is matched using the propensity scores. Third, covariate balance is assessed using diagnostics that have been described in the literature. Finally, the data can be analysed to estimate the causal treatment effect. Matching procedures were implemented using the MatchIt package in R [20].

3.2 PLANNING

Matching methods involve the choice of an estimand, which is the average treatment effect with reference to a specified population. Common estimands include the average treatment effect in the entire population (ATE), the average treatment effect among the treated (ATT) and the average treatment effect among the control (ATC). The choice of estimand depends on the target population to which the treatment effect is to generalise and on the specific question of interest [21].

The ATT seeks to answer how treated participants outcomes would differ if the treatment had been withheld, it is relevant when deciding if a treatment currently implemented for a group of participants should continue to be implemented for that group. The ATT is estimated by simulating what would have happened had the treated cases not received treatment by using information from the untreated group. In contrast, the ATC can be thought of as the effect of expanding the treatment to participants who do not currently receive it, the ATC is relevant when deciding if a treatment not currently implemented for a group of participants should be extended to that group. The ATC is estimated by simulating what would have happened had the untreated cases been treated by using information from the treatment group. The ATE is the effect of the treatment for the entire population used in the study, the ATE seeks to answer how all participants outcomes would differ if treatment was given to all cases compared to if treatment was withheld from all cases [21].

We choose to estimate all three of the estimands described above, as the control and treatment groups are different in several ways, the population in which they generalise will be different. It is important to choose a matching method that is appropriate for the chosen estimand, the most common forms of matching are best suited for estimating the ATT, though some methods exist for estimating the ATE such as optimal full matching [17].

3.3 Estimating the propensity scores

Firstly, we must specify the propensity score model. Selecting the covariates carefully to be included in in the propensity score model is critical for ensuring the resulting treatment effect is free of confounding. However, there is a distinct lack of consensus in the literature as to which variables should be included within the model. Some argue that researchers should include all known variables in the propensity score model, meanwhile others caution that the inclusion of certain variables can instead amplify bias [22]. Some studies have shown that there are merits to including only the confounding variables in the propensity score model, as it results in a greater precision in estimates of the treatment effects without introducing additional bias [18], [23]. As such, it was decided that only the confounding variables would be included in the propensity score model.

Several methods exist for estimating the probability of receiving treatment given the set of covariates. In this analysis the propensity score was estimated using logistic regression with the known confounders listed in section 2.7 and 2.8 of this report as the predictor variables. Two separate models were used for intervention level ‘complete’ and intervention level ‘partial’. The estimated propensity score is then the probability of receiving the treatment resulting from the regression of treatment status on the confounding variables. The only confounding variable to be excluded from the model was Hub location due to multicollinearity with provider, one of five providers runs each Hub, with different providers each covering different locations. As the two variables are strongly correlated it is sufficient to include only one of them in the propensity score model [24]. We chose to discard Hub location over provider as this resulted in an improved balance.

3.4 Select a Matching Method and Create Matches

Now we must select a matching method and create the matched sets of treated and untreated cases with similar propensity scores. The MatchIt package implements a wide range of matching methods with a variety of options to choose from [20]. Inferences about the treatment effects are only valid if the matched sets of treated and untreated cases have similar distributions of the covariates, therefore it is critical that we achieve adequate balance in our matched sample. As the outcome variable is not included in the matching procedure, we are able to consider several different methods and estimate the treatment effect based on the method that obtains the matched sample with the best covariate balance [25].

Additional considerations to take into account when selecting a matching procedure include the ratio of control cases to be matched to each treatment case and whether replacement is allowed. These choices often involve a trade-off between bias and precision. For example, increasing the matching ratio will increase the size of the matched sample, resulting in an increased precision when estimating the treatment effect. However, increasing the number of control cases matched to each treated case may increase bias as each additional matched control case will be more dissimilar to the treated case [25]. There is no one matching method that is considered the most effective and each method can be appropriate given certain circumstances. Therefore, the matching method should be chosen on a case-by-case basis to best suit specific properties of dataset and research problem at hand.

We first perform the matching procedure and outcome analysis for when the treatment variable is intervention level 'complete', we then repeat the procedure for when the treatment variable is intervention level 'partial'.

The first method attempted was nearest neighbour matching without replacement and using a 1:1 ratio, meaning each treated case was matched to at most one control case. The remaining unpaired cases are then dropped from the sample. Nearest neighbour matching is a "greedy" algorithm which cycles though the treated cases one at a time and selects the control unit with the smallest distance measure (in this case the distance measure is the propensity score difference), once a match between a treatment and control is created, the control case is removed and cannot be considered for further matches, even if it would have been a better match for a subsequent treated case. As each match occurs without reference to how other cases will be paired, the first few matches are often good and the final matches poor [25], [26]. This matching specification yielded poor balance, so we look to other matching methods.

The next method used was optimal matching without replacement and using a 1:1 ratio. Optimal matching is similar to nearest neighbour matching but instead the pairing of untreated to treated cases is "optimal" rather than greedy, meaning that average absolute distance across all the matched pairs is minimised. This is achieved by considering all potential matches to find the best result [26]. Optimal matching provided a better covariate balance than the greedy nearest neighbour matching technique. However, covariate balance was still considered inadequate using this method.

The final method used was optimal full matching, often just called full matching. Full matching makes use of all available cases in the data, units will only be discarded if a discard option is specified. Optimal matching works by grouping the cases into a series of matched sets known as subclasses, with each subclass containing one treated and one or more control case (or one control case and one or more treated case). Full matching forms the subclasses in an optimal way, such that the weighted average of the absolute within-subclass distances is minimised. Treated cases who have many control cases with similar pre-treatment characteristics (based on the propensity score) will be grouped with more control cases, whereas treated cases with fewer similar control cases will be grouped with relatively fewer control cases [26], [27]. We chose to proceed with full matching as it yielded better balance than the other methods attempted.

Full matching produces a set of matching weights which can be incorporated in subsequent analysis to estimate the treatment effect, how the weights are calculated depends on the target estimand. For the ATE, cases are weighted by the inverse probability of receiving the treatment they received, such that both the treated cases whose probability of receiving treatment is low, and the control cases whose probability of receiving treatment is high are given more weight for the purpose of making the treatment and control groups more comparable. For the ATT, treated cases are given a weight of 1 and the control cases with a high propensity score are given more weight such that the control group more closely resembles the treatment group. Similarly, for the ATC the control cases are given a weight of 1 and the treated cases with a low propensity score are given more weight such that the treatment group more closely resembles the control group. As full matching can target the ATT, ATE and ATC as the esitmand, we will compute all three [21].

One drawback of full matching is the wildly varying ratios of control to treated cases, for example, in our matched sample it was found that the ratio of treatment:control cases ranged from 9:1 to 1:44. One way to deal with this is specifying the max controls option within MatchIt in order to limit the ratio of treated to control cases within each subclass. Min controls is the minimum ratio of controls to treatments permitted within the matched set and max controls is the maximum ratio of controls to treatments permitted within the matched set. The literature suggests that the ratio of treated:control should be limited to less than half and no more than double what it was in the original data [27].

First, when considering the intervention level 'complete' as the treatment variable, we see that for every treated individual there are approximately 4 controls, the literature suggests that the ratio of treated:control cases should be allowed to vary from 1:2 to 1:8. However, we tested several different ratios and it was found that the best balance (determined using standardised mean difference, described in more detail in the subsequent section) was achieved when ratio of treated:control cases was allowed to vary from 1:1 to 1:9. Meanwhile, when considering the intervention level 'partial' as the treatment variable, using the same method described above we found that the best balance was achieved when the ratio of treated:control cases was allowed to vary from 1:1 to 6:1. It should be noted that using the chosen ratios comes with a cost; the variance of the effect size increases due to a reduction in the effective sample size. However, a marginal increase in variance was deemed acceptable in this case due to the decrease in bias that comes from achieving good balance.

The full matching specification can be further customised by specifying a discard option, which is a method for discarding units outside a region of common support. The region of common support being the range in which the propensity scores for the treatment and control group overlap. When setting the discard option to "both", participants from both treatment and control groups whose propensity scores fall outside the corresponding region are discarded, this ensures that the remaining cases have a sufficient overlap of pre-treatment characteristics. However, we must take care when discarding units and keep in mind the trade-off between having an improved balance and the bias introduced from the exclusion of cases from the matched sample. Using the discard option set to "both" yielded a better balance, as it drops cases in which no basis exists for comparison. As the number of cases discarded comprise a small percentage of the entire sample, we chose to proceed using the discard option. However, a disadvantage is that estimand no longer corresponds to the ATE as it no longer reflects the effect for the entire study population, and rather a subgroup of the targeted population [17]. As the number of cases that were discarded is relatively few, we assume that the estimate of the ATE is likely very close to the true ATE such that the change to the estimand from discarding cases is negligible.

For the 'complete' intervention analysis, the sample size before matching was 308 treated and 1186 control cases. A total of 78 control cases and 1 treated case were discarded from the sample. As we have discarded 78 control cases, we re-estimate the propensity score using only the remaining cases. For the 'partial' intervention analysis, there are more individuals in the treatment group than the control group. Thus, we swap the treatment and control groups for the matching procedure. The rest of the analysis is then performed as normal; however, we must take care regarding the estimand we are computing as swapping the treated and control cases results in the ATT and ATC being swapped. The sample size before matching was 1649 treated and 1186 control cases. A total of 4 control cases and 0 treated cases were discarded from the sample.

3.5 Assessing Balance

After obtaining our matched sample, we must evaluate our chosen matching specification using diagnostics that have been described in the literature to ensure that adequate covariate balance has been achieved. We should only move onto the next step of estimating the treatment effect after we have achieved a satisfactory covariate balance between treatment and control groups, as otherwise the study will not be able to robustly estimate the effect of the treatment.

The standardized mean difference (SMD) is the most commonly used statistic to examine the balance of covariate distribution between treatment groups after propensity score matching. The SMD is the difference in the means between treatment and control groups for each covariate standardised so that it is on the same scale for all covariates. We also look at the effective sample size (ESS) which denotes the remaining sample size after adjusting for weighting and serves as a measure of loss in precision due to the weights generated by full matching [28].

Several recommended thresholds have been described in the literature, with most studies considering balance to have been achieved when the absolute value of the SMD is less than 0.1. However, it should be noted that this threshold of 0.1 is somewhat arbitrary. We use the following criterion based on the SMD as a benchmark of excellent balance.

1.) The mean absolute SMD, including interactions, is less than 0.05
2.) At least 50% of absolute SMDs, including interactions, are less than 0.05
3.) At least 75% of absolute SMDs, including interactions, are less than 0.1
4.) At least 75% of absolute SMDs, excluding interactions, are less than 0.05
5.) All absolute SMDs, excluding interactions, are less than 0.1

Table 3-1 shows the covariate balance before propensity score matching, as well as the covariate balance for when the estimand is the ATT, ATE and ATC, respectively. It can be seen in Table 3-1 that the SMD is largely reduced in the matched samples, although they do not meet the above specified criteria, missing it by a narrow margin. However, the matched sample was considered adequate with respect to balance as no serious imbalance (SMD > 0.2, excluding interactions) remained.

Looking at the initial imbalance we can see that when the treatment is 'complete' intervention, 64.89% of the matching variables showed significant differences (absolute SMDs > 0.1). Meanwhile, when the treatment is 'partial' intervention, 36.67% of the matching variables showed significant differences (absolute SMDs > 0.1). There is less initial imbalance between the 'none' and 'partial' groups, this is as we expected since the 'partial' group is somewhere between 'none' and 'complete' with respect to pre-treatment characteristics.

Table 3-2 and Table 3-3 show the largest SMDs after matching for intervention levels 'complete' and 'partial', respectively. Figure 3-1 to Figure 3-6 show the distributional balance for covariates with the largest SMDs. For example, from Figure 3-1 we can see that sexual offences are unequally distributed between the treatment and control groups, with offences of this type making up a larger proportion of the treatment group. Although the SMD remains greater than 0.1, it can be seen that matching has largely improved the balance. However, as some imbalance remains, we must be aware that there will be some remaining bias in the effect estimate.

As units have been discarded, we must ensure that the effective sample size (ESS) is adequate in the matched sample else the effect estimates will lack precision. When comparing 'complete' intervention to no intervention, we find that after matching the ESS when the target estimand is the ATE is 188.27 treated cases and 1035.21 control cases, the ESS when the target estimand is the ATT is 307 treated cases and 444.29 control cases and the ESS when the target estimand is the ATC is 151.49 treated cases and 1108 control cases. Whilst the ESS may appear to be a large decrease from the original sample size, it indicates that a large number of cases were dissimilar and hence not useful for estimating the treatment effect in the population in which we are interested.

When comparing 'partial' intervention to no intervention, we find that after matching the ESS when the target estimand is the ATE is 1572.82 treated cases and 935.95 control cases, the ESS when the target estimand is the ATT is 1649 treated cases and 657.61 control cases and the ESS when the target estimand is the ATC is 1290.45 treated cases and 1182 control cases.

Table 3-1 Covariate Balance Summary

Intervention level: Complete	Initial	ATT	ATE	ATC
Mean Absolute SMD (including interactions)	0.107	0.056	0.060	0.067
Median Absolute SMD (including interactions)	0.080	0.045	0.051	0.055
Max Absolute SMD (including interactions)	1.053	0.396	0.278	0.374
% Absolute SMDs < 0.1 (including interactions)	59.20%	84.95%	82.37%	78.70%
% Absolute SMDs < 0.05 (including interactions)	32.49%	54.28%	49.32%	46.11%
% Absolute SMDs < 0.1 (excluding interactions)	35.11%	88.30%	88.30%	80.85%
% Absolute SMDs < 0.05 (excluding interactions)	19.15%	55.32%	60.64%	52.13%
Intervention level: Partial	**Initial**	**ATT**	**ATE**	**ATC**
Mean Absolute SMD (including interactions)	0.063	0.038	0.035	0.037
Median Absolute SMD (including interactions)	0.048	0.032	0.030	0.030
Max Absolute SMD (including interactions)	0.732	0.401	0.299	0.345
% Absolute SMDs < 0.1 (including interactions)	81.52%	95.89%	97.31%	96.20%
% Absolute SMDs < 0.05 (including interactions)	52.02%	72.77%	75.38%	72.66%
% Absolute SMDs < 0.1 (excluding interactions)	63.33%	94.44%	96.67%	95.56%
% Absolute SMDs < 0.05 (excluding interactions)	31.11%	80.00%	76.67%	66.67%

Table 3-2 Largest SMDs after matching (excluding interactions) – Intervention level: Complete		
Estimand	**SMD**	**Comments**
ATT	0.183	Offence: Sexual – Treatment group has a higher mean score than the control group
ATE	0.167	How much support does the participant get from family: 5 – Treatment group has a lower mean score than the control group
ATC	0.181	Provider name: Reed – Treatment group has a higher mean score than the control group

Table 3-3 Largest SMDs after matching (excluding interactions) – Intervention level: Partial		
Estimand	**SMD**	**Comments**
ATT	0.192	Provider name: Seetec – Treatment group has a lower mean score than the control group
ATE	0.185	Provider name: Seetec – Treatment group has a lower mean score than the control group
ATC	0.196	Provider name: Seetec – Treatment group has a lower mean score than the control group

Figure 3-1 Matched vs Unmatched balance for Offence – Complete – Estimand: ATT

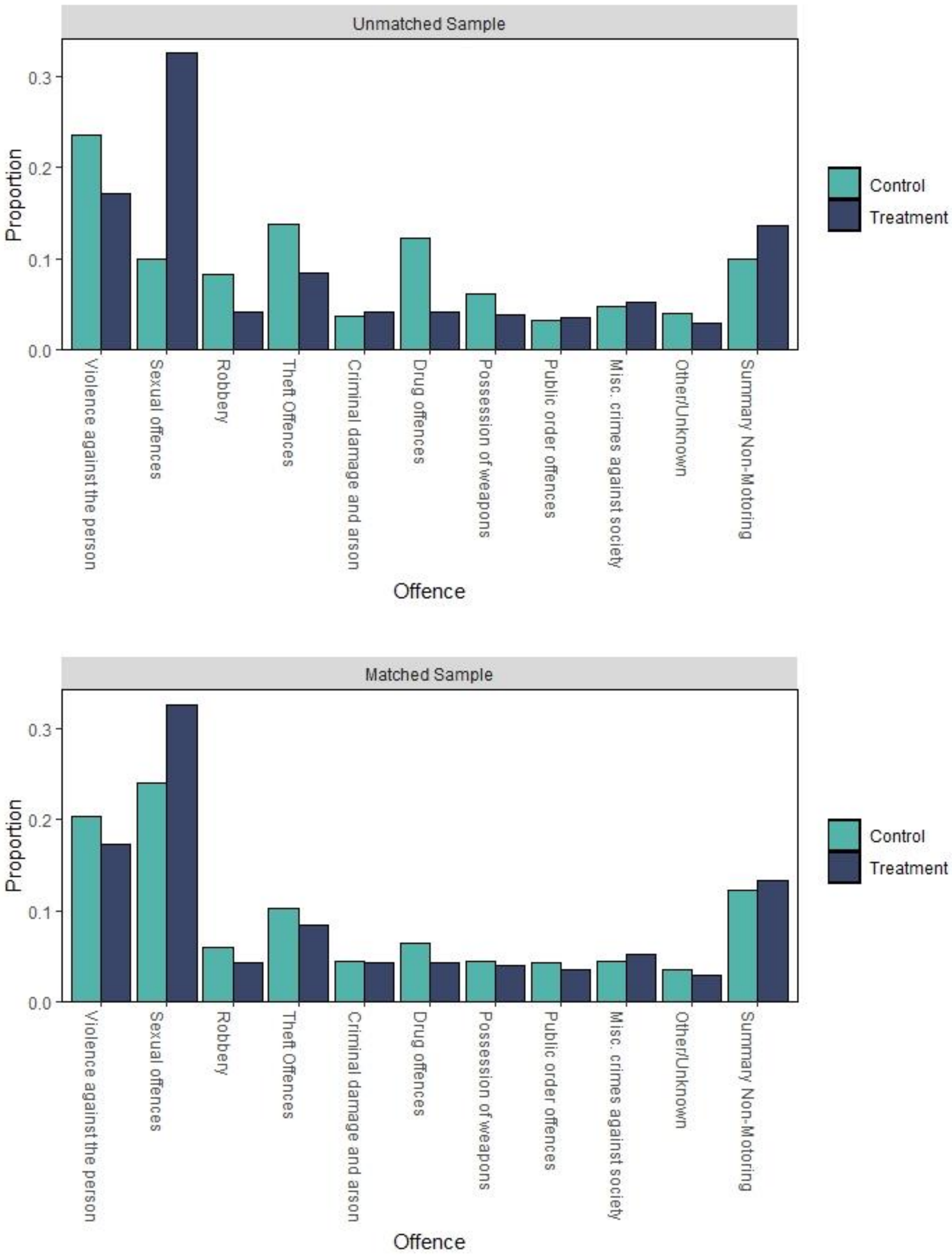


Figure 3-2 Matched vs Unmatched balance – Complete – Estimand: ATE

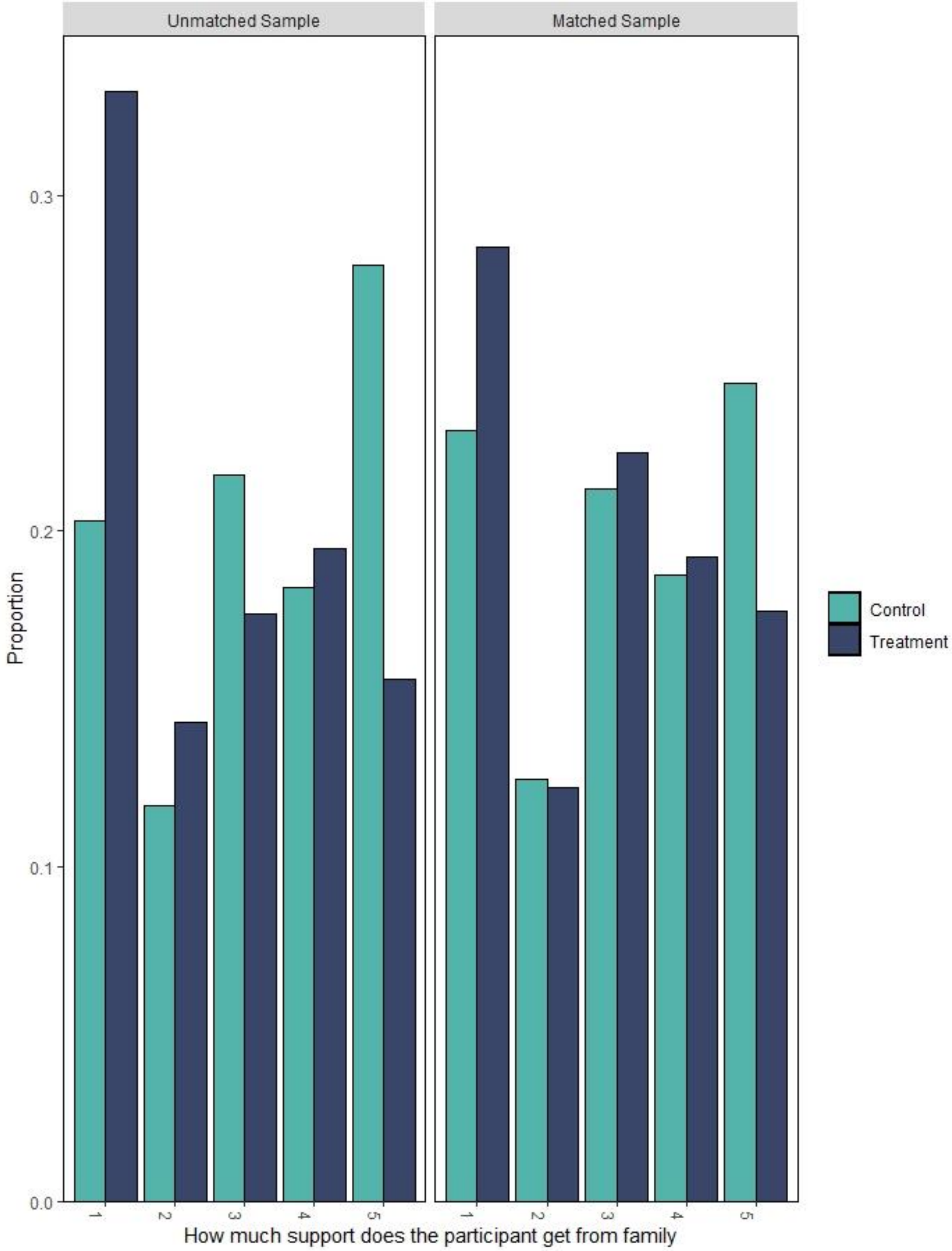


Figure 3-3 Matched vs Unmatched balance – Complete – Estimand: ATC

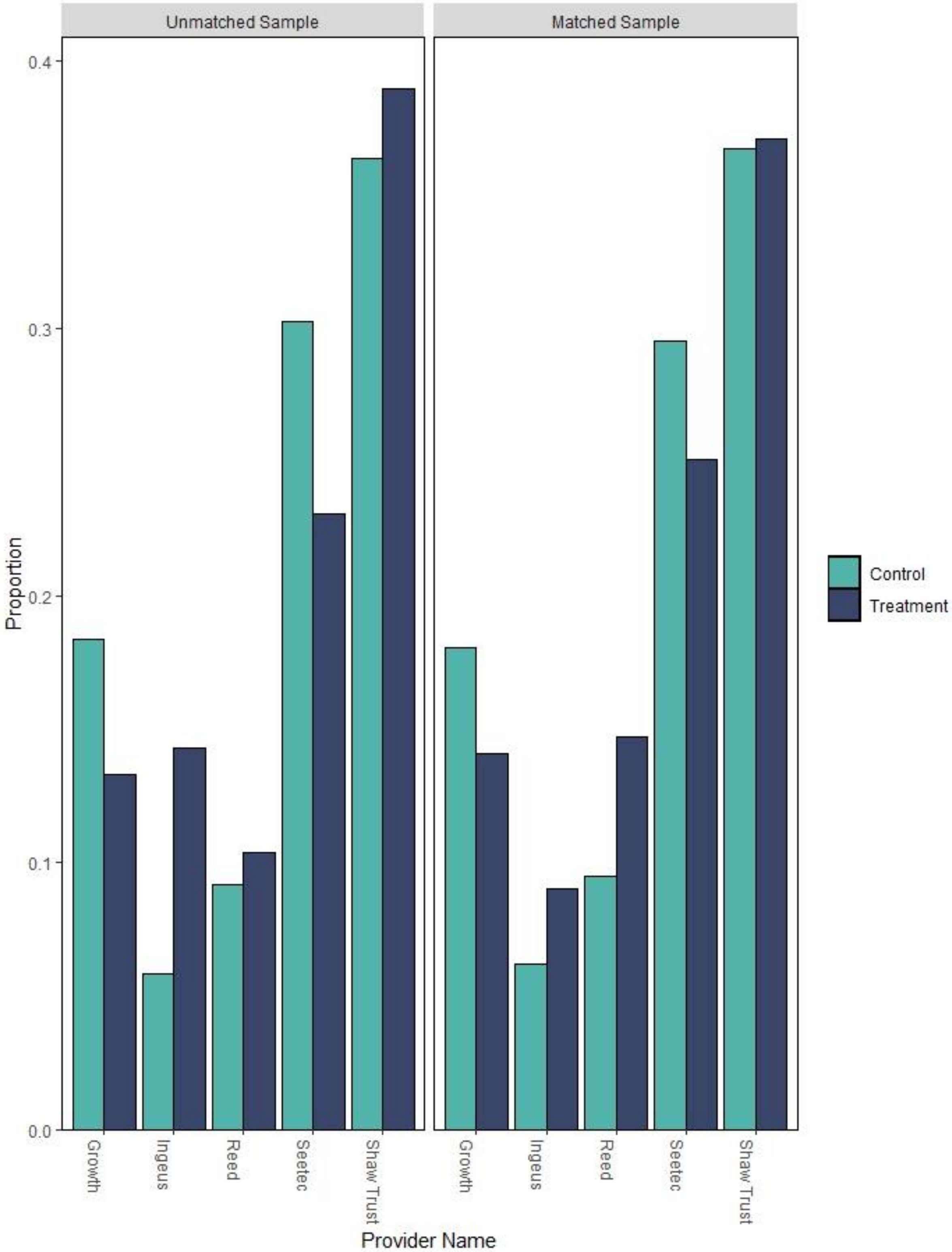


Figure 3-4 Matched vs Unmatched balance – Partial – Estimand: ATT

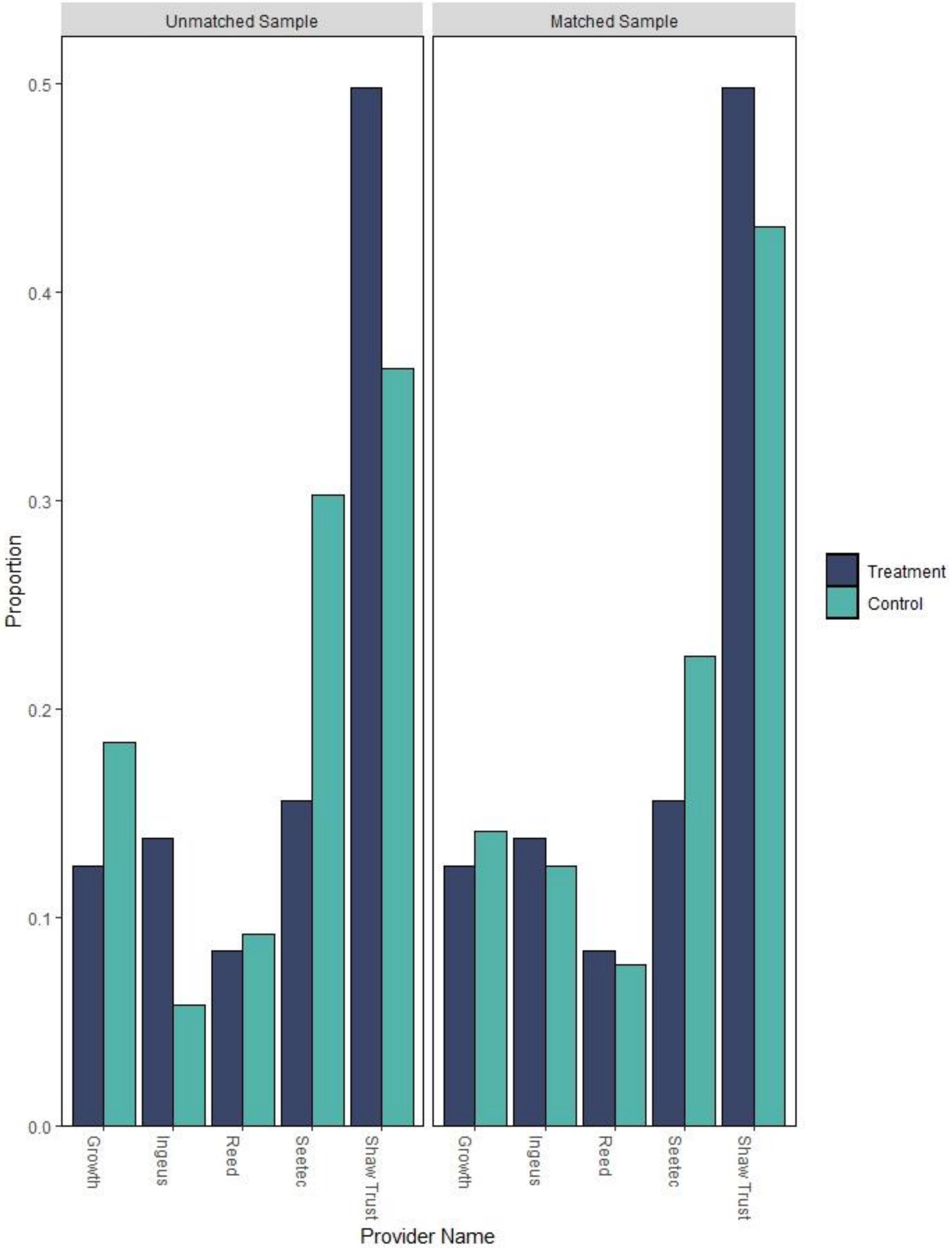


Figure 3-5 Matched vs Unmatched balance – Partial – Estimand: ATE

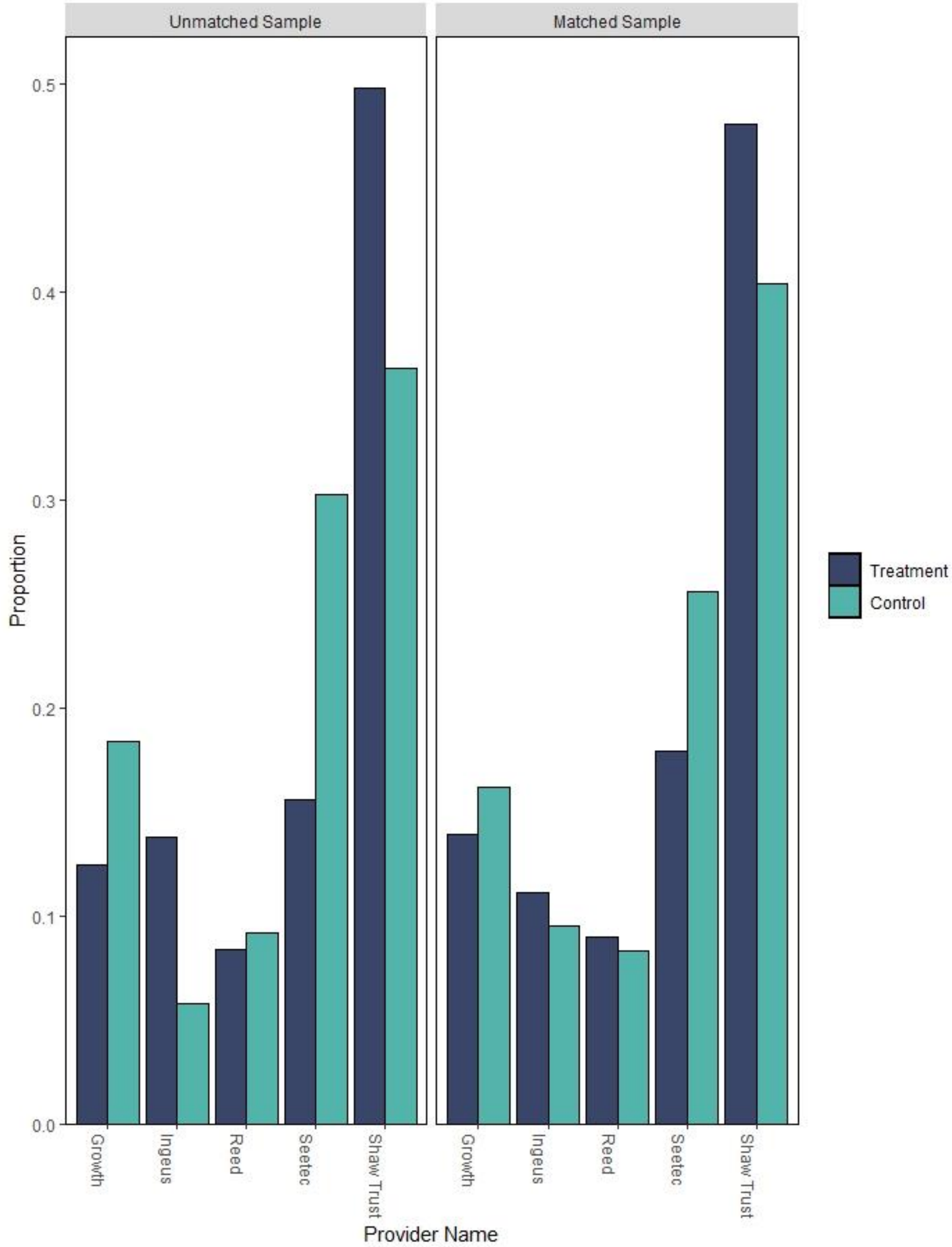
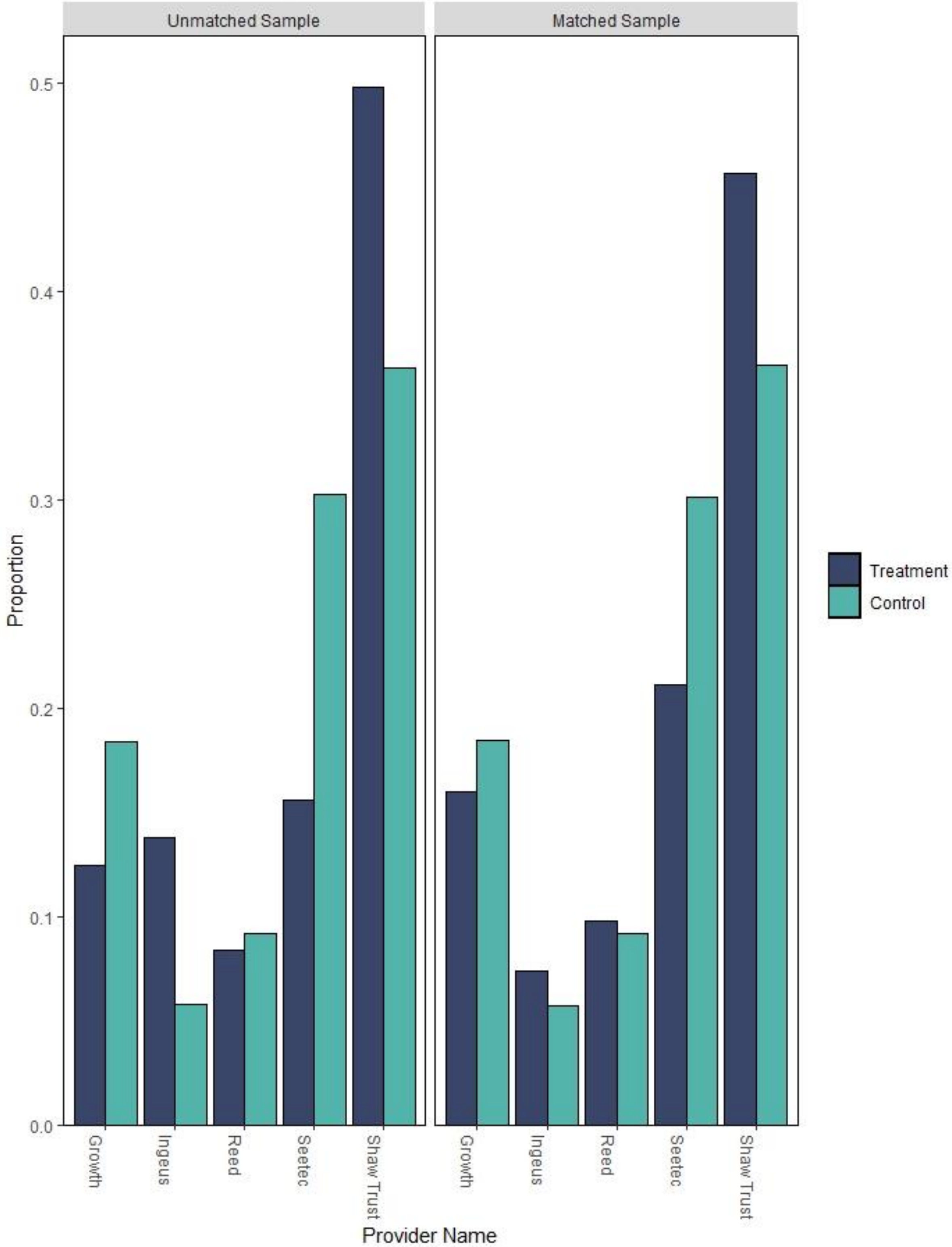


Figure 3-6 Matched vs Unmatched balance – Partial – Estimand: ATC



3.6 Specifying an Effects Model

Now we have achieved adequate balance within the matched sample, implying that the covariate distribution is similar across the treatment and control groups. The next step is to specify an effects model that can be trained on the matched dataset and used to estimate the marginal treatment effect by simulating two expected potential outcomes for each case, one for under treatment and the other under control. The marginal effect being the difference between these expected potential outcomes [29].

Two separate models were used, one for intervention level 'complete' and another for 'partial'. We use a weighted logistic regression model to regress the occurrence of reoffending (as defined in Section 2.2) on the treatment variable ('complete' intervention for part 1, 'partial' intervention for part 2), with the confounding variables identified earlier in the report included as predictor variables for further adjustment on residual imbalance. The models incorporate the weights generated by full matching to ensure that the matched treated and control groups are weighted up to be similar, the weights differing according to our choice of estimand. Both models included an interaction term between the treatment and the moderating factors, as the treatment effects vary depending on levels of the moderating factor.

3.7 Estimating Effects

Now that we have specified the model, we can now estimate marginal effects using the g-computation method. When the target estimand is the ATE, we predict outcomes for every case in the sample, first setting their status to treated, and then repeating with their status set to control, such that we have two predicted outcomes for each case. From this we then compute the weighted mean of these predicted outcomes, incorporating the matching weights, for both the treatment and control such that we have two average estimated potential outcomes. We then compute the relative risk, which compares the risk of reoffending between the treatment and control group, given by dividing the average estimated potential outcome for when the status is set to treated by the average estimated potential outcome for when the status is set to control. A relative risk of 1 suggests that the treatment has no effect, whilst a relative risks less than or greater than 1 represent a positive or negative effect on reoffending, respectively. When the target estimand is the ATT or ATC, we follow the same procedure as for the ATE but rather than predicting potential outcomes for every case, we instead restrict our predictions to the subset of treated or control cases, respectively [29].

3.8 Bootstrapping Confidence Intervals

The final step is to construct confidence intervals for the estimated effect size using the block bootstrap resampling method which we implement in R using the boot package [30]. The idea of bootstrapping is to generate many new datasets by drawing random samples with replacement from our matched dataset. The block bootstrap slightly modifies the procedure, we instead resample over blocks, in this case the subclasses generated by full matching, rather than individual cases. The original matched dataset consists of $\boldsymbol{n}$ subclasses, we randomly sample these $\boldsymbol{n}$ subclasses with replacement to make up each bootstrap sample, such that each sample also consists of $\boldsymbol{n}$ subclasses. We take 9999 bootstrap replications and within each of these bootstrap samples we estimate the marginal relative risk using the g-computation method described above. The distribution of the resulting relative risk estimates across the replications is then used to construct the 95% confidence intervals. To do so we take the 2.5th percentile and the 97.5th percentile of the estimated relative risk across all the bootstrap samples [17], [29].

3.9 Part 2 Results – Full CFO Activity Hubs experience

Table 3-4 Estimated Effect Size – Intervention level: Complete – Estimand: ATT

Model	RR	LL*	UL*
Crude Observed Effect	0.622	-	-
Controlling for Confounders	0.822	0.741	0.915

*2.5 and 97.7 percentile based on 9999 bootstrap replicates

Table 3-5 Estimated Effect Size – Intervention level: Complete – Estimand: ATE

Model	RR	LL*	UL*
Crude Observed Effect	0.622	-	-
Controlling for Confounders	0.619	0.591	0.650

*2.5 and 97.7 percentile based on 9999 bootstrap replicates

Table 3-6 Estimated Effect Size – Intervention level: Complete – Estimand: ATC

Model	RR	LL*	UL*
Crude Observed Effect	0.622	-	-
Controlling for Confounders	0.560	0.533	0.587

*2.5 and 97.7 percentile based on 9999 bootstrap replicates

3.10 Part 2 Results – Partial CFO Activity Hubs experience

Table 3-7 Estimated Effect Size – Intervention level: Partial – Estimand: ATT

Model	RR	LL*	UL*
Crude Observed Effect	0.650	-	-
Controlling for Confounders	0.784	0.757	0.811

*2.5 and 97.7 percentile based on 9999 bootstrap replicates

Table 3-8 Estimated Effect Size – Intervention level: Partial – Estimand: ATE

Model	RR	LL*	UL*
Crude Observed Effect	0.650	-	-
Controlling for Confounders	0.801	0.782	0.821

*2.5 and 97.7 percentile based on 9999 bootstrap replicates

Table 3-9 Estimated Effect Size – Intervention level: Partial – Estimand: ATC

Model	RR	LL*	UL*
Crude Observed Effect	0.650	-	-
Controlling for Confounders	0.854	0.828	0.880

*2.5 and 97.7 percentile based on 9999 bootstrap replicates

3.11 Sensitivity analysis

Even with adjustment for the observed confounding variables, observational studies are still subject to biases from residual and unmeasured confounding, meaning that an unknown third factor may explain the association between the treatment and the outcome, compromising the validity of the results. The impact of unmeasured confounding on causal associations can be evaluated by means of sensitivity analyses. If there is an important but unmeasured confounder missing from the propensity score model then this results in bias; a sensitivity analysis can tell us how strong the relationship between the unmeasured confounder and the treatment would have to be, as well as between the unmeasured confounder and the outcome, in order for there to no longer be a causal association [32].

To conduct sensitivity analyses for unmeasured confounding we use the R package, EValue [32]. The E-value is defined as the minimum strength of association, on the risk ratio (relative risk) scale, that unmeasured confounder(s) would need to have with both the treatment and the outcome in order to negate the observed treatment-outcome association [32], [33].

Let E and D represent the treatment and outcome variables, respectively, and let U represent the possible binary unmeasured confounder(s). Let RR_{UD} be the maximum relative risk of the outcome comparing cases with ($U = 1$) and without ($U = 0$) the unmeasured confounding variable across both the treatment and control groups. Let RR_{EU} be the maximum relative risk of $U = 1$ (or $U = 0$) comparing cases in the treatment group to cases in the control group. Essentially, RR_{UD} captures how much the unmeasured confounder influences the outcome and RR_{EU} captures the imbalance between the treatment and control groups for the unmeasured confounder U [32]. For example, if 30% of the control group have $U = 1$, compared with 20% of the treatment group, then we have $RR_{EU} = 1.5$.

As can be seen in Table 3-10, the E-value is 1.730 for when the treatment variable is intervention level 'complete' and the estimand is the ATT. This E-value means that in order for the observed relative risk of 0.822 to be explained away by unmeasured confounding, the confounding variable(s) would have to: first, increase the risk of reoffending 1.730-fold among either the treatment or control group and second, be 1.730 times more prevalent in either the treatment or control group, with any weaker confounding being unable to explain away the treatment-outcome association. Similarly, the E-value for the upper confidence limit, which is the confidence limit closest to the null (RR=1) is 1.412, meaning that an unmeasured confounder associated with both the treatment and outcome by a relative risk of 1.412-fold each could move the upper limit of the confidence interval to the null, but weaker confounding could not.

The lowest E-value for the upper confidence limit is 1.412 for when the treatment variable is intervention level 'complete' and the estimand is the ATT. An unmeasured confounder that is associated with both intervention level and reoffending by a risk ratio of 1.412-fold does not seem implausible, though we thought this unlikely. Overall, the evidence for causality from the E-values looks to be reasonably strong, especially for when the treatment variable is intervention level 'complete' and the estimand is the ATE or ATC as substantial unmeasured confounding would be needed in order to move the upper limit of the confidence interval to null.

Error! Reference source not found. illustrates the different combinations RR_{EU} and RR_{UD} can take that would have the joint minimum strength to explain away the treatment-outcome association for when the treatment variable is intervention level 'complete' and the estimand is the ATT. The E-value is the point on the curve when $RR_{EU} = RR_{UD}$.

Table 3-10 E-value – Intervention level: Complete – Estimand: ATT

	Point	Lower	Upper
RR	0.822	0.741	0.915
E-values	1.730	-	1.412

Table 3-11 E-value – Intervention level: Complete – Estimand: ATE

	Point	Lower	Upper
RR	0.619	0.591	0.650
E-values	2.613	-	2.449

Table 3-12 E-value – Intervention level: Complete – Estimand: ATC

	Point	Lower	Upper
RR	0.560	0.533	0.587
E-values	2.970	-	2.798

Table 3-13 E-value – Intervention level: Partial – Estimand: ATT

	Point	Lower	Upper
RR	0.784	0.757	0.811
E-values	1.868	-	1.769

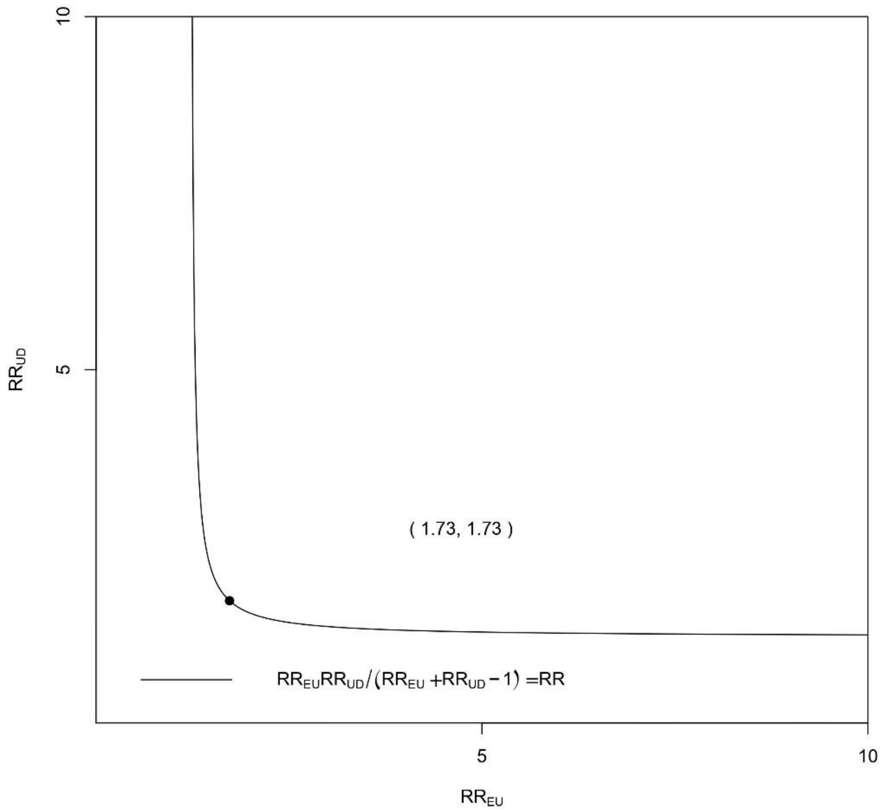
Table 3-14 E-value – Intervention level: Partial – Estimand: ATE

	Point	Lower	Upper
RR	0.801	0.782	0.821
E-values	1.805	-	1.733

Table 3-15 E-value – Intervention level: Partial – Estimand: ATC

	Point	Lower	Upper
RR	0.784	0.757	0.811
E-values	1.618	-	1.530

Figure 3-7 Plot of combinations of RR_{EU} and RR_{UD} for $RR = 0.822$ – Complete – Estimand: ATT



3.12 Potential bias and other limitations

There are a number of limitations to a study of this kind. Observational studies are subject to biases due to the lack of randomisation, as such we must be cautious when inferring causal effects, as not all bias and confounding cannot be accounted for in this analysis.

Since the CFO Activity Hubs programme is voluntary, it is likely that those who volunteer will not be representative of the general population of offenders, effecting the generalisability of the study results to the wider offender population. Also, some cases have been discarded during the matching procedure due to not having an adequate match, changing the population which the effect is meant to generalise. We must also consider that reoffending alone is a narrow measure of impact as not all crime is detected and sanctioned.

Propensity score matching relies on the assumption of conditional independence, which is the assumption that all confounding variables are observable and measured accurately [31]. However, in practice this is hard to satisfy as it is possible that important contextual information that may help explain the results has not been accounted for. For example, we have no information on drug and alcohol misuse, which is a potential confounding factor as substance misuse is known to amplify reoffending risk and is also suspected to influence engagement with the CFO Activity Hubs programme [3]. There is also the fact that many of the variables measured are self-declared or up to the interpretation of the provider. For example, some of the variables are measured on a numerical scale of one to five which can be subject to personal inclination to prefer certain response styles such as only giving answers on the extreme end of the scale.

Another limitation is selection bias, those who more actively engage with the Hubs are likely to be highly motivated, resulting in a positive selection bias. This means that we could expect a reduced reoffending rate amongst this group as they are more motivated to change. Alternatively, the programme specifically targets those who are known to have more complex needs and barriers. For this reason, we could expect an increased reoffending rate amongst this group as addressing their needs is more challenging, resulting in a negative selection bias. This means that participants may have particular characteristics relating to motivation and barriers that are not represented in the data, which could lead to selection bias in either direction.

Full matching resulted in a smaller effective sample size (ESS) than some of the other matching methods attempted and resulted in an ESS that was up to 50% smaller than the initial sample size (for when the estimand is the ATT and the treatment variable is intervention level 'complete'). This can result in imprecise estimates and loss of statistical power. Another limitation is the use of broad categories for some of the confounders in the analysis. For example, the number of times a participant has been in custody prior to enrolling is stratified into three groups only, which likely results in some residual confounding. It would have been preferable to stratify more finely but due to the small sample size this was not possible.

A further limitation is that the confidence intervals for the estimated effect size are large due to the small sample size; therefore, more participants are necessary to precisely estimate the effects the CFO Activity Hubs programme has on reoffending rates. However, this should not be taken to mean that the programme does not reduce reoffending rates as the null is not contained within the 95% confidence intervals.

4 CONCLUSIONS

Individuals who have received 'complete' intervention are observed to be 37.8% less likely to reoffend than those who have received no intervention. However, individuals in this group are observed to be on an average 5 years older than the 'none' group, have served longer sentences (on average 449 days more) and have a lower offending intensity, all of which are factors associated with a reduced reoffending risk. Controlling for confounders we find that individuals who received an intervention level of 'complete' are 8.5%-25.9%, 35.0%-40.9% and 41.3%-46.7% less likely to reoffend than those who received an intervention level of 'none' when looking at the ATT, ATE and ATC, respectively. Hence, the preventative effect 'complete' intervention has on reoffending is the greatest for those who have not received the treatment, i.e., have received no intervention.

Individuals who received 'partial' intervention are observed to be 35.0% less likely to reoffend than those who have received no intervention. As was the case with the 'complete' intervention group, those that received 'partial' intervention are on average 3 years older than the 'none' group, have served longer sentences (on average 272 days more) and have a lower offending intensity, but greater than that of the 'complete' group. All of which are factors associated with a reduced reoffending risk. Controlling for confounders we find that individuals who have received 'partial' intervention are 18.9%-24.3%, 17.9%-21.8% and 12.0%-17.2% less likely to reoffend than those who received no intervention when looking at the ATT, ATE and ATC, respectively. The preventative effect of receiving 'partial' intervention on reoffending appears to be somewhat similar for both those have received the treatment and those who have not.

Using the ATT, we can predict what the reoffending rate would have been, had the intervention been withheld (i.e., those who received 'partial' or 'complete' intervention instead received an intervention level of 'none'). The six-month observed reoffending rate of the 1,649 individuals who received 'partial' intervention was 10.86% compared with 13.85% (95% CI [13.40%, 14.33%]) for a matched group of similar individuals who did not receive intervention. This is a 2.99 (95% CI [2.54, 3.47]) percentage-point difference. The six-month observed reoffending rate of the 308 individuals who received 'complete' intervention was 10.39% compared with 12.64% (95% CI [11.31%, 13.97%]) for a matched group of similar individuals who did not receive intervention. This is a 2.25 (95% CI [0.92, 3.58]) percentage-point difference.

Overall, the six-month observed reoffending rate all 3,143 individuals enrolled onto the programme was 13.01%, compared with 14.80% (95% CI [14.43%, 15.18%]) for a matched group of similar individuals who did not receive intervention. This is a 1.79 (95% CI [1.42, 2.17]) percentage-point difference (1.79 participants prevented from reoffending per 100 participants) or approximately a 12% reduction in the rate of reoffending. This represents the impact of CFO Activity Hubs intervention in real world practice. The overall effectiveness of the intervention is diminished as only a portion (62%) of those enrolled are receiving either 'partial' or 'complete' intervention at present.

Using the ATC, we can predict what the reoffending rate would be if 'complete' intervention was expanded to all participants enrolled on the programme (i.e., those who had received an intervention level of 'none' or 'partial' intervention, instead received 'complete' intervention). The six-month observed reoffending rate of the 1,186 individuals who received no intervention was 16.69%, compared to 9.34% (95% CI [8.91%, 9.82%]) for a matched group of similar individuals who received 'complete' intervention. This is a 7.35 (95% CI [6.87, 7.78]) percentage-point difference. The six-month observed reoffending rate of the 1,649 individuals who received 'partial' intervention was 10.86%, compared to 9.69% (95% CI [8.91%, 10.57%]) for a matched group of similar individuals who received 'complete' intervention. This is a 1.17 (95% CI [0.29, 1.95]) percentage-point difference.

The six-month observed reoffending rate of the 3,143 individuals in the sample was 13.01%, compared with 9.62% (95% CI [9.05%, 10.27%) for a matched group of similar individuals who received 'complete' intervention. This is a 3.39 (95% CI [2.74, 3.96]) percentage-point difference (3.39 participants prevented from reoffending per 100 participants) or approximately a 26% reduction in the rate of reoffending should all participants receive 'complete' intervention. This represents the potential outcomes under the most ideal circumstances, which in this case is 100% uptake of 'complete' intervention. This result overestimates the effect of the intervention, as it is unlikely that all those that enrol onto the programme would receive 'complete' intervention in practice.

Table 4-1 and Figure 4-1 show some projections on how the effectiveness of CFO Activity Hubs intervention could improve if uptake was increased. At the current level of uptake, 1.79 participants are prevented from reoffending for every 100 participants. However, if an additional 10% of those currently receiving no intervention instead received 'partial' or 'complete' intervention, the number of participants prevented from committing a reoffence would increase by 5% and 15%, respectively. Similarly, if an additional 50% of those currently receiving no intervention instead received 'partial' or 'complete' intervention, the number of participants prevented from committing a reoffence would increase by 26% and 77%, respectively.

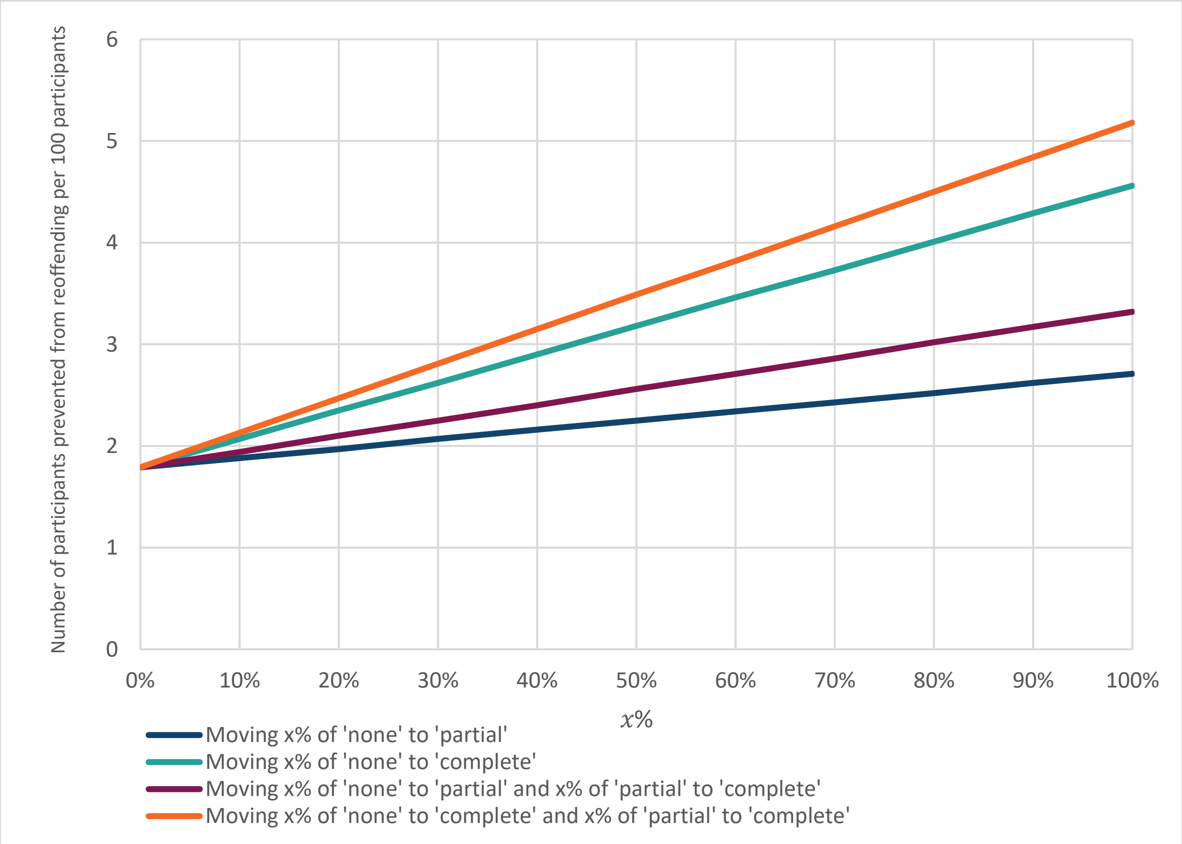
If 10% of those currently receiving no intervention instead received ‘partial’ intervention in addition to 10% of those currently receiving ‘partial’ intervention instead received ‘complete’ intervention, the number of participants prevented from committing a reoffence would increase by 9%. Similarly, if 50% of those currently receiving no intervention instead received ‘partial’ intervention in addition to 50% of those currently receiving ‘partial’ intervention instead received ‘complete’ intervention, the number of participants prevented from committing a reoffence would increase by 43%.

If 10% of those currently receiving no intervention instead received ‘complete’ intervention, in addition to 10% of those currently receiving ‘partial’ intervention instead receiving ‘complete’ intervention, the number of participants prevented from committing a reoffence would increase by 19%. Similarly, if 50% of those currently receiving no intervention instead received ‘complete’ intervention, in addition to 50% of those currently receiving ‘partial’ intervention instead receiving ‘complete’ intervention, the number of participants prevented from committing a reoffence would approximately double (increase by 95%).

Table 4-1 Projected number of participants per 100 prevented from offending

	$x\%$			
	0%	10%	50%	100%
Moving $x\%$ of 'none' to 'partial'	1.79	1.88	2.25	2.71
Moving $x\%$ of 'none' to 'complete'	1.79	2.07	3.18	4.56
Moving $x\%$ of 'none' to 'partial' and $x\%$ of 'partial' to complete	1.79	1.94	2.56	3.32
Moving $x\%$ of 'none' to partial and $x\%$ of 'partial' to complete	1.79	2.13	3.49	5.18

Figure 4-1 Projected number of participants per 100 prevented from offending



It should be noted that the projected number of participants per 100 prevented from reoffending exceeds that which was estimated using the ATC (3.39 participants prevented from reoffending per 100 participants). This is because we are combining the estimate of how many participants were prevented from reoffending by receiving an intervention level of 'partial' or 'complete', with the estimate of how many additional participants would be prevented from reoffending if uptake of 'partial' and 'complete' intervention were to further increase.

There is some preliminary evidence suggesting that there is no significant effect of ethnicity, gender, age, homelessness, or disability on the efficacy of CFO Activity Hubs intervention. This means that the benefits of receiving intervention can be considered equal across these groups. Although, some of these groups do not have comparable 'deliverability', for example, approximately 59% of men received 'partial' intervention, compared to 53% of women. For age, 67% of those aged 50 to 59 received 'partial' intervention compared to 52% of those aged under 30. Similarly, homeless participants and those that do not have a disability are more likely to receive 'partial' intervention that their counterparts. We did not examine the difference in efficacy amongst subgroups that received 'complete' intervention due to the small sample sizes.

In contrast, the impact of CFO Activity Hubs intervention differs across certain groups. With those who have received CFO3 intervention prior to, or whilst enrolled on the Activity Hubs programme benefiting particularly from CFO Activity Hubs intervention. Similarly, the impact of intervention is greater for participants that have dependent children, are a lone parent, do not have basic literacy skills, feel they do not possess good personal qualities or have spent greater than three years in custody for their index offence (most recent proven offence prior to enrolling). In addition to greater efficacy, deliverability is also higher amongst these groups. It is unclear from the data alone why such effects are present and further work would be necessary to provide clarity in this area.

Overall, the findings of this study suggest that over a 6-month follow up period after enrolling onto the CFO Activity Hubs programme, participants who received 'partial' or 'complete' intervention, were less likely to reoffend compared to those that received no intervention. The impact of receiving 'complete' intervention for those who ultimately received it is small by comparison to what it would have been for otherwise equivalent participants who did not receive intervention. In other words, the expected reduction in reoffending would have been greater for those that did not receive intervention than it was for those that ultimately received 'complete' intervention. This is due to the distribution of factors (i.e., age, offence) differing between those that received and intervention level of 'complete' intervention and those that did not receive any intervention. The 'complete' intervention group possess more factors associated with a reduced reoffending risk, thus had less to gain from receiving intervention. There would therefore seem to be a definite need to maximise retention to ensue those that stand to benefit the most from the intervention are receiving it.

The gain from moving participants from 'partial' intervention to 'complete' intervention is minimal when compared to moving participants currently receiving no intervention to 'partial' or 'complete'. This again suggests that a focus on retention should be prioritised. Further analysis is required in order to identify the point of diminishing returns, in which any further Hub activity results in minimal reduction to reoffending rates.

5 RECOMMENDATIONS

The following are a list of recommendations based on the findings of this study.

- The marginal treatment effect has been found to be greatest for those who have enrolled onto the Activity Hubs programme but received no intervention. However, the 26% reduction to participants reoffending rate will not start to materialise until all receive complete intervention, suggesting the focus should be on maximising retention. Therefore, a qualitative study to investigate what keeps participants returning to the Hubs is recommended.

- The precision of the estimates could be improved if the size of the study cohort used in the analysis was increased. Therefore, it is recommended that the analysis is repeated on a larger sample, six months after the programme has concluded. On this basis, it is also recommended that the analysis is repeated with individuals who have not previously had custody sentences included also.

- This analysis revealed that impact of CFO Activity Hubs intervention is greater for participants who have previously been enrolled on CFO3, have dependent children, are a lone parent, do not have basic literacy skills, feel they do not possess good personal qualities or have spent greater than three years in custody for their index offence. In contrast, there was no significant effect of ethnicity, gender, age, homelessness, or disability on the reoffending impact of the intervention, although deliverability was not comparable for these groups. On this basis, a further study is recommended in order to better understand the efficacy and deliverability of the CFO Activity Hubs intervention amongst different subgroups and the intersectionality of these groups, for example, where gender and ethnicity interact.

6 References

[1] "Bigger, better Probation Service to cut crime," 28 June 2021. [Online]. Available: https://www.gov.uk/government/news/bigger-better-probation-service-to-cut-crime.

[2] A. Newton, X. May, S. Eames and M. Ahmad, "Economic and social costs of reoffending," Ministry of Justice Analytical Series, 2019. [Online]. Available: https://www.gov.uk/government/publications/economic-and-social-costs-of-reoffending.

[3] I. Brunton-Smith and K. Hopkins, "The factors associated with proven re-offending following release from prison: findings from Waves 1 to 3 of Surveying Prisoner Crime Reduction (SPCR)," Ministry of Justice Analytical Series, 2013. [Online]. Available: https://www.gov.uk/government/publications/results-from-the-surveying-prisoner-crime-reduction-survey.

[4] "Transforming Rehabilitation: a summary of evidence on reducing reoffending," Ministry of Justice Analytical Series , 2013. [Online]. Available: https://www.gov.uk/government/publications/transforming-rehabilitation-a-summary-of-evidence-on-reducing-reoffending.

[5] "Basic housing to keep offenders off streets and cut crime," 18 July 2023. [Online]. Available: https://www.gov.uk/government/news/basic-housing-to-keep-offenders-off-streets-and-cut-crime.

[6] S. Lundrigan, N. Mann and D. Specht, "The National MAPPA Research: Proven Reoffending Report," April 2023. [Online]. Available: https://www.aru.ac.uk/policing-institute/research/national-mappa-research.

[7] "Proven reoffending statistics: January to March 2021," 26 January 2023. [Online]. Available: https://www.gov.uk/government/statistics/proven-reoffending-statistics-january-to-march-2021.

[8] M. Sapouna, C. Bisset, A.-M. Conlong and B. Matthews, "What Works to Reduce Reoffending: A Summary of the Evidence," 8 May 2015. [Online]. Available: https://www.gov.scot/publications/works-reduce-reoffending-summary-evidence/pages/4/.

[9] D. P. MacKinnon, J. L. Krull and C. M. Lockwood, "Equivalence of the Mediation, Confounding and Suppression Effect," 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819361/.

[10] G. Tripepi, K. J. Jager, F. W. Dekker and C. Zoccali, “Stratification for Confounding – Part 1: The Mantel-Haenszel Formula,” [Online]. Available: https://doi.org/10.1159/000319590.

[11] P. H. Lee, “Is a Cutoff of 10% Appropriate for the Change-in-Estimate Criterion,” 2013. [Online]. Available: https://doi.org/10.2188/jea.JE20130062.

[12] A. J. Fairchild and S. D. McQuillin, “Evaluating mediation and moderation effects in school psychology: A presentation of methods and review of current practice,” 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5488867/.

[13] N. Breslow and N. Day, “Statistical Methods in Cancer Research Volume I: The Analysis of Case-Control Studies,” 1980. [Online]. Available: https://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Statistical-Methods-In-Cancer-Research-Volume-I-The-Analysis-Of-Case-Control-Studies-1980.

[14] V. Bewick, L. Cheek and J. Ball, “Statistics review 8: Qualitative data – tests of association,” 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC420070/.

[15] K. H. Tilaki, “Methodological issues of confounding in analytical epidemiologic studies,” 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755849/.

[16] S. Rakes, S. G. Prost and S. J. Tripodi, “Recidivism among Older Adults: Correlates of Prison Re-entry,” 2018. [Online]. Available: https://www.cjcj.org/media/import/documents/recidivism_among_older_adults_correlates_of_prison_reentry.pdf.

[17] N. Greifer and E. A. Stuart, “Matching Methods for Confounder Adjustment: An Addition to the Epidemiologist’s Toolbox,” 2021. [Online]. Available: https://doi.org/10.1093/epirev/mxab003.

[18] P. C. Austin, “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies,” 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

[19] A. Burden, N. Roche, C. Miglio, E. V. Hillyer, D. S. Postma, R. M. Herings, J. A. Overbeek, J. M. Khalid, D. v. Eickels and D. B. Price, “An evaluation of exact matching and propensity score methods as applied in a comparative effectiveness study of inhaled corticosteroids in asthma,” 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5367458/.

[20] K. I. G. K. E. A. S. Daniel E. Ho, “MatchIt: Nonparametric Preprocessing for Parametric Causal Inference,” 2011. [Online]. Available: https://www.jstatsoft.org/v42/i08/.

[21] N. Greifer and E. A. Stuart, “Choosing the Causal Estimand for Propensity Score Analysis of Observational Studies,” 2021. [Online]. Available: https://arxiv.org/abs/2106.10577.

[22] K. Kainz, N. Greifer, A. Givens, K. Swietek, B. M. Lombardi, S. Zietz and J. L. Kohn, "Improving Causal Inference: Recommendations for Covariate Selection and Balance in Propensity Score Methods," 2017. [Online]. Available: https://www.journals.uchicago.edu/doi/full/10.1086/691464.

[23] P. C. Austin, P. Grootendorst and G. M. Anderson, "A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study," 2007. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16708349/.

[24] G. Simoneau, F. Pellegrini, T. P. Debray, J. Rouette, J. Muñoz, R. W. Platt, J. Petkau, J. Bohn, C. Shen, C. d. Moor and M. E. Karim, "Recommendations for the use of propensity score methods in multiple sclerosis research," 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9260471/.

[25] E. A. Stuart, "Matching Methods for Causal Inference: A Review and a Look Forward," 2010. [Online]. Available: https://doi.org/10.1214/09-STS313.

[26] W. Leite, "Practical Propensity Score Methods Using R," SAGE Publications, 2016, pp. 87-109.

[27] E. A. Stuart and K. M. Green, "Using Full Matching to Estimate Causal Effects in Nonexperimental Studies: Examining the Relationship Between Adolescent Marijuana Use and Adult Outcomes," 2008. [Online]. Available: https://doi.org/10.1037/0012-1649.44.2.395.

[28] Z. Zhang, H. J. Kim, G. Lonjon and Y. Zhu, "Balance diagnostics after propensity score matching," 2019. [Online]. Available: https://atm.amegroups.org/article/view/22865.

[29] N. Greifer, "Estimating Effects After Matching," 2023. [Online]. Available: https://cran.r-project.org/web/packages/MatchIt/vignettes/estimating-effects.html.

[30] A. Canty and B. Ripley, "Package 'boot'," 2022. [Online]. Available: https://cran.r-project.org/web/packages/boot/boot.pdf.

[31] C. Heinrich, A. Maffioli and G. Vázquez, "A Primer for Applying Propensity-Score Matching," 2010. [Online]. Available: https://publications.iadb.org/en/primer-applying-propensity-score-matching.

[32] T. J. VanderWeele and P. Ding, "Sensitivity Analysis in Observational Research: Introducing the E-Value," 2017. [Online]. Available: https://doi.org/10.7326/M16-2607.

[33] M. B. Mathur, P. Ding, C. A. Riddell and T. J. VanderWeele, "Website and R Package for Computing E-Values," [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29912013/.