A critical review of the BOSCAR report: *An evaluation of the Suspect Target Management Plan*

Ian Watson
Social Policy Research Centre
UNSW Sydney*

9 November 2020

Introduction

This paper is a critical review of a recent BOSCAR Bulletin, *An evaluation of the Suspect Target Management Plan* by Steve Yeong.[1] For ease of expression I will refer to it from now on as the *STMP Report*. My commentary covers the following issues:

- the nature of the STMP-II data;
- research design and the issue of causality;
- some technical weaknesses in the modeling.

I focus on both methodological issues as well as the interpretation of the findings, and my conclusion is that the *STMP Report* has serious weaknesses. I am particularly critical of the author's argument that his modeling shows that the SMPT-II has reduced criminality in NSW.[2] I conclude that a more accurate assessment of this study is that methodological weaknesses in the analysis have prevented any reasonable assessment being made regarding the outcomes of the STMP-II program.

Email: mail@ianwatson.com.au. *Website*: ianwatson.com.au

1. Steve Yeong (2020), *An evaluation of the Suspect Target Management Plan*, Crime and Justice Bulletin Number 233, Sydney NSW: NSW Bureau of Crime Statistics and Research

2. All of my comments refer only to the STMP-II data and analysis; I do not discuss the DV-STMP data or analysis.

The nature of the STMP-II data

Cameos and typical persons

Why does the STMP-II data matter? When it comes to interpreting the findings, the nature of the data affects the reader's perception of who the STMP-II is applied to. The *STMP Report* makes it very clear who the author thinks this is:

> By the time that the typical individual is placed on either form of STMP, he has almost 10 prior court appearances, half of which relate to the use of violence, one relating to the use of weapons and two relating to the use of drugs. He has also had a sentence of imprisonment and five community orders, all by age 26 …[3]

This is essentially a cameo drawn from the sample extracted by the author from the Reoffending Database (ROD). The terminology of 'typical' would lead the reader to assume that this cameo, while not constituting the majority of the STMP-II sample, is nevertheless reasonably common. Phrased in this way, the cameo presents a disturbing and threatening picture of 'criminality' in the community. Is it an accurate account of those people subject to the STMP-II program?

The core problem here is that the report provides no information on how many people subject to STMP-II actually fit this cameo. Fortunately, it is feasible to construct *synthetic data* based on the descriptive statistics (sample size, means and standard deviations) provided in Table 1 ('First day on STMP') in the *STMP Report* for the count variables in the sample.[4] I do not argue that this synthetic data is a reconstruction of the real data; rather I argue that the distribution of possible values in the synthetic data is close to those in the real data. We don't know, for example, how these variables combine at the unit level, that is, how many individuals have a certain combination of the characteristics represented by these variables. We cannot accurately estimate, therefore, how many people are likely to fit the cameo outlined above. (There is a method, however, for simulating a unit-record dataset from these synthetic data, and I will discuss that below.)

For the moment, it is worth asking whether basing a cameo on the sample means is appropriate? In the real data, all of the count variables have standard deviations that are large relative to their means. In the synthetic data, this gives rise to distributions such as those shown in Figure 1. Even

3. Yeong 2020, p. 6.

4. The synthetic data for the count variables (all of which are overdispersed) are simulated using R's `rnegbin` function with `n` equal to sample size, `mu` (μ) equal to the mean and `theta` equal to a dispersion parameter calculated as $(\mu + \mu^2)/sd^2$.

a cursory glance shows that these data are heavily right skewed (they are overdispersed count data) with the median lower than the mean in all cases. The percentages shown on the vertical axes are illuminating: the vast majority (over 80 per cent) of the synthetic sample have zero weapons offences and a clear majority (over 60 per cent) have zero drug offences and zero prison sentences. One conclusion that can be draw from these distributions is that for individual offences (or court appearances/sentences), the STMP-II data is best characterised as: *a large number of people have a small number of offences (or court appearance/prison sentences) and a very small number of people have a large number of offences (or court appearance/ prison sentences).*

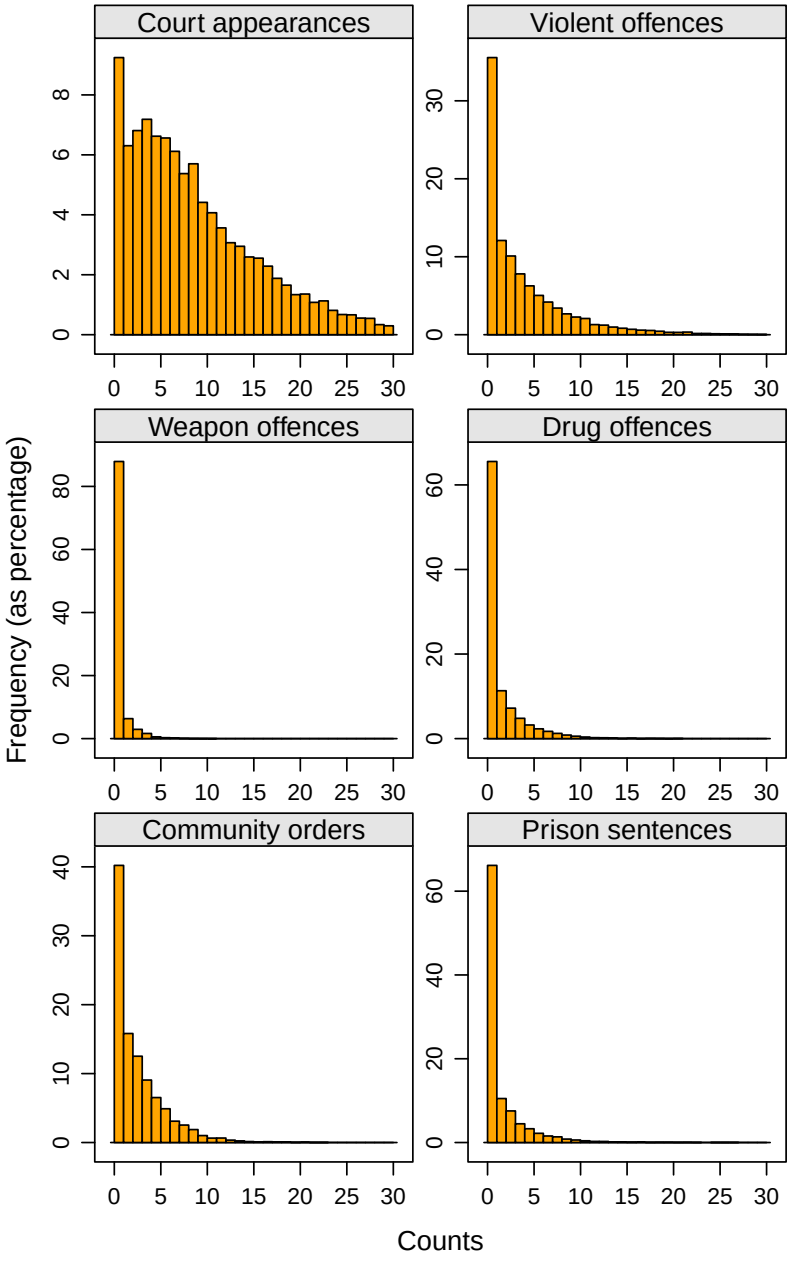

Figure 1: Distribution of count variables in synthetic data

In simulating this synthetic data—based on the descriptives shown in Table 1 of the *STMP Report*—there is little scope to change these distributions. It is certainly not possible to alter the skewed nature of the distributions. For example, 1000 iterations of the simulation for the court appearance variable consistently reproduces a strongly right skewed density. Furthermore, for these kinds of data, even if one manually alters the lower counts (replaces the 0s and 1s with 2s and 3s, for example) so as to shift the data further to the right, this increases the mean, but it reduces the standard deviation. If one attempts to retain the standard deviation by enlarging the range of high values, then this shifts the mean well above the value reported for the real data. In other words, for overdispersed count data like these there is little scope to maintain the location and scale of these variables (the mean and standard deviations) and yet to reshape this distribution away from a highly skewed shape.

The 'typical' person cameo mentioned above (referred to from now on as the '*stylised STMP-II cameo*') uses the *mean* rather than the *median*, an inappropriate measure with highly skewed data. As Table 1 shows, the means of these variables in both the original and the synthetic data differ from the median and the mode in the synthetic data. The mode is the measure which probably comes closest to the everyday notion of 'common' or 'typical' but the median is generally preferred as the most accurate reflection of the central tendency in data like these. In both cases, these figures are lower than the means, yet the means are used to construct the cameo discussed above. In other words, the severity of these interactions with the criminal justice system (CJS) is inflated in the *stylised STMP-II cameo*.[5]

Table 1: Measures of central tendency in original and synthetic data

	Original		*Synthetic data*			
Variables	*Mean*	*Rounded mean*	*Mean*	*Rounded mean*	*Median*	*Mode*
Court appearances	9.70	10	9.63	10	8	4
Violent offences	4.50	4	4.33	4	3	0
Weapon offences	0.54	1	0.52	1	0	0
Drug offences	1.66	2	1.64	2	1	0
Community orders	2.93	3	2.89	3	2	0
Prison sentences	1.64	2	1.64	2	1	0

Notes: Rounded original means are those used in the report's cameo. Except for the number of community orders and prison sentences. It is not clear where the cameo draws those figures from.

5. The cameo in the report appears to draw all its figures from Table 1 ('First day on STMP') so this has been the basis for the simulations. The number of community orders and prison sentences differ and it is unclear where these figures are drawn from.

These inflated counts are a problem, but a minor one. The more serious problem lies in the assumption in the *STMP Report* that it is reasonable to construct a typical individual from summary measures for the whole sample in an additive fashion. Constructing cameos may be intended as a device to make the descriptive statistics more vivid to a lay audience, but it can be a highly misleading device, particularly when the characteristics are combined in this additive way.

Is it possible to gain a more realistic sense of the prevalence of court appearances, criminal offences and prison sentences in the the STMP-II data rather than rely on this misleading cameo? I mentioned above that there is a method for simulating a synthetic unit record dataset from these synthetic data and, in so doing, estimate the size of the gap between what is most likely to be the case and what this *stylised STMP-II cameo* presents.

The simulation exercise proceeds as follows. Four cameos are constructed, the first of which matches the *stylised STMP-II cameo*. The other three are variations on this first one in which a more 'relaxed' definition of the combination of offences is constructed. I will say more about these shortly. The next stage of the exercise involves constructing four synthetic datasets. These reflect a number of different approaches to combining vectors (the variables) into matrices which reflect different combinations of characteristics at the unit record level. The first dataset—called the 'random dataset'—is based on repeatedly randomly shuffling the vectors so that different combinations emerge, and then counting the number of observations in the dataset for each of these four cameos. This simulation is repeated 10,000 times to produce a collection of counts, and the maximum number is then tabulated for each cameo. Why select the maximum counts? Basically, in order to favour an outcome similar to the *stylised STMP-II cameo* taking the maximum count across all 10,000 iterations makes it more likely that we will find people who combine these characteristics in a way which might approximate the *stylised STMP-II cameo*.

Of course, a random dataset like this ignores the likely correlation between these offences within individual observations. For example, individuals may be more likely to have prison sentences if they have committed violent offences. Three datasets are constructed which incorporate such correlations.[6] There is a 'low correlation' dataset in which we assume only weak correlations between all the variables. Another dataset is a 'high correlation' one, where strong correlations are assumed. Finally, a 'real

6. These datasets are constructed using copulas which preserve the marginal distributions of the variables whilst inducing correlations between them. The pairs plots in the Appendix illustrate the outcome.

world' dataset is constructed, in which correlations are differentiated in an attempt to match likely real-world conditions.[7] See the pairs plots in the Appendix for a visual representation of these four datasets.

The four cameos to which the counting exercise is applied are shown in Table 2. The first cameo matches the *stylised STMP-II cameo*. The second cameo relaxes the requirement for an exact match, by allowing any number of elements (ie. courts appearances, offences, prison etc) up to the numbers shown in the stylised cameo. The third cameo also relaxes the requirement for an exact match by allowing the elements to all be greater than those in the stylised cameo. Finally, the fourth cameo also relaxes the exact match by setting boundaries around the numbers, for example, between 8 and 12 court appearances. This last cameo is also notable in allowing for both weapons offences and prison sentences to vary from none at all (which is very common) through to two such outcomes. In other words, a range of variations on the original *stylised STMP-II cameo* are constructed in two of these cameos which favour higher counts than does this original; and one cameo (number three) which looks for 'dangerous' combinations of elements.

The results from this exercise are shown as percentages[8] in Figure 2. Despite the best efforts to match the *stylised STMP-II cameo*—and a series of alternatives—all these numbers fall way short of anything which could be regarded as 'typical'. The highest number of observations is 699 (or 7 per cent) is for cameo three—the 'dangerous' combination of characteristics—and this only applies to the dataset where all these variables are highly correlated. In other words, this number is partly an artefact of the dataset, since by construction it maximises such combinations.

In summary, despite relaxing the definition of the original *stylised STMP-II cameo* in a variety of ways, the largest proportion of people to whom it might apply is less than 7 per cent. Applying the definition as it originally appeared in the *STMP Report* sees virtually no-one fitting this cameo. By way of contrast, if one creates a cameo for someone with a handful of court appearances and just *one* other offence (or prison sentence or a few community orders), then one finds a match for 24 per cent of the 'real-world' synthetic dataset.[9] In other words, offenders with only

7. These draw on information from experts in criminology and from data in Patrizia Poletti et al. (2010), 'Common offences in the NSW higher courts', in: *Judicial Commission of NSW: Sentencing Trends & Issues*, URL: https://www.judcom.nsw.gov.au/wp-content/uploads/2016/07/sentencing_trends_41.pdf and Georgia Brignell et al. (2010), 'Common offences in the NSW local court', in: *Judicial Commission of NSW: Sentencing Trends & Issues*, URL: https://www.judcom.nsw.gov.au/wp-content/uploads/2016/08/sentencing_trends_40.pdf.

8. Since these are based on 10,000 observations, conversion into counts is simple: multiply the percentage shown by 100.

9. This cameo is not shown in Figure 2 or Table 2 but consists of 2380 observations. The

a single offence make up one quarter of this synthetic STMP-II dataset. Clearly, these simulations reinforce the view expressed earlier, that the people subject to STMP-II consist of *a small group of people with a large number (or range) of interactions with the criminal justice system (CJS) and a large group of people with a small number (or range) of interactions with the CJS*.

Table 2: Definitions of cameos

Category	*Definition*
Cameo 1	court==10 & viol==5 & weap==1 & drugs==2 & comm==5 & pris==1
Cameo 2	court %in% 1:10 & viol %in% 1:5 & weap==1 & drugs %in% 1:2 & comm %in% 1:5 & pris==1
Cameo 3	court > 10 & viol > 5 & weap > 1 & drugs > 2 & comm > 5 & pris > 1
Cameo 4	court %in% 8:12 & viol %in% 3:7 & weap %in% 0:2 & drugs %in% 0:3 & comm %in% 3:7 & pris %in% 0:2

Notes: Abbreviations: == equal to; %in% in the range; 1:10 1 to 10; > greater than. Note that Cameo 1 matches the *stylised STMP-II cameo*.

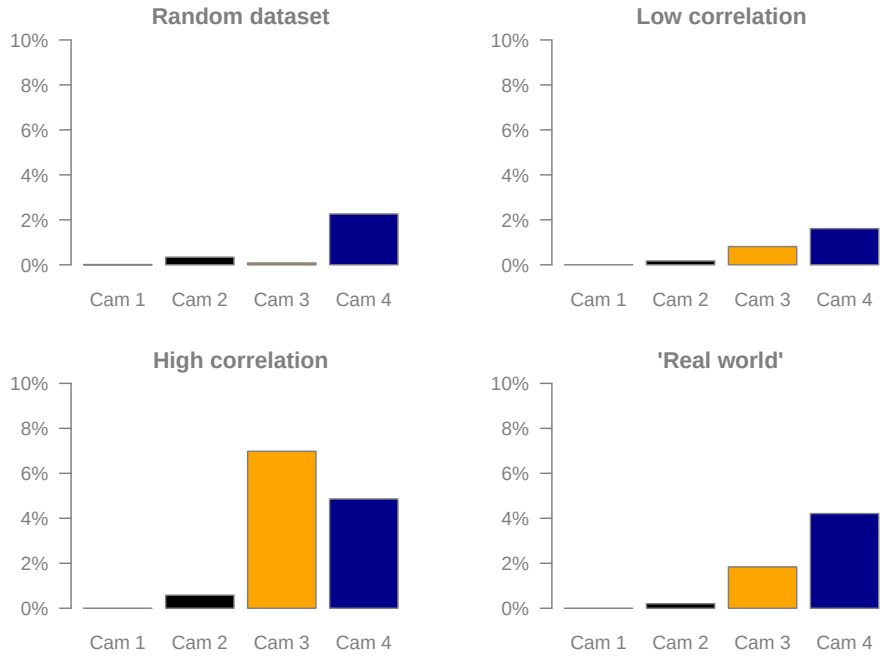

Figure 2: Percentage of observations in each dataset which match cameos
Definitions of cameos are shown in Table 2

The report argues that the descriptive statistics in Table 1 show that 'the police are identifying high-risk individuals for STMP'.[10] As suggested earlier, this claim mistakenly extrapolates from overall sample averages to construct a 'typical' individual, and this is then used to argue that a large number of people have been legitimately placed on STMP-II. Clearly, the synthetic data suggests such a claim is completely unwarranted and that the

'handful' of court appearances are for 1 to 4 and the 'few' community orders are for 1 to 3.
10. Yeong 2020, p. 6.

STMP-II program is wide-ranging in its application rather than precisely targeted.

The onus lies with the author of the *STMP Report* to refute this conclusion by *using the real data* to generate counts for the cameo in that report and to display the distributions shown in the real data along the lines of Figure 1 above. In other words, it rests with the author of the *STMP Report* to show that the cameo of the typical person placed on the STMP constitutes more than a small handful of people. Otherwise, one can only conclude that the *stylised STMP-II cameo* is a complete fiction.

Research design and causality

The *STMP Report* is located within the treatment effects tradition, in which a a treatment group (participants in a program) is exposed to a treatment to which a control group (non-participants) is not exposed, and one then compares outcomes across the two groups. While this approach can be applied reasonably well within an experimental setting, for observational data this approach can be fraught with difficulties, particularly when regression modeling is solely relied upon for establishing causality.[11] The usual procedure is to include a dummy variable (treated or not treated) and test whether it has a significant association with the outcome (such as offending). A range of confounding variables are also included in order to isolate the 'effect' of treatment on participants.

Research design

Counterfactuals are fundamental to assessing treatment effects within observational studies. They address the obvious question: what would have happened in the absence of treatment? It is the counterfactual which confers 'causality' on the research findings.[12] For a counterfactual to have validity the control group must be comparable on a range of variables, with the only notable difference being exposure to the treatment. The author of the *STMP Report* recognises at the outset selection bias makes it difficult to construct a valid comparison, because 'individuals on STMP are likely to be at a higher risk of offending, irrespective of whether STMP has any impact on offending'.[13] The *STMP Report* takes two approaches to this

11. Paul R. Rosenbaum (2002), *Observational Studies*, New York: Springer.
12. As the Neyman-Rubin causal model puts it: 'A causal effect is defined as the difference between an observed outcomes and its counterfactual.'Alexis Diamond and Jasjeet S. Sekhon (2013), 'Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies', in: *Review of Economics and Statistics* Vol. 95. No. 3, pp. 932–945, URL: http://sekhon.berkeley.edu/papers/GenMatch.pdf, p. 4
13. Yeong 2020, p. 7.

problem and while findings for each are presented in the report, the second approach is relegated to the appendix. I will return to this issue below.

In the first approach the study does not employ a 'conventional' control group. Rather the research design involves a time-shift strategy in order to create a control group. This group consists of another group of individuals who will be *subsequently* placed on the STMP-II but who, during the 'observation period', are not yet subject to that program.[14] However, because the treatment variable is causally dependent on the dependent variable, pre-STMP individuals cannot serve as a control group for a post-STMP target group. Because of the time-shift imposed by the study design, the number of court appearances and offences etc are *rising* among the control group while they are *falling* among the treatment group. This artefact of the study design will inevitably bias the regression towards finding a larger gap between the two groups than might otherwise be the case. This problem is not a minor one but is inherent to the research design because of the construction of the control group. Since prior offending is listed as a trigger for an STMP in the report, this study only shows that offences cause STMPs.

The author is aware of this problem. The dummy variable in the regression modeling which represents placement on STMP-II is identified through variation in the timing of when individuals become subject to SMTP-II. To interpret this dummy variable as causal relies on the risk of offending being unconditionally related to the timing, something which is not the case with these data.[15] As the author concedes, the risk of offending by individuals is not time-invariant, but rather appears to be relate to when such individuals are placed on STMP-II. The author of the *STMP Report* acknowledges this weakness in the research design:

> If [STMP] were to have a causal interpretation, we would expect so see no trend in offending prior to STMP, followed by a sharp (downward) trend after placement on STMP … However, from Figures 1a and 1b we can see sizable upward trends in the year leading up to STMP, followed by sharp downward trends immediately after being placed on STMP.[16]

The second approach—the one relegated to the appendix—entails using a matching estimators strategy to explicitly create a control group who are *not* subject to the STMP-II. These consist of a group drawn from the Reoffending Database (ROD) but who have *not been placed on STMP-II at all* (as opposed to a group placed on STMP-II in the next time period as

14. That is, the time shift involves multiple periods with pre-treatment and post-treatment groups aligned.
15. Yeong 2020, p. 8.
16. Ibid., p. 8.

happens in the time-shift strategy). Research designs based on matching estimators are well established in the literature, and as long as the researcher achieves good balance on the covariates between treatment and control groups, then regression modeling may proceed with reasonable confidence.

In implementing the matching estimators, the study combined Coarsened Exact Matching (CEM) and Propensity Score Matching (PSM), but the author was dissatisfied with the matching results: 'the matched groups are not statistically or practically equivalent to their respective treatment groups'.[17] It is not clear what 'practically equivalent' means. Perhaps the author is referring to an earlier footnote where he observed:

> Interestingly, I was not able to find a credible match for individuals subject to STMP using the entire Reoffending Database (which contains information for every person charged by the NSW Police Force since 1996). This suggests that the people the police select for STMP are truly distinct from other offenders they interact with.[18]

However, it is also likely that the matching strategy employed by the author was inadequate. I return to this issue below. The lack of a 'statistically equivalent' match is not explained in any detail. The descriptive comparison of treatment and control groups in Table A1 of the appendix does not provide compelling evidence that the two groups are not reasonably comparable. The means are shown to three decimal points, whereas if they were shown to one decimal point, the impression of how well they matched might be quite different. For example, age differences (26.7 to 26.3) equate to a few months apart, and differences for prior court appearances (10.6 to 10.1), prior prison sentences (1.8 to 1.7) and prior community orders (3.2 to 3.1) are all fairly minor. Moreover, with over 9,000 observations in each group, minor differences are almost bound to be 'statistically significant'. What matters with the matching estimator approach is whether the differences between a treatment group and a control group substantively shrink during the matching process such that one is ultimately comparing 'like with like' across a large majority of the variables employed.

It is more likely that the author's matching strategy has let him down. It is well known that propensity score approaches can worsen the matching outcome, and recent literature has reiterated this criticism.[19] Far better matching estimators are available which the author might have employed, such as 'genetic matching', an approach which invariably improves

17. Yeong 2020, p. 23.
18. Ibid., fn. 21, p. 7.
19. See, for example, Gary King and Richard Nielsen (2019), 'Why Propensity Scores Should Not Be Used for Matching', in: *Political Analysis* Vol. 27. No. 4, pp. 1–20

on propensity score outcomes.[20] In other words, instead of giving up on the matching estimator approach—and relegating the findings to the appendix—the author should have persisted with this strategy.

What is particularly disturbing is that the regressions fit to these matched data provided results opposite to those in the main body of the report. It showed higher offending among the STMP-II group compared with their counterparts. Rather than view these results as casting doubt on the main findings in the report, the author speculated that the weaknesses in the matching process may be the reason: 'One explanation for this finding is that there is some form of unobserved heterogeneity that matching cannot address'.[21] While it can be difficult to make matching estimators work well, the finding here is *not* that there is *no difference* between the groups, but that the results are the *reverse* of the findings in the main report. This anomaly surely warranted further investigation rather than a curt dismissal of the matching estimator procedure, particularly when better approaches were available.

Causality

The interpretation of the study's finding is one of the most worrying aspects of the *STMP Report*. Having concluded that he had failed in his efforts to construct a valid counterfactual, the author concluded: 'my estimates do not have a causal interpretation. Instead, they must be interpreted as the association between STMP and offending'.[22] However, a second conclusion immediately contradicted this:

> And second, this would suggest that the police are both correctly identifying individuals at a high risk of offending for STMP, and that once placed on STMP, an individual's risk of offending drops dramatically.[23]

The wording of this last sentence is clearly a causal one. This is not an isolated lapse in expression. The author repeats the caveat about causality in the discussion section of the report ('the estimates do not have a causal interpretation') but again negates this by discussing the possible direction

20. See, for example, Jasjeet S. Sekhon (2011), 'Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R', in: *Journal of Statistical Software* Vol. 42. No. 7, pp. 1–52, URL: http://www.jstatsoft.org/v42/i07/. A recent Productivity Commission report on the youth labour market made extensive use of this approach. See Catherine de Fontenay et al. (2020), *Climbing the jobs ladder slower: Young people in a weak labour market*, Staff Working Paper, July, Productivity Commission.
21. Yeong 2020, p. 24.
22. Ibid., p. 8.
23. Ibid., p. 8.

of bias in these estimates by referring to the 'true crime reduction benefit associated with STMP'. He concludes the paragraph with 'It is, therefore, likely that STMP-II is reducing other types of crime in addition to those examined in this paper'.[24] In the report's overview, the Results section is careful to stick with the language of 'association' but the Conclusion section immediately overturns this: 'Both STMP-II and DV-STMP are effective in reducing crime. Both programs predominately reduce crime through deterrence'.[25] Clearly, causal language is endemic to the author's interpretation of his results.

Even were the author to avoid the language of causality, and stay strictly with the language of 'association', the conclusion that the STMP-II had a positive and sizable association with a reduction crime is completely unfounded. There are several reasons for this:

- the strength of any association is indeterminate because of the bias in the time-shift comparison between treatment and control group (as outlined above);
- the association is the opposite in the matching estimators approach, and no serious engagement with these results is offered;
- the specific regression findings are quite diverse, but a *single* policy conclusion is drawn.

This last point is an important one. Using the time-shift strategy, the author finds mixed results. The associations between STMP-II and subsequent offending are:

- negative for property crime by the whole sample;
- indeterminate for violent crime by the whole sample;
- positive for imprisonment by the whole sample;
- negative for both violent and property crime for juveniles;
- positive for violent crime for Aboriginal participants;
- negative for property crime for Aboriginal participants;

When it came to the matching estimators strategy, as just noted, the associations were positive for violent and property crime (combined). In other words, negative associations (that is, a 'reduction' in crime) was far from universal, yet the report's main conclusions ignore this unevenness in the results and assert confidently that STMP-II is 'effective in reducing

24. Yeong 2020, p. 17. This constant lapsing into causal language is found throughout the report, sometimes in the context of discussing technical points. For example, the author suggests on page 8 that the estimates in the study might be conservative and underestimate 'the true impact of the STMP's crime reduction benefit.' 'Reduction' is clearly a causal term. The author then proceeds to ask: 'If STMP is generating a reduction in crime, the question is how?' The terminology of 'generating' is thoroughly causal.
25. Ibid., p. 1.

crime'. In other words, the policy implications of the report amount to an endorsement of the STMP-II, yet the regression modeling fails to support this blanket conclusion. The most accurate conclusion to this report would be: *methodological weaknesses in the analysis have prevented any reasonable assessment being made regarding the outcomes of the STMP-II program*.

It is worth noting that one of the key insights in the author's overview of the literature is that overseas programs which target subpopulations in an effort to reduce crime often supplement the policing strategy with increased social support (eg. housing, education, employment) for those subpopulations. The author cites the example of the community Initiative to Reduce Violence (CIRV) in Glasgow as one successful program that has been 'rigorously evaluated'.[26] However, towards the end of the *STMP Report* when discussing the importance of his findings, the author ignores this insight and argues:

> The first is to illustrate that offender-focused policing programs work in Australia. This is an interesting finding given that STMP differs markedly from most focused deterrence programs overseas. Focused deterrence programs typically involve working with community organisations to communicate an explicit message of deterrence. Focused deterrence programs also generally involve increasing access to social services as an adjunct to intensive policing.

In claiming that his study has shown that the STMP-II has 'worked' and that it 'caused' crime to fall the author dispenses with the relevance or need for social support in addition to policing activity. His report can be seen as an endorsement of a policing-only approach, even though it is clear that the study has not established such a causal link.

Technical weaknesses

Model fit

How well do these models fit the data? The author offers very little information on model diagnostics. The adjusted R-squared figures are nearly all below 0.1, which means that some 90 percent of the variability in the outcome is not accounted for by the predictors used in these models. A great deal else is going on in these data that is not captured well in this modeling.[27]

Other measures of fit, in particular, predictive adequacy (for example, cross-validation) are not canvassed. To some extent, the author addresses

26. Ibid., p. 4.

27. It is interesting to note that the adjust R-squared figures are higher for the regressions of the matched estimators approach compared with the regressions in time-shift approach.

this in the footnote which contrasts the objectives of prediction versus causal inference.[28] His objective is the latter and this focus can be used to justify a lack of concern with the predictive accuracy of the modeling. However this does not mean that a poor fitting model is acceptable. As Hilbe cautions:

> The problem is that predictor p-values may all be under 0.05, or may even all be displayed as 0.000, and yet the model can nevertheless be inappropriate for the data. A model that has not undergone an analysis of fit is, statistically speaking, useless.[29]

Heterogeneity

The concept of heterogeneity—diversity—is an important one in statistics and its relevance has been increasing in recent years.[30] When it comes to the treatment effects literature—the field in which the *STMP Report* can be located—there is an increasing recognition that the average treatment effect (ATE) of some intervention on the treated is not necessarily very useful. A more interesting question is: for whom did the treatment work? and for whom didn't it work? And why?

This suggests that focusing on heterogeneity should be a major focus whenever the subjects in a dataset show diversity. It is clear from the discussion above that the people subject to STMP-II are indeed quite heterogeneous. The *STMP Report* report does acknowledge heterogeneity and it does this by running separate regressions for young people and for Aboriginal people and separate regressions for a number of cohorts (based on the duration of their sentences). Unfortunately, the author's implementation of this may be unsound: he compares coefficients from separate regressions, a procedure whose validity depends on assumptions about the sample variances. A more rigorous way to deal with heterogeneity is to fit a *single* model and use either interaction terms or a multilevel model. In this way, one can answer the question: how does the relationship between outcomes and predictors vary across subgroups? In so doing, it is legitimate to make direct comparisons because all the coefficients (or predictions) come from the same model.

Another source of heterogeneity in this study are the Police Area Commands (PAC). Not only does selection into the STMP-II depend on

28. Yeong 2020, p. 10, fn 25.
29. Joseph M. Hilbe (2011), *Negative Binomial Regression*, Second edition, Cambridge: Cambridge University Press, p. 64.
30. See the emphasis on moving beyond averages in the field of quantile regression (Roger Koenker (2005), *Quantile Regression*, New York: Cambridge University Press) or the emphasis on multilevel models for investigating heterogeneity (Gelman and Hill 2007.

decisions made at the PAC level, but the subsequent interactions between the PAC and these people would appear to be quite fundamental. The modeling in the *STMP Report* regards the PAC as a 'control', specifically as a fixed effect. But is this an adequate way to deal with such heterogeneity? The characteristics of the PACs are extremely diverse, given their geographical basis. The *STMP Report* certainly recognises that heterogeneity arises from 'PAC-specific considerations such as their priority crimes, annual budgeting allocations, variation in the application of STMP-II, local labour market conditions and the demographic characteristics of civilians living within the jurisdiction of each PAC'.

The author deals with this by including the PAC as a fixed effect, a statistical device for adjusting for this variability in so far as the outcome is concerned. For example, do the characteristics of the PAC relate to the outcomes such as committing a violent or property crime or being imprisoned (see Table 2 in the *STMP Report*). But the variability in the predictors are ignored with this approach. Fixed effects cannot answer questions such as: how do the variability in age, Aboriginality, cautions, court appearances, PAC and so forth *interrelate*? How do these different covariates operate for different subgroups within the model? In other words, many of the various subgroup effects for the key predictors are not canvassed in these regressions.

To achieve this one needs interactions in a model. However, introducing the PACs as fixed effect interaction terms in not feasible—given how many PACs there are—so the obvious solution is a multilevel model in which the PAC is a grouping term (or level). From the model equation ($y_{ipt} = \beta_0 + \beta_1 Post_{ipt} + \gamma X'_{it} + \lambda_{pt} + u_{ipt}$) and the accompanying description, it is clear that the data are already indexed by PAC, so using multilevel models to accommodate this hierarchical structure is completely feasible. It is also evident that there is clustering in the sample: observations drawn from the same PAC in the sample will have greater similarity to each other than to those in other PACs. This can violate the regression assumption regarding independent error terms. The *STMP Report* acknowledges the clustering for the PAC variables and presents robust standard errors to deal with this. This approach, while adjusting the naïve standard errors, leaves the coefficient estimates unchanged. By contrast, multilevel models not only adjust the standard errors, but also improve the accuracy of the coefficient estimates.[31] In other words, not only would multilevel modeling

31. The increased accuracy comes from the 'partial pooling' which multilevel models employ. By contrast, the classical regression model, as employed in the *STMP Report* is essentially a 'complete pooling' model. For further elaboration on this distinction see ibid. One view of a multilevel model is that it operates as a 'giant interaction machine' (Richard McElreath (2020), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Second Edition, Boca Raton: CRC Press, Taylor & Francis Group.

offer insights into the heterogeneity in these data, but such an approach would provide better estimates.

As mentioned earlier, the author's main approach to this heterogeneity is separate regressions for subgroups. Yet when different model results are found for one of these subgroups—Aboriginal people—-the author gains little insight from his modeling and instead resorts to speculation which has no grounding in the data itself: 'Aboriginal people may react negatively to STMP-II interactions with police which results in increased offending.'[32]

In summary, there is insufficient material in the *STMP Report* to assess the adequacy of the modeling. The appendix is more of a supplement than a compendium of detailed model results and it is unclear what diagnostics the author used to assess the models. While his use of classical regression models (OLS) is standard practice in econometrics, among statisticians there is an increasing recognition that better model estimates come from using multilevel models.

References

Brignell, Georgia, Zeinab Baghizadeh, and Patrizia Poletti (2010), 'Common offences in the NSW local court', in: *Judicial Commission of NSW: Sentencing Trends & Issues*, URL: https://www.judcom.nsw.gov.au/wp-content/uploads/2016/08/sentencing_trends_40.pdf.

Diamond, Alexis and Jasjeet S. Sekhon (2013), 'Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies', in: *Review of Economics and Statistics* Vol. 95. No. 3, pp. 932–945, URL: http://sekhon.berkeley.edu/papers/GenMatch.pdf.

Fontenay, Catherine de, Bryn Lampe, Jessica Nugent, and Patrick Jomin (2020), *Climbing the jobs ladder slower: Young people in a weak labour market*, Staff Working Paper, July, Productivity Commission.

Gelman, Andrew and Jennifer Hill (2007), *Data analysis using regression and multilevel / hierarchical models*, Cambridge: Cambridge University Press.

Hilbe, Joseph M. (2011), *Negative Binomial Regression*, Second edition, Cambridge: Cambridge University Press.

King, Gary and Richard Nielsen (2019), 'Why Propensity Scores Should Not Be Used for Matching', in: *Political Analysis* Vol. 27. No. 4, pp. 1–20.

32. Yeong 2020, p. 17.

Koenker, Roger (2005), *Quantile Regression*, New York: Cambridge University Press.

McElreath, Richard (2020), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Second Edition, Boca Raton: CRC Press, Taylor & Francis Group.

Poletti, Patrizia, Zeinab Baghizadeh, and Pierrette Mizzi (2010), 'Common offences in the NSW higher courts', in: *Judicial Commission of NSW: Sentencing Trends & Issues*, URL: https://www.judcom.nsw.gov.au/wp-content/uploads/2016/07/sentencing_trends_41.pdf.

Rosenbaum, Paul R. (2002), *Observational Studies*, New York: Springer.

Sekhon, Jasjeet S. (2011), 'Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R', in: *Journal of Statistical Software* Vol. 42. No. 7, pp. 1–52, URL: http://www.jstatsoft.org/v42/i07/.

Yeong, Steve (2020), *An evaluation of the Suspect Target Management Plan*, Crime and Justice Bulletin Number 233, Sydney NSW: NSW Bureau of Crime Statistics and Research.

Appendix

The figures on the following pages show pairs plots for the synthetic datasets. Correlations are shown numerically in the upper triangles and visually as regression lines fit to scatter points in the lower triangles. The distributions of each variable are shown along the diagonal.

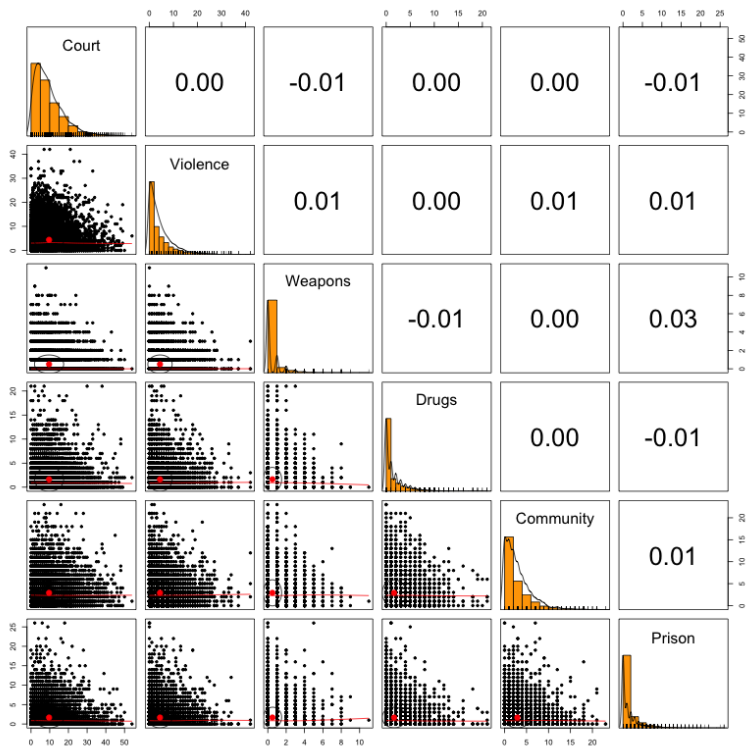

Figure 3: Pair plots of random dataset

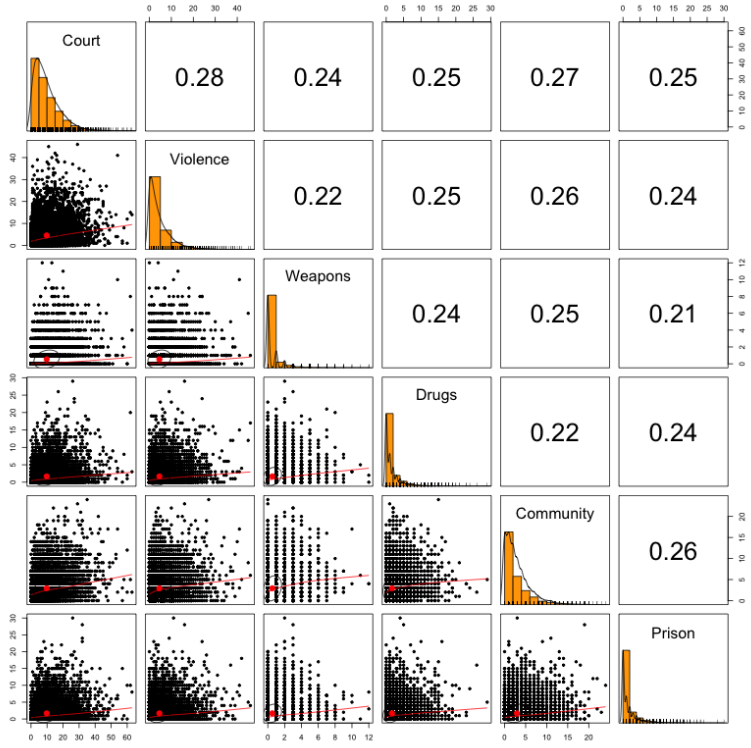

Figure 4: Pair plots of low correlation dataset

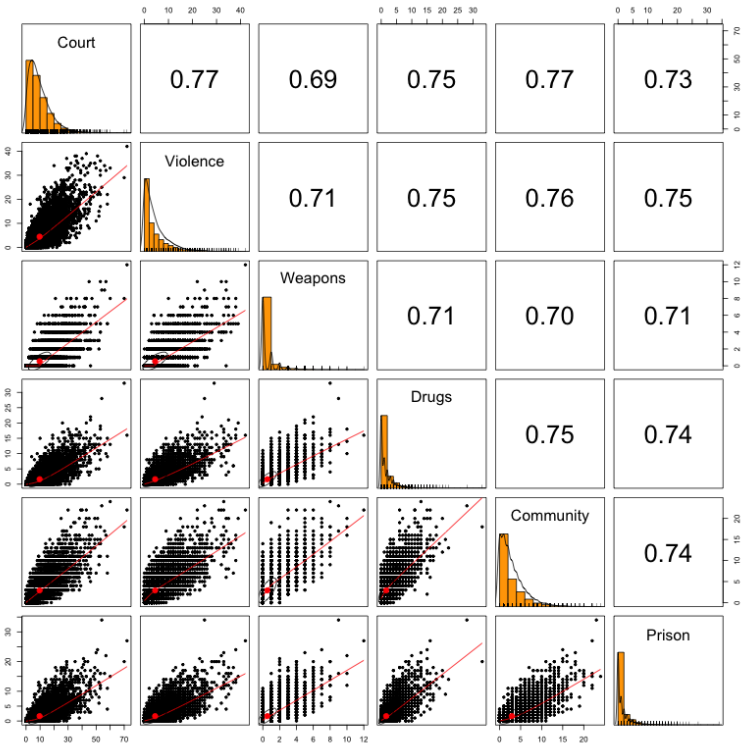

Figure 5: Pair plots of high correlation dataset

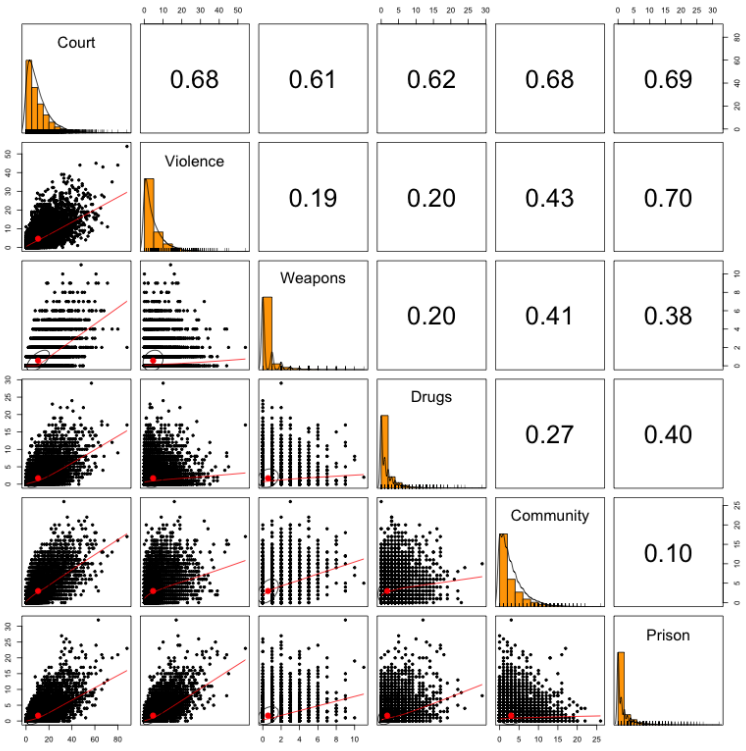

Figure 6: Pair plots of 'real-world' dataset