Linking serial sex offences using standard, iterative, and multiple classification trees

Bennell, Craig; Mugford, Rebecca ; Woodhams, Jessica; Beauregard, E; Blaskovits, Brittany

DOI:
10.1007/s11896-021-09483-6

Linking Serial Sex Offences Using Standard, Iterative, and Multiple Classification Trees

Craig Bennell[1,‡], Rebecca Mugford[1], Jessica Woodhams[2], Eric Beauregard[3], and Brittany Blaskovits[1]

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: *https://doi.org/10.1007/s11896-021-09483-6*

[1] Department of Psychology, Carleton University, Ottawa, Ontario, Canada
[2] School of Psychology, University of Birmingham, Birmingham, United Kingdom
[3] School of Criminology, Simon Fraser University, Burnaby, British Columbia, Canada
‡Correspondence to: Craig Bennell, Department of Psychology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6; (613) 520-2600 ext. 1769; craig.bennell@carleton.ca

Abstract

Studies have shown that it is possible to link serial crimes in an accurate fashion based on the statistical analysis of crime scene information. Logistic regression (LR) is one of the most common statistical methods in use and yields relatively accurate linking decisions. However, some research suggests there may be added value in using classification tree (CT) analysis to discriminate between offences committed by the same vs. different offenders. This study explored how three variations of CT analysis can be applied to the crime linkage task. Drawing on a sample of serial sexual assaults from Quebec, Canada, we examine the predictive accuracy of standard, iterative, and multiple CTs, and we contrast the results with LR analysis. Our results revealed that all statistical approaches achieved relatively high (and similar) levels of predictive accuracy, but CTs produce idiographic linking strategies that may be more appealing to practitioners. Future research will need to examine if and how these CTs can be useful as decision aides in operational settings.

Keywords: crime linkage; comparative case analysis; classification trees; serial crime; sexual assaults

Linking Serial Sex Offences Using Standard, Iterative, and Multiple Classification Trees

Studies over the last two decades have shown that serial offenders exhibit a reasonable degree of behavioural similarity and distinctiveness across their crimes, which makes it possible to link serial offences in a reasonably accurate fashion. Indeed, in a review of 19 studies that used receiver operating characteristic (ROC) analysis to assess the accuracy of statistical linking methods, Bennell et al. (2014) found accuracy levels (defined as the area under the ROC curve, or AUC) ranging from .45 (just below chance level accuracy) to .96 (near perfect accuracy). Using pre-established guidelines for interpreting AUCs (Swets, 1988), 36% of the AUCs included in the review were in the low range (.50-.70), 49% were in the moderate range (.70-.90), and 14% were in the high range (.90-1.00).

One of the most common statistical methods examined in crime linkage studies is logistic regression (LR) analysis (Bennell et al., 2014). This method is used to determine the linkage status of crime pairs (whether they were committed by the same vs. different offenders) based on the degree of behavioural similarity exhibited across the crimes (Bennell & Canter, 2002). In a typical study, across-crime similarity is assessed for behaviours falling into different domains (e.g., sexual, control, and escape behaviours). LR analysis is then used to assess the degree to which similarity scores from each domain can discriminate between crime pairs committed by the same vs. different offenders. When similarity scores from actual crime pairs are entered into the resulting LR equation, the output is a probability that the crimes have been committed by the same offender. To

make binary, forced-choice linkage decisions,[4] any crime pair associated with a probability value above a pre-specified threshold is predicted to be linked.[5]

While LR analysis is often associated with moderate to high AUCs when used for this purpose (Bennell et al., 2014), it typically results in nomothetic linking strategies (i.e., the same predictor variables are applied the same way in every case) that will miss potential links (Tonkin et al., 2012b). For example, a LR model might always result in a prediction that two crimes are linked when the distance between those crimes is small (e.g., Bennell & Canter, 2002), whereas practitioners will know, and research has consistently confirmed, that some crimes committed far apart are the work of the same offender (e.g., Lundrigan et al., 2010). Concerns regarding this "one-size-fits-all" approach has prompted research into other statistical crime linkage methods that might be more useful for capturing the heterogeneity of serial offender behaviour (Tonkin et al., 2012b).

For example, Tonkin et al. (2012b) compared linking models produced through LR analysis and classification tree (CT) analysis. As Tonkin and his colleagues discuss, a CT consists of a structured set of questions (related to specific predictor variables) that can be used by practitioners to systemically decide whether crime pairs have been committed by the same offender (e.g., Is the distance between these two crimes > or < 2.4 km?). Questions in the tree are organized hierarchically based on their level of predictive power. The first question encountered in the tree is asked of all crime pairs; subsequent questions differ depending on the preceding answer, which gives CTs their idiographic

[4] Other, non-binary types of linkage decisions can be made (e.g., rank-ordering crime pairs based on probability values), but we will focus in this paper on binary, forced-choice linkage decisions.
[5] A range of empirical methods are available to determine what threshold should be used for this purpose. See Swets et al. (2000) for a discussion of these methods.

quality (i.e., different questions, or predictor variables, can be applied to different crime pairs). Questions are asked until a linkage decision has been made for every crime pair.

Drawing on samples of residential burglaries from Finland and car thefts from the UK, Tonkin et al. (2012b) found a statistically comparable level of predictive accuracy for their LR and CT models across both crime types. However, while both statistical approaches considered the same predictor variables for the residential burglary data, the CT resulted in three distinct pathways to a linked decision, making it slightly more idiographic in nature. More specifically, linked decisions could be made by relying on: (1) relatively short distances between crime pairs, (2) very short distances between crime pairs in combination with a high degree of similarity in entry behaviours, or (3) moderate distances between crime pairs in combination with a high degree of similarity in internal behaviours. Of note, the residential burglary CT did not generalize well to a hold-out sample (this was not the case for car theft). This suggests that overfitting might be a potential problem for CT analysis, but not LR analysis.[6]

More recently, a large-scale study examining over 3,000 sexual assaults committed in five countries was conducted (Tonkin et al., 2017). Both LR and iterative classification tree (ICT) analysis (i.e., a more complex version of standard CTs; described in more detail below) were compared, along with Bayesian methods.[7] All approaches achieved moderate to high levels of discrimination accuracy. The ICT and LR performed at comparable levels, and the Bayesian analysis achieved the highest level of accuracy.

Refinements to Standard Classification Trees

[6] Overfitting occurs when a statistical model applies well to the data used to construct the model, however, it is a poor fit for data that was not used to construct the model (Wang et al., 2010).
[7] Note that Tonkin et al. (2012b) also attempted to examine ICTs, but their data did not allow such a model to be developed.

Our study contributes to existing research by comparing several variations CT analysis, some of which have been shown to outperform LR analysis in other domains (Monahan et al., 2001). In this section, we describe in detail what these variations are. However, before doing this, it is necessary for readers to have a more detailed understanding of how CT analysis works.

Technically, the structure of a CT is determined through the use of chi-square tests of independence. As shown in Figure 1, a CT begins with all crime pairs, regardless of linkage status, being contained in the 'root node.' The CT algorithm then runs a chi-square analysis for all predictor variables and the dichotomous outcome and selects the most significant predictor variable to split the data into sub-groups (sub-groups that best distinguish crime pairs that have been committed by the same vs. different offenders). Each resulting sub-group is comprised of crime pairs that fall into the different categories of this predictor variable. These sub-groups are referred to as 'child nodes' because they arise from earlier nodes, called 'parent nodes.' This splitting process continues until the crime pairs can no longer be distinguished from one another on the basis of the predictor variables. If no additional predictors can differentiate between crime pairs in a given child node, the splitting process ends; that child node is then labelled a 'terminal node'.

[Insert Figure 1 here]

Analogous to the process of selecting and applying a probability threshold to make linking decisions based on a LR model, some classification criteria must be established and applied to the terminal nodes in a CT to make a 'linked' versus 'unlinked' decision. In other domains, it is common practice to make these decisions in reference to the base rate of the target decision outcome (e.g., violence in the community in risk

assessment contexts; Monahan et al., 2001). In crime linkage research, this would be equivalent to the base rate of crime pairs in the entire sample that have been committed by the same offender. For example, a particular terminal node could be designated as representing the target decision outcome (i.e., a node categorised as containing linked crime pairs) if the proportion of crime pairs comprising that node is more than twice the overall base rate.

Much like Tonkin et al.'s (2012b) study, studies from other domains that have compared LR analysis to standard CT analysis have reported similar levels of accuracy for the two procedures. For example, in the McArthur Risk Assessment Study, Monahan et al. (2001) found comparable levels of predictive accuracy when using these two procedures to make risk assessment decisions for individuals discharged from mental health hospitals in the US. Drawing on 134 potential risk factors for violence in the community post-release, their LR and CT models were associated with AUCs of .81 and .79, respectively. However, hypothesizing that predictions based on the CT approach could be enhanced, researchers involved in the MacArthur Study introduced three novel methodological refinements to CT modelling, each of which will be examined in the current study: (1) adopting two decision thresholds, (2) iterating the standard CT, and (3) constructing multiple iterative CT models.

Adopting Two Decision Thresholds

In the MacArthur Study, researchers acknowledged that risk assessment predictions (or any predictions for that matter) are not necessarily as simple as a single decision threshold approach implies (violent/not violent or linked/unlinked). In fact, there are likely to be cases that one cannot unequivocally classify as high or low risk because

the risk of violence displayed by these ambiguous cases cannot be differentiated from the base rate of violence using the sequence of predictors initially identified by the CT (Monahan et al., 2001). For this reason, Monahan and his colleagues opted to use the violence base rate to construct two decision thresholds. In addition to labelling high-risk groups as those exhibiting twice the violence base rate, they also used a second threshold identifying the low-risk groups as those exhibiting less than half the violence base rate. Using this two-threshold approach, they found that their LR model could classify 57.1% of cases into either the high- or low-risk group, whereas their CT could classify 50.8% of the cases into either group. Thus, approximately half of the sample remained unclassified because they represented an indistinguishable level of violence risk when using either of the statistical approaches.

The same argument can be made concerning the crime linkage task. Although a certain prediction model may be successful at classifying some crime pairs, for other crime pairs, the level of behavioural similarity observed across the crimes may not be extreme enough (in the direction of either high or low similarity) for accurate 'linked' or 'unlinked' decisions to be made (using the initial set of predictor variables). When dealing with these ambiguous cases, it may be desirable to leave the cases unclassified. It may also be possible to subject these cases to further CT analysis to determine if any of them can be re-classified. By doing this, iterative CTs (ICTs) can be developed.

Iterating Standard Classification Trees

Although relatively good levels of predictive accuracy could be achieved using a standard CT, researchers involved in the MacArthur Study were concerned with the fact that only half of all cases could be classified into a high- and low-risk group by applying

the two decision thresholds to the CT (Monahan et al., 2001). To resolve this issue, they devised an extension to the standard CT known as ICTs. The ICT approach involves the re-analysis of cases deemed unclassifiable by the standard CT when the two-threshold approach is used. What resulted from this second analysis was a CT similar in structure to the first one but containing different combinations of predictor variables to differentiate between the included cases. This process can be repeated until the unclassified cases can no longer be distinguished from one another using different combinations of risk factors.

In the MacArthur Study, the predictive accuracy achieved by the ICT (AUC = .82) was slightly better than that achieved by the standard CT (AUC = .79) and the LR model (AUC = .81) (Monahan et al., 2001). However, 25.8% more cases could be classified compared to the standard CT. Thus, the ICT approach maintained a similar level of predictive accuracy compared to the LR model and standard CT, but the iterative process made it possible to classify a considerably higher number of cases into definitive groups.

Constructing Multiple Iterative Classification Trees

Finally, researchers involved in the MacArthur Study also hypothesized that the accuracy achieved with an ICT model could be further enhanced by constructing numerous ICT models and combining the risk predictions from these models to provide a more robust ‘combined’ estimate of violence risk for each case (Monahan et al., 2001). Using a different predictor variable to initially split the cases for each CT, ten different ICT models were developed. Next, all cases were scored on each of the 10 models, with each score reflecting whether the case was classified as low, high, or unclassifiable on the corresponding ICT model. For each case, these scores were then summed to create an

overall risk estimate of violence.[8] Conducting an ROC analysis on the predictions made by the multiple ICT model, predictive accuracy was found to be significantly higher (AUC = .88) than that achieved by the single ICT model (AUC = .82). As explained by the researchers, improved predictions may be observed when combining multiple ICT models because the approach "may capture a different but important facet of the interactive relationship between the measured risk factors and violence" (Banks et al., 2004, p. 324).

The Current Study

The goal of the current article is to determine if various types of CTs outperform LR analysis when applied to serial sexual assault data consisting of crime pairs that are the work of the same vs. different offenders. A relatively novel feature of the current study is its examination of ICTs, which have only been examined in one previous crime linkage study as far as we are aware (Tonkin et al., 2017), and no other study in the crime linkage context has examined the predictive accuracy of multiple ICTs. The research also differs from previous research by examining crimes committed in Canada, a jurisdiction that has largely been ignored in crime linking research to date (with some notable exceptions; e.g., Deslauriers-Varin & Beauregard, 2013).

Hypotheses

Based on previous crime linkage research, it is expected that crime pairs committed by the same offender will be characterized by a higher degree of behavioural similarity than crime pairs committed by different offenders. As a result, it is also

[8] For each of the 10 models, a score of -1 was provided to a participant if they were in the low-risk category for that model, a score of 0 was provided if they were unclassified, and a score of +1 was provided if they were in the high-risk category. For each participant, these 10 model scores were then summed to provide the combined risk score.

hypothesized that it will be possible to distinguish crime pairs committed by the same vs. different offenders in a relatively accurate fashion (Hypothesis 1). This is expected to be the case regardless of what statistical model is used (i.e., LR, standard CT, ICT, or multiple ICTs). While previous linking research conducted by Tonkin and his colleagues (2012b, 2017) suggests that LR, standard CTs, and ICTs will result in similar levels of linking accuracy, we expect that multiple ICTs may perform best based on research from other domains (e.g., Monahan et al., 2001) (Hypothesis 2). Consistent with the results of Tonkin et al. (2012b), we also expected that all CT-based models will be less robust (i.e., have more issues with generalizability) than the LR model (Hypothesis 3). Finally, due to the interactive nature of CT analysis, patterns of behavioural similarity and distinctiveness that remain hidden when using LR are expected to emerge when using the various CT approaches. However, given the dearth of research that has examined the CT-based approach to linking crimes of an interpersonal nature, a hypothesis was not developed regarding the nature of these patterns.

Methodology

Data

Data collected as part of a large-scale Canadian study examining the characteristics of sex offenders and their offending patterns were used in this study (see Beauregard [2005] for a more detailed description of these data). This dataset has been used in several published studies to date (e.g., Beauregard et al., 2012; Hewitt et al., 2012; Reid et al., 2014). Many of these studies have focused on variables not examined in the current research (e.g., crime site selection similarity; e.g., Deslauriers-Varin &

Beauregard, 2013) and none of the previous studies have examined the range of methodological approaches used in this study.

The data were primarily collected through semi-structured interviews with federally incarcerated offenders, supplemented by crosschecks of police investigative reports included in each offender's correctional file (Beauregard, 2005). To facilitate data collection, Beauregard identified all offenders who had committed a minimum of two stranger sexual assaults and were serving a sentence of two or more years in a Quebec penitentiary between 1995 and 2004. A total of 92 sex offenders met these criteria, with 72 offenders agreeing to participate. However, only 69 offenders provided sufficient information to be included in the current analyses. These offenders committed a total of 347 stranger serial sexual assaults between 1975 and 2003. Crime series ranged in length from 2 to 37 crimes (Mode = 2.00; Median = 3.00; M = 5.00, SD = 6.04).

Exclusion and inclusion of variables. The original dataset contained over 500 variables pertaining to various aspects of the offence, offender, and victim. Given that the goal of the current study is to develop statistical models that could potentially be used for crime linkage in practice, variables reflecting information that would likely be unknown to police at the time the crime occurred were omitted from the current research (e.g., offenders characteristics). Also, while the dataset had a 'date of crime' variable and a number of location-related variables, large amounts of these data were missing. Because of this, we decided the most appropriate course of action was to exclude the temporal and spatial similarity variables from the analyses.

All remaining variables were separated into the following six domains prior to calculating similarity scores and carrying out the linkage analysis: (1) control behaviours

(e.g., knife was used during the crime; 32 variables), (2) environmental behaviours (e.g., crime occurred on a weekday; 52 variables), (3) escape behaviours (e.g., offender used a disguise; 5 variables), (4) sexual behaviours (e.g., vaginal intercourse with fingers; 17 variables), (5) style behaviours (e.g., offender complimented victim; 23 variables), and (6) victim selection (e.g., victim was male; 19 variables). The full variable list can be found as supplemental material here: LINK [to maintain anonymity this link will be provided if the paper is accepted]. All behaviours were dichotomously coded, with a 1 indicating the presence of the behaviour in a given crime and a 0 indicating the absence of the behaviour. Unfortunately, no index of inter-rater reliability is available for the data used in this study.

Controlling the impact of prolific offenders. A common concern in crime linkage research is that prolific offenders may disproportionately influence the linking models developed (Bennell & Canter, 2002). Researchers have typically controlled for prolific offenders by selecting a constant number of crimes per offender (Bennell & Canter, 2002). More recently, however, some researchers have started recommending that all identified crimes committed by offenders be included in the study to ensure the ecological validity of the linking models (Tonkin & Woodhams, 2015). Given that both these issues reflect legitimate concerns, a different approach to selecting the final sample was used in the current study.

Prolific offenders in the current dataset were first identified by detecting outliers on the variable ‘number of crimes in series.’ A variation on winsorizing (Field, 2013) was then used to bring the crime series of prolific offenders within a “normal” range of crimes. That is, for each offender/crime series that was identified as an outlier, they were provided with a new ‘number of crimes per series’ value that was one unit higher than the

highest non-outlying value (i.e., non-prolific offender). A random sample of each prolific offender's crimes that corresponded to this new number of crimes per series was then selected for inclusion in the final dataset. Using this approach, a total of eight series/offenders were identified as outliers (i.e., prolific offenders). The highest non-outlying value was 7. As such, a random sample of 8 crimes was selected from the eight prolific offenders' original crime series. This resulted in a reduced final dataset of 260 sexual assaults (series ranging in length from 2 to 8 crimes, $M = 3.74$, $SD = 2.00$).

Data Analysis

The sexual assaults contained in the dataset were submitted to a program called B-LINK (Bennell, 2002). This program creates all possible pairs of offences and uses the behavioural profile associated with each crime (0's and 1's indicating which behaviours were present or absent in each crime) to calculate a similarity score for each crime pair. In the current study, behavioural profiles existed for each of the six domains described above and thus, each crime pair was associated with six different similarity scores (one per domain). Jaccard's coefficient was used to measure behavioural similarity between crimes in each pair (Jaccard, 1908). This coefficient ranges from 0 (no similarity) to 1 (total similarity). These similarity scores, along with the dichotomous outcome variable (linked/unlinked), served as the input for both the LR analysis and CT analysis. All these analyses were conducted using SPSS.

Split-half validation was used in the current study, as it has in previous crime linkage research (e.g., Bennell & Jones, 2005; Tonkin et al., 2012a), to determine the extent to which all linking models generalize to crime pairs not used to create the models. Using the crime pairs in the development sample, LR analyses were used to examine the

individual and combined ability of the six behavioural domains to distinguish between crime pairs committed by the same vs. different offenders. Forward stepwise LR analysis was also used to determine the optimal combination of behavioural domains to distinguish between these crime pairs. Additionally, classification trees were developed using crime pairs in the development sample, with the similarity scores from each behavioural domain acting as potential predictors of the dichotomous outcome (crime pairs committed by the same vs. different offenders). The procedures adopted by Monahan et al. (2001), which were described above, were used to develop ICTs and multiple ICTs.[9]

To assess the accuracy of all prediction models when applied to the test sample, ROC analysis was used, along with an assessment of several other performance metrics (e.g., percentage of crime pairs correctly classified). Comparisons were made between the LR models and the CT models, and between the development and test samples for each statistical method to examine generalizability. We conducted significance tests to compare the AUCs across all models and sub-samples. The procedure outlined in Hanley and McNeil (1982) was used for this purpose.

Results

[9] In addition to these general procedures, a variety of user-specified decisions related to model parameters needed to be made when constructing CTs for the development and test samples. These included the selection of the particular chi-square test used for splitting the data according to the predictors, the level of significance set for these tests, the maximum number of intervals that the continuous predictors can be separated into, the minimum number of cases that must be present in each successive node, and the maximum tree depth allowed. First, in terms of the chi-square test used for determining node splitting, the likelihood ratio chi-square test was selected. Second, although the default significance level for partitioning nodes in SPSS is $p < .05$, it was decided to adjust this to $p < .01$ because the samples are relatively large, and a more conservative significance level makes it less likely that the resulting models capitalize on chance. Third, the default level of 10 was used as the maximum number of categories permitted to separate the continuous predictors. Fourth, we decided to maintain the default levels of 100 and 50 cases for parent and child nodes, respectively. Finally, although the SPSS default for tree depth is three, tree depth was set to the number of predictors involved in the analysis to ensure that each predictor had at least one chance to be included in the tree.

Descriptive statistics are displayed in Table 1 as a function of behavioural domain. Significance tests were carried out to examine the differences between crime pairs committed by the same vs. different offenders. A subset of crime pairs committed by different offenders equal to the total number of crime pairs committed by the same offender (n = 495) was randomly extracted from the complete dataset to facilitate these analyses.[10] Kolmogorov-Smirnov tests of normality confirmed that both distributions for each behavioural domain were significantly different from normal. As a result, Mann-Whitney U tests were conducted to examine the median differences between the distributions for each domain. These results indicate that *J*-scores were significantly larger for crime pairs committed by the same offender for all behavioural domains. Effect sizes were in the medium range.

[Insert Table 1 here]

Developing and Evaluating Main Effects LR Models

The results of the single linking feature LR models constructed using the development sample data are presented in Table 2. The model statistics suggest that all behavioural domains were significant predictors of linkage status.[11] Next, a forward stepwise LR model was constructed using the development sample to determine the optimal combination of variables for predicting linkage status. Steps continued so long as the inclusion of additional variables significantly improved the predictive power of the model. The analysis proceeded through six steps, with style behaviours, control

[10] This was only done for this analysis. The full set of linked and unlinked crime pairs were used to develop and evaluate the LR, CT, ICT, and multiple ICT models.

[11] Readers may note that the rank-order of variables (based on their predictive power) changes depending on whether the results from the LR analysis or ROC analysis are relied on. Similar findings have been highlighted by others (e.g., Demler et al., 2011). We tend to rely on the results from ROC analysis in these cases, but future research should explore this issue in more depth.

behaviours, victim selection, sexual behaviours, environmental behaviours, and escape behaviours all being retained in the final model (all *p*'s < .01; see Table 2).

The results from ROC analyses conducted on the development and test samples are also presented in Table 2. For the test sample, the AUC for the stepwise model was significantly higher than the AUCs for all other models (all *p*'s < .01) with the exception of the victim selection AUC ($z = 1.6$, $p = .10$). When examining the pattern of AUCs across the development and test samples, all LR models generalize well to the test sample crime pairs.

[Insert Table 2 here]

Developing and Evaluating Standard CT and ICT Models

The standard CT for the development sample had 4 levels with 37 nodes, 23 of which were terminal nodes.[12] All variables appeared at least once in the CT, with the exception of the escape domain. The base rate of crime pairs committed by the same offender in the development and test samples was 1.50%. Following Monahan et al. (2001), terminal nodes containing greater than 3.00% crime pairs committed by the same offender were classified as linked, nodes containing less than 0.75% crime pairs committed by the same offender were classified as unlinked, and nodes containing a proportion of crime pairs committed by the same offender that was equal to or fell between these two thresholds (0.75–3.00%) were deemed unclassified.

For the development sample, this resulted in eight linked nodes, nine unlinked nodes, and six unclassified nodes. A similar pattern was found when applying these cut-offs to the same model in the test sample, although some nodes were labelled differently.

[12] The full CT graphic is too large to present in its entirety here.

Specifically, one node moved from being labelled linked to unclassified, two nodes moved from being unclassified to unlinked, and one node moved from being labelled unlinked to unclassified. Importantly, no nodes moved from being linked to unlinked (or vice versa). In total, 15.18% of crime pairs in the development sample and 7.25% of crime pairs in the test sample were deemed unclassified in this standard CT.

To construct an ICT, the unclassified cases were analyzed a second time. The second iteration produced a much simpler CT with two nodes, both of which were terminal nodes. The ICT was able to classify an additional 1.50% of crime pairs in the development sample and 1.20% of crime pairs in the test sample. No further cases could be classified.

The results of the ROC analyses for the standard CT and ICT are presented in Table 3. Comparisons of the AUCs across the development and test samples suggest that generalizability was not an issue with any of the CT models. No significant difference was found between the standard CT and the ICT in the test sample ($z = .61$, $p = .54$).

[Insert Table 3 here]

Developing and Evaluating Multiple CT/ICT Linking Models

The final phase of analysis involved constructing the multiple ICTs. Six separate ICT models were first developed, each one forcing a different predictor as the initial splitting variable in the first iteration of each CT. All CTs proceeded through two iterations, except for the CT where the style domain was forced as the first variable (only a standard CT was produced). The characteristics of the CTs/ICTs produced are summarized in Table 4. Each model's AUC, and its overall ability to classify cases into linked versus unlinked subgroups, are also displayed in Table 4. Most CTs/ICTs

classified a high number of the crime pairs as linked or unlinked. AUCs were also high for the development and test sample. Although some "shrinkage" was evident for all models (i.e., the models appear to perform slightly less well on the test sample; Everitt, 2002), the only model with a significant amount of shrinkage was the victim selection model ($z = 2.37$, $p = .02$).

[Insert Table 4 here]

Next, each crime pair was provided a score based on how they were classified in each model using the two-threshold classification approach described previously. Specifically, for each model, crime pairs labelled as linked were assigned a score of 1, crime pairs labelled as unlinked were assigned a score of -1, and crime pairs labelled as unclassified were assigned a score of 0. A composite score was then created for each crime pair based on how they were classified on the models combined (summing across all their scores; Cronbach's alpha was .86 for these six scores indicating a satisfactory level of internal reliability). Composite scores could range from -6 (indicating that the crime pair was classified as unlinked on all six models) to +6 (indicating that the crime pair was classified as linked on all six models), with a median score of -6.00 ($SD = 2.55$).

The total number of crime pairs possessing each composite score value, and the corresponding percentage of those cases that are crime pairs committed by the same vs. different offenders, was examined (see Table 5). As expected, the majority of crime pairs committed by the same offender in both the development and test samples (82.66% and 79.35%, respectively) had a score of 1 or higher (indicating they were in the linked category more often across the six models than the other categories). Likewise, the majority of all crime pairs committed by different offenders in the development and test

samples (92.99% and 93.40%, respectively) possessed a score of -1 or lower (indicating they were in the unlinked category more often across the six models than the other categories).

[Insert Table 5 here]

To create an empirically optimal multiple CT/ICT, a forward stepwise LR analysis was conducted using the scores for each model entered as the predictors (six predictors) and linkage status as the outcome variable. Scores for five of the six models were retained in the stepwise model (the scores for the model beginning with the control domain were excluded). Cronbach's alpha ($\alpha = .83$) for these five score variables indicated a satisfactory level of internal reliability. As such, a modified composite score was calculated for each crime pair using only the scores forming these five models. Similar to the original composite score, the majority of crime pairs committed by the same offender in the development and test samples (82.66% and 81.16%, respectively) had a score of 1 or higher, and the majority of crime pairs committed by different offenders in the development and test samples (93.51% and 92.73%, respectively) had a score of -1 or lower on the modified composite score variable (see Table 5).

ROC analysis was conducted to examine the predictive accuracy of the original and modified composite scores (i.e., the original multiple CT/ICT model and the 'optimal' CT/ICT model). The results are presented in Table 6. Both the original and optimal multiple CT/ICT models resulted in the same level of predictive accuracy for both the development and test samples.

[Insert Table 6 here]

Comparing the Performance of all Linking Models

The final step was to compare all the models. Table 7 includes the "confusion matrix" for each model, which indicates the frequencies of the various decision outcomes that resulted when each model was applied to the test sample (i.e., true positives [TP], false negatives [FN], true negatives [TN], and false positives [FP]). Table 8 includes the predictive accuracies for the LR model, the standard CT, the ICT, and the original and optimal multiple CT/ICT models, along with a range of other performance metrics associated with each model when they were applied to the test sample, including the percentage of crime pairs left unclassified, the percentage of crime pairs correctly classified, and the TP, FN, TN, and FP rates. Finally, a ROC graph that includes curves for each of the models (when they were applied to the test sample) is included in Figure 2.

In general, these results demonstrate that the statistical models examined in this study can accurately differentiate between crimes committed by the same vs. different offenders, although a reasonable number of false positive decisions are made with each model. The results also suggest that there are no significant differences between the AUCs achieved by any of the models (all p's > .05), although potentially important differences emerged suggesting that the CT-based approaches could classify a greater percentage of crime pairs than the LR model and classify crime pairs correctly to a greater degree. Finally, comparisons of the AUCs across the development and test samples suggest that generalizability is not a concern for any of the models we examined; the AUCs across the two samples were almost identical.

[Insert Table 7 here]

[Insert Table 8 here]

[Insert Figure 2]

Discussion

This study set out to examine the relative accuracy and generalizability of various statistical linking methods, including LR and models based on CT analysis. In support of Hypothesis 1, the results indicated that it is possible to use statistical models to distinguish crimes committed by the same vs. different offenders. On the other hand, Hypothesis 2 was only partially supported. We hypothesized that LR, standard CT, and ICT models would have comparable levels of predictive accuracy and they did have very similar AUCs (although the CT-based models had higher classification rates and higher classification accuracy). However, we predicted that multiple CT/ICT models would exhibit superior performance, but the AUCs associated with these models were not significantly higher than any other model. Likewise, contrary to Hypothesis 3, we found no serious issues with shrinkage across any of the statistical methods we tested, including the CT models, which stands in contrast to the findings of Tonkin et al. (2012b). Finally, although no formal hypotheses were developed around this issue, multiple pathways to making a 'linked' decision were identified in the CT analysis, highlighting the idiographic nature of CTs relative to the LR analysis we conducted.

The Consistency and Distinctiveness of Canadian Serial Sexual Offenders

Generally speaking, previous research has found that it is possible to distinguish sexual offences that have been committed by the same vs. different offenders to a moderate degree (AUCs ranging from .75 [Bennell et al., 2009] to .89 [Winter et al., 2013]). The results of the current study add further support to this literature; however, a particularly high degree of predictive accuracy was achieved by the multi-variable

models developed in this study (all AUCs were > .90). This finding suggests that it may indeed be possible to distinguish between serial sexual assaults committed by the same vs. different offenders in Canada, possibly with a high degree of accuracy.

There are at least two possible explanations for the increased predictive accuracy observed in the current study. First, we know from additional data collected by Beauregard (2005) that at least some of the offenders in the sexual assault dataset had a history of psychiatric problems (n = 16; 23%). Research conducted by Woodhams and Komarzynska (2014), which examined the offence behaviours exhibited by mentally disordered sexual offenders, suggests that these individuals exhibit highly similar and distinctive behaviours across their offences, which can result in high levels of linking accuracy. To the degree that these types of individuals exist within the current dataset, this might explain the high AUCs we found. Future research can test this possibility if researchers are able to collect more detailed information about the mental health status of offenders.

Second, unlike previous studies that relied on police data, the data in the current study were collected through offender interviews. The use of semi-structured interviews may have allowed Beauregard (2005) to collect richer information (Brookman, 2010), while also standardizing the data collection process, which could account for the improved predictive accuracy observed in the current study. That being said, it is important to note that offender interviews may also be problematic (e.g., offenders may distort their accounts, have little insight into their own behaviours, fail to remember certain things, etc.). Likewise, reliance on interview data arguably decreases the ecological validity of the current study compared to crime linkage studies that have relied

on police data. Future research should examine the ways in which these different data collection protocols impact the findings of linking studies.

The Accuracy of CT-based versus LR-based Models

Although we found no differences between the statistical models we tested with respect to their AUCs, we did find potentially important differences in some of the other metrics we examined. Perhaps most notably, CT-based models were able to classify a greater percentage of crime pairs than the LR model and they were able to classify those crime pairs more accurately. Indeed, the percentage of crime pairs left unclassified by the LR model was more than twice the percentage of crime pairs left unclassified by the best performing CT-based models, and the best performing CT-based models achieved a classification accuracy rate than was more than 10% higher than the LR model. These findings suggest that CT-based models may be preferable to LR models that are more commonly examined in the crime linkage literature. The practical implications of these findings should be explored by testing the models under more ecologically valid conditions (e.g., see Woodhams et al., 2019).

The fact that the multiple CT/ICT models were not associated with significantly higher AUCs than all the other models was unexpected. Indeed, consistent with Monahan et al. 2001), we expected that these models would be better able to capture the complexity of sexual assault behaviour, and as a result, significantly outperform the main effects LR model, the standard CT, and the ICT. Why were these results not found? Two explanations seem plausible. First, because our LR models, standard CT, and ICT were associated with such high levels of predictive accuracy, there may be little room for improvement. Second, the multiple CT/ICT models used in the current study may not be

complex enough to reveal the sorts of findings reported by Monahan et al. For example, the models we developed involved fewer iterations than those reported on by Monahan and his colleagues. These differences are likely the result of methodological issues, most notably the limited number of predictors available for use in the current research compared to the 134 risk predictors examined by Monahan and his colleagues.

Although slight increases in predictive accuracy were observed across the different CT approaches, it appears as though a standard CT approach might be the best model to use for linking purposes. Indeed, the results from our study suggest that, despite there being complex behavioural patterns underlying serial sexual assaults, attempting to capture this complexity using the ICT or multiple CT/ICT method may not add much value in terms of our ability to link these crimes. While it may not take more effort to use multiple CT/ICT models in practice, given that all models would likely be automated through some sort of decision support system, the additional work involved in developing (and interpreting) a multiple CT/ICT model for linking sexual assaults may not be worth it if such models do not result in great levels of linking performance.

Finally, it is important to reiterate that, even though the CTs produced from the sexual assault data were relatively idiographic compared to the LR models, we found no issues with generalizability. It is unclear why the results from the split-half validation results differ from those reported by Tonkin et al. (2012b). One possibility is that the parameters used for CT development in the current study were slightly different, with those employed by Tonkin and his colleagues resulting in CTs that might not generalize as well to crimes committed by different offenders. Regardless of why the differences between the two studies emerged, shrinkage was not found to be an issue with the current

linking models (i.e., the models performed at comparable levels for the development and test samples). This suggests that all the models developed may apply comparably well to other sexual assaults committed in Quebec.

Capturing the Complexities of Serial Sexual Assaults

A key argument that has been presented in favour of a CT-based approach to crime linking is that CTs will be better able to capture the complexity of offending behaviour than a traditional main effects LR approach; that is, assuming that subsets of offenders do in fact differ in the extent to which they commit crimes that are behaviourally similar and distinctive, the interactive nature of CTs will be able to capture these differences, leading to multiple, tailored pathways for making linkage decisions. These pathways clearly emerged in the current study.

Indeed, as shown in Table 9, a total of seven pathways to identifying a crime pair as linked were identified when applying the standard CT to the test sample.[13] The pathways in Table 7 are rank ordered by the percentage of crime pairs committed by the same offender found within each pathway. As shown, if one were to rely on high similarity in control behaviours, high similarity in style behaviours, and high similarity in environmental behaviours (Pathway 1), they could be relatively confident that they were dealing with crimes committed by the same offender (84.9% chance). In contrast, if one were to rely on moderate similarity in control behaviours, moderate similarity in victim selection behaviours, and lower similarity in sexual behaviours (Pathway 7), they should

[13] There were an additional 6 pathways to making a linked decision using the ICT. Given that the CT and ICT resulted in similar levels of predictive accuracy, only the standard CT pathways will be discussed here for the sake of brevity.

be much less confident that they were dealing with crimes committed by the same offender.

[Insert Table 9 here]

Although it is not clear what these pathways mean at this time, what can generally be concluded is that serial sexual assaults do seem to differ from one another in terms of the types of behaviours for which they are similar and distinctive. For example, offenders falling along Pathway 1 are highly similar in their control, style, and environmental behaviours. This may arguably reflect the fact that these offenders engage in high levels of pre-offence planning and fantasy (Gee & Belofastov, 2014). However, other sexual offenders are much less consistent in their control behaviours, yet they seem to be consistent with respect to the types of victims they select (e.g., Pathways 5 and 6). It is possible that these offenders are not predisposed to engage in specific control behaviours across their crimes (e.g., they are opportunistic rather than planners), even though they have a highly specific preference for a certain type of victim (consequently allowing us to link their crimes on the basis of victim selection similarity).

It is important to stress that the explanations attached to these pathways are only speculative at this time and that future research is needed before the true meaning of these pathways can be known. Future research should also explore why CTs differ in their level of behavioural complexity. While it is obvious that the CTs produced in the current study are much more complex than those produced by Tonkin et al. (2012b), in that their CT involved only two pathways to a linkage decision and they could not create an ICT, it is not necessarily clear why this difference was found. One possibility is that sexual assaults are more behaviourally complex than property crimes, perhaps due to the involvement of

a victim or because there are simply a greater range of behaviours that can be displayed in a sexual crime. However, another possibility, which we touch on below, is that the differences in CT complexity between our study and Tonkin et al.'s study are the result of methodological decisions related to the construction of the CTs.

Limitations of the Current Research

There are a number of limitations of the current research that warrant discussion. First, only a small subset of classification methods was tested in the current study and the methods that were examined are unlikely to produce the most accurate linkage decisions. Indeed, even though CT-based approaches are being examined by researchers for the purpose of crime linkage (e.g., Tonkin et al., 2012b, 2017), more sophisticated tree methodologies exist that would likely perform better than the ICT and multiple CT/ICTs examined in this paper. Examples of such methodologies include AdaBoost (short for Adaptive Boosting) and Random Forest algorithms (e.g., Thongkam, Xu, & Zhang, 2008). Future research should explore these methodologies (and other machine learning algorithms) to determine if they perform better on crime linkage classifications tasks than the approaches we examined.[14]

Second, the fact that split-half validation was used in the current study to determine the extent to which the linking models generalized is not ideal since it ultimately means that half the data were not available to estimate (or test) the model. This validation procedure was chosen because it is commonly used in crime linkage research (e.g., Bennell & Jones, 2005; Tonkin et al., 2012a), but more robust procedures should be examined in future research. While it may not be possible to apply some validation

[14] We would like to thank one of the anonymous reviewers for bringing these issues to our attention.

procedures to all the models examined in this paper (e.g., leave-one-out cross-validation [Tonkin & Woodhams, 2015] would be challenging to use with CT-based models), other procedures are likely to be more viable. For example, k-fold cross-validation, where multiple development and test samples from the same dataset are generated, is a promising approach that should be considered.

Third, in hindsight, another possible limitation of the current research is the way in which the behavioural domains were operationalized. Although a popular approach in the crime linking literature (Bennell et al., 2014), defining the behavioural domains in an atheoretical manner (based on their assumed function) may have led to lower levels of linking accuracy. This approach to defining behavioural domains may have also contributed to difficulties interpreting the different pathways that arose. If a goal of crime linking research is to further our understanding of offender behaviour, then future research should examine ways to use a CT approach (which does appear to capture the complexity in offending behaviour) with more theoretically informed predictors. Alternatively, it may be useful to explore the value of creating empirical behavioural domains using different statistical procedures (e.g., multidimensional scaling, cluster analysis, principal component analysis, etc.).

A final limitation of the current research concerns the parameters that were used to construct the CT models (see footnote 9). It is important to note that the combination of parameters selected in the current research, and the combination of parameters selected by Tonkin et al. (2012b, 2017), are among countless options for parameter selection. The parameters in the current study were ultimately selected because it was believed they simultaneously created a parsimonious model, while still also capitalizing on the ability

of the CT approach to capture more complex ‘signals’ in the data than the traditional main effects linking approach. However, it must be appreciated that CT results are entirely dependent on the parameters that are selected and it is certainly plausible that the CT models developed in this study (or Tonkin et al.’s studies) were not constructed in the most optimal way. Similarly, it is also important to highlight that the thresholds chosen to classify the nodes of the CTs (as linked, unlinked, or unclassified) could obviously have a large impact on model performance. We chose one set of thresholds, but another set would have resulted in different results. Future research should more thoroughly explore the impact of different parameter-setting and threshold-setting methods on model performance when using CT-based approaches.

Conclusion

The current study adds to the scant literature aimed at determining the most suitable statistical method to link crimes. The current study’s comparison of LR, standard CTs, ICTs, and multiple CT/ICTs found no significant differences between the approaches, although multiple CT/ICTs were marginally better. That being said, we believe that the marginal increase in accuracy observed with the multiple CT/ICT approach does not necessarily outweigh the complexities involved in developing or interpreting these types of crime linking models.

These findings are important theoretically in that they offer a better understanding of criminal offending patterns (e.g., offenders appear to vary in terms of which of their behavioural domains remain consistent). More practically, examination of CTs as a method aimed at capturing such idiosyncrasies also provides a possible alternative for analysts interested in a more flexible approach to crime linkage analysis. If the crime

linkage methods we examined could be built into decision support systems, analysts could then discover first-hand which approach best meets their needs (e.g., in terms of being accurate, user-friendly, transparent, etc.).

Before analysts are consulted though, future field research may be required to identify if (and how) CTs can actually aid the crime linking process. We expect that some improvements can be made to how practitioners link serial crimes using the statistical approaches discussed in this article, and our hope is that very soon, the sorts of empirically informed structured decision-making approaches that are becoming common place in other domains (e.g., risk assessment; Hart et al., 2017) will become more common in the crime linkage context.

References

Banks, S., Robbins, P. C., Silver, E., Vesselinov, R., Steadman, H. J., Monahan J., … Roth, L. H. (2004). A multiple-models approach to violence risk assessment among people with mental disorder. *Criminal Justice and Behavior, 31*(3), 324-340.

Beauregard, E. (2005). *Hunting process of serial sex offenders: A rationale choice approach* (Unpublished doctoral dissertation). University of Montreal, Canada.

Beauregard, E., Leclerc, B., & Lussier, P. (2012). Decision making in the crime commission process: Comparing rapists, child molesters, and victim-crossover sex offenders. *Criminal Justice and Behavior, 39,* 1275-1295.

Bennell, C. (2002). Behavioural consistency and discrimination in serial burglary (Unpublished doctoral dissertation). University of Liverpool, UK.

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. *Science & Justice, 42*(3), 153-164.

Bennell, C., & Jones, N. J. (2005). Between a ROC and a hard place: A method for linking serial burglaries by modus operandi. *Journal of Investigative Psychology and offender profiling*, *2*(1), 23-41.

Bennell, C., Jones, N. J., & Melnyk, T. (2009). Addressing problems with traditional crime linkage methods using receiver operating characteristic analysis. *Legal and Criminological Psychology*, *14*(2), 293-310.

Bennell, C., Mugford, R., Ellingwood, H., & Woodhams, J. (2014). Linking crimes using behavioural clues: Current levels of linking accuracy and strategies for moving forward. *Journal of Investigative Psychology and Offender Profiling, 11*, 29-56.

Brookman, F. (2010). Beyond the interview: Complementing and validating accounts of incarcerated violent offenders. In W. Bernasco (Ed.), *Offenders on offending: Learning about crime from criminals* (pp. 84-105). Portland, OR: Willan.

Demler, O. V., Pencina, M. J., & D'Agostino Sr., R. B. (2011). Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in Medicine*, *30*(12), 1410-1418.

Deslauriers-Varin, N., & Beauregard, E. (2013). Investigating offending consistency of geographic and environmental factors among serial sex offenders A comparison of multiple analytical strategies. *Criminal Justice and Behavior, 40*(2), 156-179.

Everitt, B. S. (2002). *The Cambridge dictionary of statistics* (2nd edition). Cambridge, UK: Cambridge University Press.

Field, A. (2013). *Discovering statistics using SPSS (and sex, drugs, and rock 'n' roll)* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Gee, D., & Belofastov, A. (2014). Sex crime linkage: Sexual fantasy and offense plasticity. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 33-53). Boca Raton, FL: CRC Press.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Hart, S. D., Douglas, K. S., & Guy, L. S. (2017). The structured professional judgement approach to violence risk assessment: Origins, nature, and advances. In D. P. Boer, A. R. Beech, T. Ward, L. A. Craig, M. Rettenberger, L. E. Marshall, & W. L. Marshall (Eds.), *The Wiley handbook on the theories, assessment, and treatment of sexual offending* (pp. 643–666). Wiley Blackwell.

Hewitt, A., Beauregard, E., & Davies, G. (2012). “Catch and release”: Predicting encounter and victim release location choice in serial rape events. *Policing, 35,* 835-856.

Jaccard, P. (1908). Nouvelle recherches sur la distribution florale. *Bulletin de la Société vaudoise des Sciences Naturelles, 44*, 223-270.

Lundrigan, S., Czarnomski, S., & Wilson, M. (2010). Spatial and environmental consistency in serial sexual assault. *Journal of Investigative Psychology and Offender Profiling, 7*, 15-30.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., … Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York, NY: Oxford University Press.

Reid, J. A., Beauregard, E., Fedina, K. M., & Frith, E. N. (2014). Employing mixed methods to explore motivational patterns of repeat sex offenders. *Journal of Criminal Justice, 42,* 203-212.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26.

Thongkam, J., Xu, G., & Zhang, Y. (2008, June). AdaBoost algorithm with random forests for predicting breast cancer survivability. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 3062-3069). IEEE.

Tonkin, M., Pakkanen, T., Siren, J., Bennell, C., Woodhams, J., Burrell, A.,…Santtila, P. (2017). Using offender crime scene behavior to link stranger sexual assaults: A comparison of three statistical approaches. *Journal of Criminal Justice, 50,* 19-28.

Tonkin, M., & Woodhams, J. (2015). The feasibility of using crime scene behaviour to detect versatile serial offenders: An empirical test of behavioural consistency, distinctiveness, and discrimination. *Legal and Criminological Psychology, 22*(1), 99-115.

Tonkin, M., Woodhams, J., Bull, R., & Bond, J. W. (2012a). Behavioural case linkage with solved and unsolved crimes. *Forensic Science International, 222*(1-3), 146-153.

Tonkin, M., Woodhams, J., Bull, R., Bond, J. W., & Santtila, P. (2012b). A comparison of logistic regression and classification tree analysis for behavioural case linkage. *Journal of Investigative Psychology and Offender Profiling, 9,* 235-258.

Wang, T., Qin, Z., Jin, Z., & Zhang, S. (2010). Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning. *Journal of Systems and Software, 83*(7), 1137-1147.

Winter, J. M., Lemeire, J., Meganck, S., Geboers, J., Rossi, G., & Mokros, A. (2013). Comparing the predictive accuracy of case linkage methods in serious sexual assaults. Journal of Investigative Psychology and Offender Profiling, 10, 28-56.

Woodhams, J., & Komarzynska, K. (2014). The effect of mental disorder on crime scene behaviour, its consistency, and variability. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 55-82). Boca Raton, FL: CRC Press.

Woodhams, J., Tonkin, M., Burrell, A., Imre, H., Winter, J. M., Lam, E. K., ... & Santtila, P. (2019). Linking serial sexual offences: Moving towards an ecologically valid test of the principles of crime linkage. *Legal and Criminological Psychology*, *24*(1), 123-140.

Tables

Table 1. Descriptive statistics and results from non-parametric tests of similarity scores for crime pairs committed by the same vs. different offenders across each behavioural domain.

Domains	Mdn (*J*)		*U*	Effect Size (*r*)
	Crime Pairs Committed by the Same Offender	Crime Pairs Committed by Different Offenders		
Control Behaviours	.67	.14	208,522	.61
Environmental Behaviours	.85	.26	198,199	.54
Escape Behaviours	.00	.00	143,912	.28
Sexual Behaviours	.80	.10	185,633	.46
Style Behaviours	1.00	.14	204,755	.60
Victim Selection	.67	.25	214,363	.65

Note. *J* = Jaccard's coefficient; *U* = Mann-Whitney U statistic; $r = z/\sqrt{N}$ where *z* is the absolute (positive) standardized test statistic from the Mann-Whitney U test (Rosenthal, 1991); all *p*'s <.001.

Table 2. Results of separate simple LR analyses, and stepwise LR analysis, for each predictor variable.

							Development Sample		Test Sample	
Model	Constant (*SE*)	*B* (*SE*)	Wald (*df*)	χ^2 (*df*)	R_N^2	R_L^2	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Control	-6.10 (0.14)	5.39 (0.21)	677.16 (1)	672.89 (1)	.28	.26	.86 (.01)	.84 – .89	.87 (.01)	.84 – .90
Environmental	-6.93 (0.18)	6.05 (0.26)	538.64 (1)	566.13 (1)	.23	.22	.79 (.02)	.75 – .82	.82 (.02)	.79 – .86
Escape	-4.46 (0.07)	4.33 (0.22)	391.48 (1)	261.45 (1)	.11	.10	.59 (.02)	.55 – .63	.58 (.02)	.54 – .63
Sexual	-5.74 (0.13)	4.70 (0.21)	508.64 (1)	484.32 (1)	.20	.19	.72 (.02)	.68 – .76	.79 (.02)	.75 – .83
Style	-6.00 (0.13)	5.36 (0.20)	728.49 (1)	709.36 (1)	.29	.27	.81 (.02)	.77 – .84	.86 (.02)	.83 – .89
Victim Sel.	-6.88 (0.17)	5.94 (0.25)	582.28 (1)	638.83 (1)	.26	.25	.86 (.01)	.83 – .88	.89 (.01)	.87 – .91
Stepwise				1243.04 (6)	.50	.48	.88 (.01)	.86 – .91	.92 (.01)	.89 – .94
Style		2.65 (0.28)	88.79 (1)							
Control		3.01 (0.29)	104.35 (1)							
Victim sel.		1.82 (0.35)	26.90 (1)							
Sexual		1.77 (0.29)	37.18 (1)							
Environmental		1.83 (0.34)	28.63 (1)							
Escape		1.76 (0.36)	24.70 (1)							
Constant		-8.02 (0.22)	1364.08 (1)							

Note. χ^2 = model chi-square; R_N^2 = Nagelkerke index; R_L^2 = Hosmer and Lemeshow's index; AUC = area under the ROC curve; *SE* = standard error; 95% CI = 95% confidence interval; all *p*'s < .001.

Table 3. Development and test sample AUCs, standard errors, and 95% CIs for the CT models.

	Development Sample		Test Sample	
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Standard CT	.94 (.01)	.92 – .95	.92 (.01)	.90 – .95
ICT	.94 (.01)	.93 – .96	.93 (.01)	.90 – .95

Note. AUC = area under the ROC curve; *SE* = standard error; 95% CI = 95% confidence interval; all *p*'s <.001.

Table 4. Characteristics of the CTs produced in Iteration 1 and 2 when each predictor was forced as the initial splitting variable for the first iteration of each tree developed.

First Forced Variable	Depth	Nodes	Terminal Nodes	Variables Included	Development Sample			Test Sample		
					Classified (%)	AUC (*SE*)	95% CI	Classified (%)	AUC (*SE*)	95% C
Iteration 1 CT										
Control	4	37	23	All Except Escape	86.40	.94 (.01)	.93 – .96	93.90	.93 (.01)	.90 – .9
Environmental	5	38	24	All Except Escape	88.10	.94 (.01)	.92 – .95	76.10	.91 (.01)	.89 – .9
Escape	5	34	21	All Variables	87.80	.95 (.01)	.93 – .96	79.50	.92 (.01)	.90 – .9
Sexual	4	43	28	All Except Escape	92.30	.95 (.01)	.93 – .96	92.50	.93 (.01)	.91 – .9
Style	5	35	21	All Except Escape	86.80	.93 (.01)	.91 – .95	82.50	.91 (.01)	.88 – .9
Victim Sel.	4	34	22	All Except Escape	88.20	.94 (.01)	.93 – .96	88.70	.90 (.01)	.88 – .9
Iteration 2 CT										
Control	1	3	2	Sexual						
Environmental	1	4	3	Sexual						
Escape	1	3	2	Sexual						
Sexual	1	3	2	Environmental						
Victim Sel.	1	3	2	Style						

Note. A second iteration was not produced when the style domain was forced as the initial splitting variable at the first iteration; AUC = area under the ROC curve; *SE* = standard error; 95% CI = 95% confidence interval; all *p*'s <.001.

Table 5. Distribution of composite (and modified composite) linkage scores for the development and test samples.

	Composite Scores				Modified Composite Scores			
	Development Sample		Test Sample		Development Sample		Test Sample	
Score	*n*	% Linked	*n*	% Linked	*n*	% Linked	*n*	% Linked
- 6	9,969	0.07	8,855	0.05	--	--	--	--
- 5	2,200	0.32	2,521	0.24	10,478	0.08	8,947	0.04
- 4	1,206	0.75	1,770	0.62	2,210	0.41	2,694	0.22
- 3	898	0.22	1,429	0.70	1,266	0.63	1,775	0.60
- 2	786	1.15	485	0.62	825	0.48	1,484	1.01
- 1	404	0.74	478	2.30	771	1.23	525	1.52
0	365	1.64	329	1.82	279	1.43	430	1.16
1	143	3.50	213	0.94	248	3.63	291	3.09
2	156	2.56	145	7.59	166	3.01	169	4.14
3	146	4.11	124	1.61	216	5.09	176	3.98
4	201	5.97	164	3.67	132	13.64	114	10.53
5	120	14.17	98	12.24	245	66.12	229	71.18
6	242	66.53	223	73.09	--	--	--	--

Table 6. Development and test sample AUCs, standard errors, and 95% CIs for the multiple CT/ICT models.

	Development Sample		Test Sample	
Multiple CT/ICT Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI
Original MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97
Optimal MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97

Note. MM = multiple model; AUC = area under the ROC curve; *SE* = standard error; 95% CI = 95% confidence interval; all *p*'s <.001

Table 7. Confusion matrices for the LR, standard CT, ICT, and multiple CT/ICT models.

LR	Predicted			
Observed	Unlinked	Unclassified	Linked	% Correct
Unlinked	13,254	2,506	827	79.91
Linked	39	15	193	78.14
Total	13,293	2,521	1,020	79.88
Standard CT	Predicted			
Observed	Unlinked	Unclassified	Linked	% Correct
Unlinked	14,680	1,194	713	88.50
Linked	38	26	183	74.09
Total	14,718	1,220	896	88.29
ICT	Predicted			
Observed	Unlinked	Unclassified	Linked	% Correct
Unlinked	14,680	1,002	905	88.50
Linked	38	16	193	79.91
Total	14,718	1,018	1,098	88.35
Original MM	Predicted			
Observed	Unlinked	Unclassified	Linked	% Correct
Unlinked	15,026	790	771	90.50
Linked	34	17	196	79.35
Total	15,060	807	967	90.42
Optimal MM	Predicted			
Observed	Unlinked	Unclassified	Linked	% Correct
Unlinked	14,864	942	781	89.61
Linked	36	13	198	80.16
Total	14,900	955	979	89.47

Note. Total N = 16,834; linked n = 247; unlinked n = 16,587; % Correct = percent of all crime pairs correctly classified (excluding unclassified crime pairs) = (TP+TN)/(TP+TN+FP+FN).

Table 8. Development and test sample AUCs, standard errors, and 95% CIs, and other performance metrics for the LR model, the standard CT, ICT, and the multiple CT/ICT models (MM).

	Development Sample		Test Sample							
Model	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	% Unclassified	% Correct	TPR	FNR	TNR	FPR
Stepwise LR	.88 (.01)	.86 – .91	.92 (.01)	.89 – .94	14.98%	79.88%	.83	.17	.94	.06
Standard CT	.94 (.01)	.92 – .95	.92 (.01)	.90 – .95	7.25%	88.29%	.83	.17	.95	.05
ICT	.94 (.01)	.93 – .96	.93 (.01)	.90 – .95	6.05%	88.35%	.84	.16	.94	.06
Original MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97	4.79%	90.42%	.85	.15	.95	.05
Optimal MM	.95 (.01)	.93 – .97	.95 (.01)	.94 – .97	5.67%	89.47%	.84	.16	.95	.05

Note. AUC = area under the ROC curve; *SE* = standard error; 95% CI = 95% confidence interval; % Unclassified = percent of crime pairs left unclassified when adopting the two-threshold approach; % Correct = percent of all crime pairs correctly classified (excluding unclassified crime pairs) = (TP+TN)/(TP+TN+FP+FN); TPR = true positive rate (excluding unclassified crime pairs) = sensitivity = TP/(TP+FN); FNR = false negative rate (excluding unclassified crime pairs) = FN/(TP+FN); TNR = true negative rate (excluding unclassified crime pairs) = specificity = TN/(TN+FP); FPR = false positive rate (excluding unclassified crime pairs) = FP/(TN+FP); all *p*'s <.001.

Table 9. The seven different pathways in the CT leading to a 'linked' decision for pairs of serial sexual assaults when using the two-threshold approach proposed by Monahan et al. (2001).

Pathway	Percent of Crime Pairs Committed by the Same Offender
1 control behaviours (high similarity: J >.429)	84.9
→ style behaviours (high similarity: J >.400)	
→ environmental behaviours (high similarity: J >.571)	
2 control behaviours (moderate-to-high similarity: J= .308-.429)	12.0
→ style behaviours (high similarity: J >.400)	
→ victim selection (mod-to-high similarity: J >.333)	
3 control behaviours (high similarity: J >.429)	11.2
→ style behaviours (low-to-mod similarity: J = .167-.200)	
→ victim selection (mod-to-high similarity: J >.333)	
→ environmental (mod-to-high similarity: J >.211)	
4 control behaviours (high similarity: J >.429)	9.5
→ style behaviours (high similarity: J >.400)	
→ environmental behaviours (low-to-mod similarity: $J \leq .571$)	
5 control behaviours (mod similarity: J = .125-.143)	5.3
→ victim selection (high similarity: J >.571)	
6 control behaviours (lower similarity: J = .100-.125)	3.7

	→ victim selection (high similarity: J >.571)	
7	control behaviours (mod similarity: J = .125-.143)	3.4
	→ victim selection (mod similarity: J = .222- .333)	
	→ sexual behaviours (lower similarity: $J \leq .000$)	

Figures

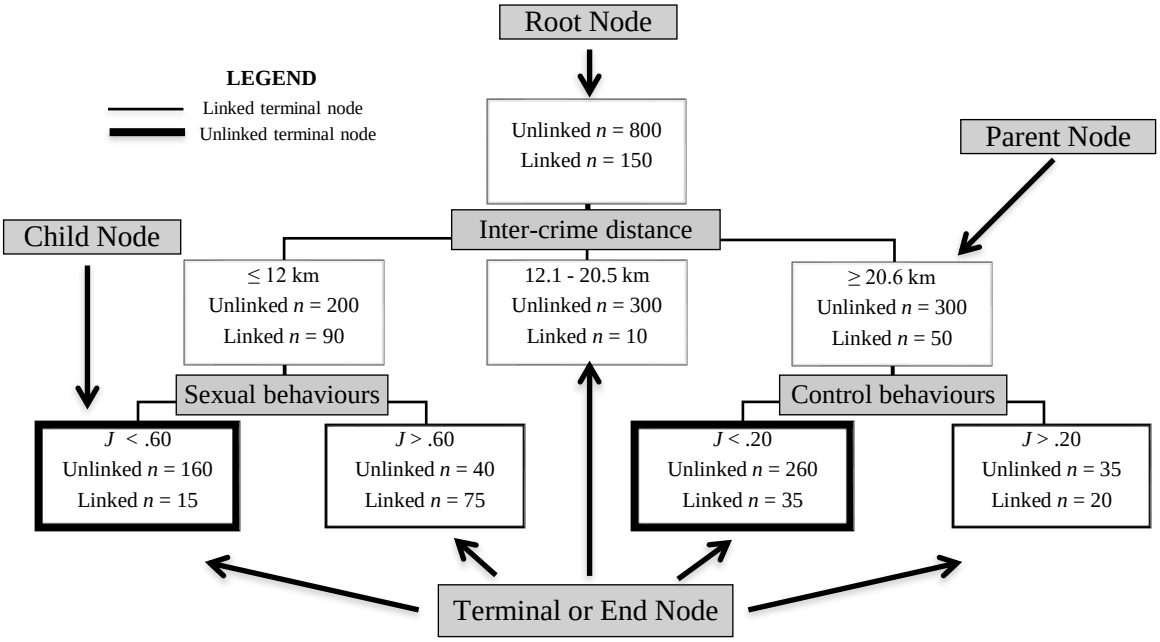

Figure 1. A hypothetical example of the structure of a standard sex offence CT with linked versus unlinked as the decision outcome (in the terminal nodes).

Note: J refers to Jaccard's coefficient, a commonly used measure of across-crime behavioural similarity in crime linking studies. Jaccard's coefficient varies from 0 to 1, with 0 indicating no similarity and 1 indicating total similarity.

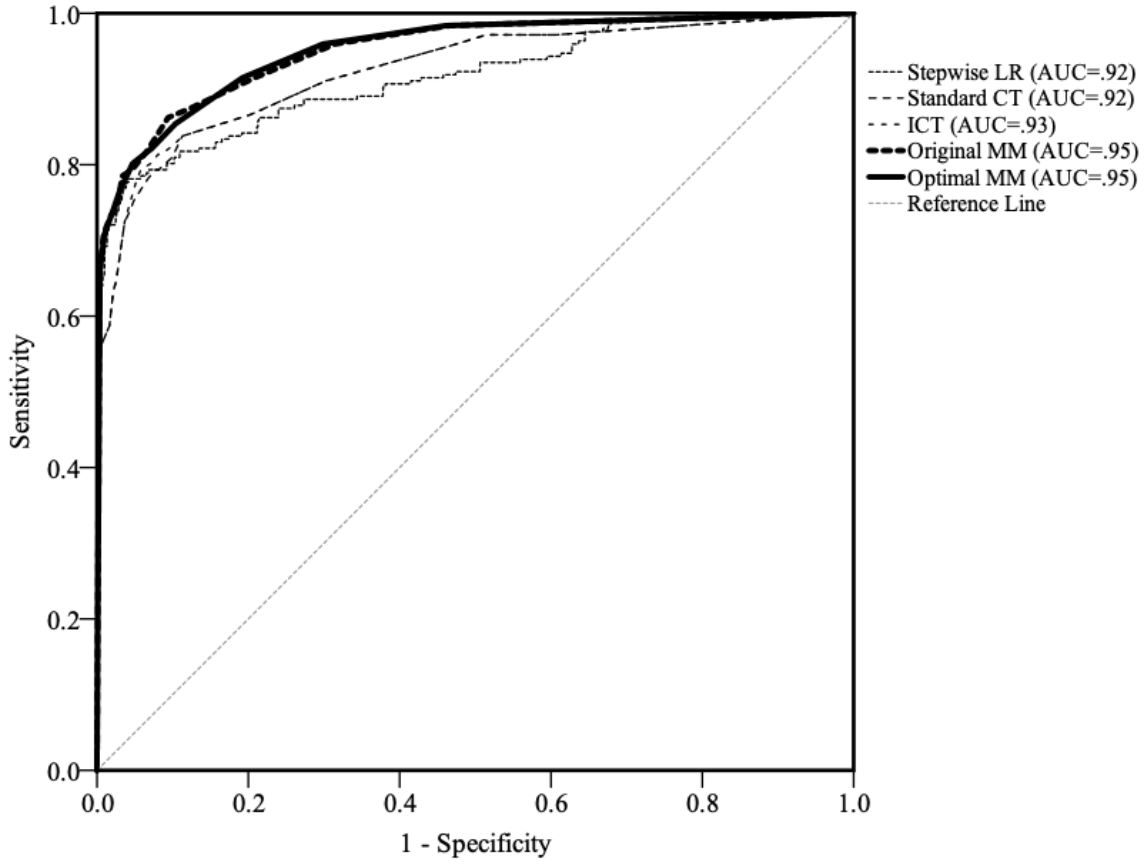

Figure 2. ROC curves indicating the performance of each linking model on the test sample.

Declarations

Funding: No source of funding was used to support this project.

Conflict of Interest: All authors declare that they have no conflicts of interest.