Received: 10 May 2023 | Revised: 18 July 2023 | Accepted: 19 July 2023

DOI: 10.1002/acp.4122

RESEARCH ARTICLE

WILEY

Receiver operating characteristic curves in the crime linkage context: Benefits, limitations, and recommendations

Logan Ewanation[1] | Craig Bennell[2] | Matthew Tonkin[3] | Pekka Santtila[4]

[1]Faculty of Social Science and Humanities, Ontario Tech, Oshawa, Canada

[2]Department of Psychology, Carleton University, Ottawa, Canada

[3]School of Criminology, University of Leicester, Leicester, UK

[4]NYU-ECNU Institute for Social Development at NYU, NYU Shanghai, Shanghai, China

Correspondence
Logan Ewanation, Faculty of Social Science and Humanities, Ontario Tech, 2000 Simcoe Street, North, Oshawa, Ontario L1G 0C5, Canada.
Email: logan.ewanation@ontariotechu.ca

Abstract

Deciding whether two crimes have been committed by the same offender or different offenders is an important investigative task. Crime linkage researchers commonly use receiver operating characteristic (ROC) analysis to assess the accuracy of linkage decisions. Accuracy metrics derived from ROC analysis—such as the area under the curve (AUC)—offer certain advantages, but also have limitations. This paper describes the benefits that crime linkage researchers attribute to the AUC. We also discuss several limitations in crime linkage papers that rely on the AUC. We end by presenting suggestions for researchers who use ROC analysis to report on crime linkage. These suggestions aim to enhance the information presented to readers, derive more meaningful conclusions from analyses, and propose more informed recommendations for practitioners involved in crime linkage tasks. Our reflections may also benefit researchers from other areas of psychology who use ROC analysis in a wide range of prediction tasks.

KEYWORDS

area under the curve, comparative case analysis, crime linkage, diagnostic accuracy, ROC analysis, serial crimes

1 | INTRODUCTION

Crime linkage analysis involves the task of determining whether a single perpetrator has committed two or more crimes (Bennell & Canter, 2002). Accurately linking crimes can improve police investigations in numerous ways. For instance, linking crimes can provide additional data for investigative techniques such as criminal and geographic profiling, while also allowing for more efficient use of police resources (Woodhams et al., 2007). Furthermore, prosecutors have used crime linkage analysis in criminal trials as inculpatory evidence to establish the guilt of an offender, and as an aggravating factor to help secure more appropriate sentences for serial offenders (Labuschagne, 2015).

Crimes committed by the same offender can typically be identified with a high degree of certainty using physical evidence, such as DNA or fingerprints. However, this sort of physical evidence is often unavailable or of too poor quality to be properly analyzed (Hazelwood & Warren, 2003). When physical evidence *is* collected, analysis can also be time-consuming and expensive, potentially leading to substantial backlogs (Davies, 1991). In contrast, crime linkage analysis can be carried out using behavioral, temporal, and/or geographic evidence that can potentially be collected at much lower cost and analyzed more quickly (Bennell et al., 2014).[i]

In recent years, there has been an increased focus on studying the processes underlying the crime linkage task, and in evaluating the degree to which it is possible to accurately link crimes using behavioral, temporal, and geographic information (Bennell et al., 2014). One of the most popular methods for assessing the degree to which this task can be accomplished accurately involves the use of receiver operating characteristic (ROC) analysis (e.g., Bennell & Canter, 2002;

Davidson & Petherick, 2020; Tonkin et al., 2011). In fact, since it was first used for this purpose in 2002 (Bennell, 2002), this method of analysis has been used in at least 32 other published studies.

In the present paper, we explain how ROC analysis has been used in the context of crime linkage research and describe some of the benefits of its use. We then move on to the two primary purposes of this paper. First, we highlight various problems with the way ROC analysis (and accuracy metrics that are derived from it) has often been used in this context. Second, we present methods that crime linkage researchers can use to mitigate these problems in future research. To help us illustrate points throughout the paper, we present analyses using a dataset of Finnish serial burglaries examined by Tonkin, Santtila, & Bull (2012) and Tonkin et al. (2019). which we describe in more detail below. While our focus in this paper is exclusively on crime linkage analysis, we hope that our reflections are also useful for researchers from other areas of psychology who use ROC analysis to examine a wide range of prediction tasks (e.g., diagnostic accuracy, risk assessment, recognition memory).

2 | USING ROC ANALYSIS TO STUDY THE CRIME LINKAGE TASK

Researchers have primarily examined the task of crime linkage in one of two ways. Both involve coding solved crimes for the presence of behaviors exhibited by offenders. One approach, which we will call the "series membership task", is to use this coded information, specifically the degree of behavioral similarity between crimes, to determine if it is possible to accurately predict whether a particular offense that is randomly selected from a database belongs to a known (i.e., solved) series of crimes included in that database (e.g., Santtila et al., 2004; Santtila et al., 2005; Santtila et al., 2008). Research using this approach has generally shown that crimes can be assigned to the correct series at a higher rate than what would be expected by chance. For instance, using a sample of 248 arson cases, Santtila et al. (2004) found that 33% of cases could be correctly linked to the series they belong to (while only 3% would have been expected by chance).

A second approach to studying crime linkage, and the one that will be focused on in this paper, has conceptualized the linkage task as one that requires a decision to be made about whether pairs of crimes are the work of the same offender or not (e.g., Bennell & Canter, 2002; Davidson & Petherick, 2020; Tonkin et al., 2011). We will call this approach the "pairwise linking task". Like other two-alternative (yes–no) type diagnostic decisions that must be made based on ambiguous evidence, the goals when using this approach are to identify what behaviors are best suited for distinguishing between crimes committed by the same offender versus different offenders, and to determine how similar two crimes should be before a decision is made that they have been committed by the same offender (i.e., establish an appropriate decision threshold; Bennell, 2005; Swets et al., 2000). Research has shown that it is possible to achieve both these goals and to accomplish the pairwise crime linkage task with a reasonable degree of accuracy (Bennell et al., 2014).

These different aspects of the pairwise crime linkage task are illustrated in Figure 1, where the probabilities of observing across-crime similarity scores for crime pairs committed by the same offender versus different offenders are graphed for a specific set of behaviors. In this case, across-crime similarity scores refer to scores derived from some type of similarity index, which is used to quantify the degree of similarity that exists across a crime pair. When relying on behavioral information, numerous similarity coefficients can potentially be used for this purpose (e.g., simple matching coefficient, taxonomic similarity index; Ellingwood et al., 2013; Melnyk et al., 2011), but Jaccard's coefficient is the most commonly used measure for crime linkage analysis (Bennell et al., 2014).[ii] When temporal or geographic information is being relied on, across-crime similarity can be established using simple measures of time differences or inter-crime distances. Similarity scores are typically calculated for all crime pairs in a sample, some of which will have been committed by the same offender. This is done in an exhaustive fashion usually with the help of specialized software that creates crime pairs from a sample of crimes, assigns each pair a code depending on whether the crimes have been committed by the same versus different offenders, and then calculates various across-crime similarity scores for each pair (e.g., B-LINK, which was created by the second author for this specific purpose; Bennell, 2002).

As is typical in crime linkage research, when across-crime similarity scores are calculated in this fashion, they tend to be larger, on average, for crime pairs committed by the same offender. Borrowing from work in other diagnostic fields such as radiology (e.g., Swets et al., 2000), it has been argued that the degree of overlap between these distributions indicates how useful the behaviors in question will be for discriminating between crimes committed by the same offender versus different offenders (i.e., the more overlap, the more difficult it will be; Bennell, 2005). For example, if the distributions overlap completely, it will be impossible for the similarity scores that gave rise to those distributions to be used for discriminatory purposes because every score is just as likely to be associated with crimes committed by

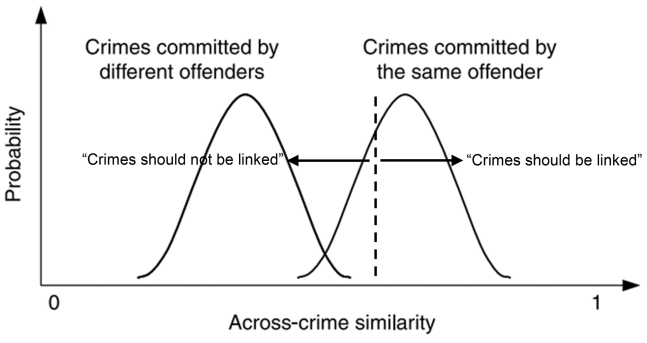

FIGURE 1 Hypothetical distributions of across-crime similarity scores for crimes committed by the same offender (the right distribution) versus different offenders (the left distribution). The *x*-axis represents the degree of similarity (from 0 to 1) between crime pairs and the *y*-axis represents the probability that a crime pair possesses any given degree of similarity.

the same offender as they are to be associated with crimes committed by different offenders.

Researchers have also argued that a threshold can be set anywhere along the x-axis in this figure (the dashed line in Figure 1) to indicate when a "linked decision" should be made for a particular pair of crimes, and that the decision outcomes resulting from possible thresholds can be examined to determine an "optimal" threshold (Bennell, 2005). More specifically, when conceptualizing crime linkage decisions in this way, there are four possible decision outcomes. These decision outcomes include: *hits* (i.e., two crimes committed by the same offender are correctly linked), *correct rejections* (i.e., two crimes committed by different offenders are correctly left unlinked), *false alarms* (i.e., two crimes committed by different offenders are incorrectly linked), and *misses* (i.e., two crimes committed by the same offender are incorrectly left unlinked).[iii] The probabilities of these decision outcomes depend on the ability of a given set of behaviors to discriminate between crimes committed by the same offender versus different offenders (i.e., distribution overlap), but also on the level of behavioral similarity used to decide when two crimes are similar enough to one another to warrant being linked (i.e., the placement of the decision threshold).

These issues can be modelled using ROC analysis (Bennell, 2005). As illustrated in Figure 2, ROC analysis plots the probabilities of hits (*p*H; also known as sensitivity) and false alarms (*p*FA; also known as 1–specificity) across a variety of decision thresholds, ranging from lenient thresholds (a low level of across-crime similarity is required to say that two crimes are linked) to strict thresholds (a high level of across-crime similarity is required to say that two crimes are linked).[iv] For each decision threshold, *p*H and *p*FA values are calculated to form a point in the ROC graph. Connecting these plotted points produces a ROC curve.

The area of the graph that lies underneath the curve (referred to as the area under the curve or the AUC) ranges from 0 (none of the graph falls under the curve) to 1 (all the graph falls under the curve) and is used as a measure of discrimination accuracy (Hajian-Tilaki, 2013). The less distribution overlap that exists in Figure 1, the better able one is to discriminate between the alternatives of interest, and the higher the AUC will be (more hits will be made relative to false alarms, across the possible decision thresholds). According to Swets (1988), AUCs between 0.50 and 0.70 "represent a rather low accuracy", values between 0.70 and 0.90 indicate values that "are useful for some purposes" (depending on context), and values above 0.90 indicate a "rather high accuracy" (p. 1292).[v] In studies that attempt to use behavioral information to discriminate between crime pairs that have been committed by the same offender versus different offenders, the range of reported AUCs is between 0.45 to 0.96 (Bennell et al., 2014).[vi]

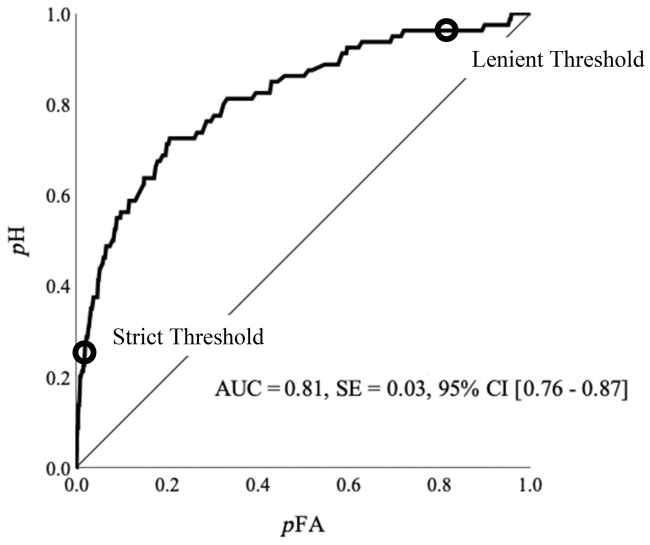

FIGURE 2 A receiver operating characteristic graph indicating the degree of linking accuracy associated with a set of serial burglary behaviors exhibited by Finnish serial burglars.

The utility of linkage decisions at any point along a ROC curve, which indicates performance at a particular threshold of across-crime similarity, can be captured by examining the *p*H to *p*FA ratio at that point. One can see from Figure 2 that, even for a set of distributions with a constant degree of overlap (i.e., a single ROC curve), *p*H and *p*FA vary widely and predictably as a function of the decision threshold. When a lenient threshold is set, low levels of across-crime behavioral similarity are required to decide that a crime pair was committed by the same offender, and *p*H and *p*FA will both be very high. In contrast, when a strict threshold is set, requiring a higher level of across-crime similarity to make an affirmative linkage decision, *p*H and *p*FA will both be very low. One of the challenges then in the pairwise crime linkage task is determining what threshold to use to find an appropriate balance between *p*H and *p*FA. We will discuss this challenge, and ways to potentially resolve it, below.

3 | POTENTIAL BENEFITS OF USING ROC ANALYSIS TO STUDY THE PAIRWISE CRIME LINKAGE TASK

Researchers who have used ROC analysis to examine the pairwise crime linkage task have highlighted numerous benefits associated with this approach (e.g., Bennell et al., 2009). The primary benefit is that ROC analysis provides a measure of linking accuracy (the AUC) that applies across different decision thresholds, rather than being specific to any single threshold, which may or may not result in desirable decisions (Bennell et al., 2009). Since the AUC is independent of any single threshold on a ROC curve, it provides an index of *overall* linkage performance, which is a more valid approach for assessing linkage accuracy than using *threshold specific* measures (e.g., percentage of correct decisions made when using a particular decision threshold).

For example, consider the approach adopted by Canter et al. (1991) who examined the pairwise crime linkage task. Based on an analysis of crime scene behaviors exhibited by serial rapists, they calculated across-crime behavioral similarity scores (ranging from 0 [no similarity] to 1 [total similarity]) for each crime pair. They then selected an arbitrary threshold of ≥0.30 to make linkage decisions (i.e., any crime pair associated with a similarity score ≥0.30 was deemed to be linked) and calculated how many correct decisions were

made. As discussed in more detail by Bennell et al. (2009), this approach is problematic because the accuracy estimate only applies to the specific threshold that Canter and his colleagues adopted, and that threshold may not be desirable. As an alternative, a ROC curve could have been generated and an AUC calculated by assessing decision outcomes for multiple thresholds. This would have provided an estimate of linkage accuracy using Canter et al.'s linking approach, independent of any specific decision threshold.

Another benefit associated with ROC analysis is that, while the AUC provides a measure of *overall* linkage accuracy, the various types of linkage decisions that can be made (i.e., hits, false alarms, misses, and correct rejections) are also still captured in the ROC graph for *every* decision threshold (i.e., the different points along the ROC curve; Bennell, 2005). Not only does this allow researchers and practitioners to understand the linkage approach they are using more fully, but it also provides the opportunity for them to identify decision thresholds that result in the desired balance between hits and false alarms. For example, threshold setting techniques have been developed that consider the base rate of the diagnostic alternative that the decision maker is interested in (crimes committed by the same offender in the current case), and the costs and benefits of the possible decision outcomes (Swets et al., 2000). Returning to the example of Canter et al.'s (1991) study, this approach could have been used to determine, for instance, that a threshold of ≥0.40 resulted in a more desirable balance between hits and false alarms, rather than their threshold of ≥0.30.

Unlike other metrics that could be used to assess crime linkage accuracy, such as correlations or odds ratios, another potential benefit of using ROC analysis that has been highlighted (especially by researchers outside of the policing context) is that the AUC is unaffected by base rates (e.g., the percentage of crime pairs committed by the same offender in a particular dataset; Douglas et al., 2012; Mossman, 1994; Rice & Harris, 1995). This is because a ROC curve (and the resulting AUC) is based on the *proportions* of the various decision outcomes, not their *frequency*. As such, the AUC can be used to compare linkage accuracy across datasets characterized by different base rates. For instance, serial burglars' crime series in police databases are often longer than those of serial rapists of the sort examined by Canter et al. (1991), which results in higher base rates of crime pairs committed by the same offender in serial burglary datasets (Bennell et al., 2014). However, this would not prevent researchers from comparing the linking method developed by Canter and his colleagues using these two datasets if ROC analysis was relied on.

Finally, ROC analysis is beneficial because it is very flexible (Bennell et al., 2009). For example, not only can the technique be used to compare the degree of crime linkage accuracy achieved across datasets that vary (e.g., by base rates), but this form of analysis can be used to examine how different analytical procedures perform on the pairwise crime linkage task. For instance, Tonkin et al. (2017) recently used ROC analysis to compare the linkage accuracy achieved by three different statistical techniques when they were applied to stranger sexual offence data from five different countries—logistic regression analysis, iterative classification tree analysis, and Bayesian analysis. Based on an analysis of AUCs, they found that each statistical approach was able to link crimes reasonably well, but that certain Bayesian procedures were particularly accurate. If a researcher wanted to extend this analysis to examine how Canter et al.'s (1991) simple linking approach compared to these more sophisticated analytical methods, ROC analysis could easily accommodate those types of comparisons as well.

4 | PROBLEMS WITH THE WAY ROC ANALYSIS HAS BEEN USED BY CRIME LINKAGE RESEARCHERS

While the potential benefits associated with ROC analysis have been described in numerous publications, less attention has been paid to potential problems with how this analysis has been used in the crime linkage context. That being said, certain issues have been discussed in other contexts (e.g., clinical psychology, medicine, radiology; see Halligan et al., 2015; Singh et al., 2013; Szmukler et al., 2012), and these issues also apply to the crime linkage context. For the remainder of this paper, we will highlight some of the problems that we have observed in some crime linkage studies and discuss ways that crime linkage researchers might mitigate these issues in future work.

As indicated previously, to help us illustrate our points, we will draw on a dataset of Finnish serial burglaries examined by Tonkin, Santtila, & Bull (2012) and Tonkin et al. (2019). The original dataset contains information related to 234 solved residential burglaries committed by 117 serial offenders in Finland between 1990 and 2001. For each crime, the geographical location is captured along with an estimated offence date, and behaviors related to the type of property burgled, the method of entry, and internal searches are available. The data we used for our analyses consisted of across-crime similarity scores ranging from 0 to 1 (based on Jaccard coefficients) for each behavioral domain, for every possible crime pair in the sample (80 crime pairs committed by the same offender and 12,640 committed by different offenders). The data also included an outcome variable for each crime pair, indicating whether the pair was actually committed by different offenders (0) or the same offender (1). For our purposes here, we use logistic regression analysis to make key points, with the across-crime similarity scores being used to predict the binary status of the crime pairs.

At the outset of this section, it is important to clarify two points to ensure readers understand what the purpose of our analysis is. Unlike other crime linkage studies, we are not examining the different types of behaviors exhibited by Finnish serial burglars to determine what the most effective predictors are for crime linkage analysis. We simply use statistical models derived from these behaviors to highlight potential problems with the way researchers studying these issues currently use ROC analysis and to illustrate methods for minimizing these problems. For readers interested in the kinds of geographic, temporal, and/or behavioral measures that have been used in crime linkage research, and their relative importance, we recommend reading the primary studies that explored the Finnish serial burglary

dataset (e.g., Tonkin et al., 2019; Tonkin, Santtila, & Bull, 2012; Tonkin, Woodhams, et al., 2012) and other reviews that have examined this topic (e.g., Bennell et al., 2014; Davies & Woodhams, 2019; Fox & Farrington, 2018).

The second issue that is important to discuss is our reliance on logistic regression analysis to fit our prediction models. We rely on logistic regression analysis because we believe it will be widely understood by most readers and because it has been the method of choice in most research examining crime linkage analysis to date (Bennell et al., 2014). Our choice to rely on this analysis should not be taken to mean that this is the only method that can be used for this purpose, or the best method. Researchers have explored a variety of analytical approaches to examine crime linkage analysis, including but not limited to, Bayesian analysis, cluster analysis, decision tree analysis, multidimensional scaling, neural networks, and sequence analysis (e.g., Bennell et al., 2021; Winter et al., 2013; Winter & Rossi, 2021). While we do not yet know which of these tools is best suited to the task, research is beginning to explore these issues (e.g., Tonkin et al., 2017).

4.1 | Definitional issues

One problem with how ROC analysis is used (and this goes beyond the crime linkage literature) relates to how the AUC is defined and interpreted (Douglas et al., 2012; Munro, 2004; Singh et al., 2013). Technically, the AUC is the percentage of times a randomly selected event (e.g., a cancerous tumor, a serious storm, a pair of crimes committed by the same offender) will have a higher (or lower, depending on the test) diagnostic score than a randomly selected non-event (e.g., a non-cancerous tumor, a benign weather pattern, a pair of crimes committed by different offenders). However, this is rarely the way in which the AUC is defined or interpreted in the research literature. For example, Singh and his colleagues (2013) conducted a systematic review exploring how predictive validity is measured in the field of forensic risk assessment. Although all the reviewed studies used ROC analysis and included the AUC as a measure of accuracy, only 34% of the studies defined and interpreted the AUC. Of these definitions and interpretations, the overwhelming majority were incorrect.

It is not unusual in published crime linkage studies for the AUC to also be left undefined (e.g., Bennell & Jones, 2005; Slater et al., 2015; Woodhams et al., 2019), and this opens up the opportunity for it to be misinterpreted. One of the most common misinterpretations of the AUC is believing that it reflects the percentage of cases where the actual outcome matches the predicted outcome (i.e., percent correct; Singh et al., 2013). Anecdotally, we have seen the AUC in the crime linkage context be interpreted in the same way, especially by police professionals. To illustrate that this interpretation is incorrect, we analyzed the Finnish burglary dataset.

We first conducted a logistic regression analysis using a combined similarity score (calculated using all behaviors in the dataset) as the predictor and linkage status (whether crime pairs were committed by the same offender versus different offenders) as the outcome, saving the model's predicted probabilities for each crime pair. We then constructed a ROC curve based on these probabilities, which was associated with a moderately high AUC (AUC = 0.81, SE = 0.03, 95% CI [0.76–0.87]; see Figure 2). Next, we selected a decision threshold along the scale of predicted probabilities that capped the probability of false alarms (*p*FA) at 0.05. This threshold equated to a probability level of ≥0.02. Applying this threshold to the Finnish data resulted in a "predicted linkage status" variable by coding crime pairs with predicted probability values ≥0.02 as "linked" and the remaining crime pairs as "unlinked."

Table 1 presents the contingency table comparing the predicted linkage status to the actual linkage status. Using data from this table, we can calculate the proportion of crime pairs that the model correctly predicted using the selected decision threshold (percent correct = [33 + 12,012]/[33 + 47 + 628 + 12,012] = 0.95). In other words, the model predicted the correct linkage status 95% of the time using the specified decision threshold, a value that clearly does not correspond with the model's overall AUC of 0.81.

The challenge with using the AUC value of 0.81 in practical contexts as a metric for *absolute* crime linkage accuracy (i.e., the level of linkage accuracy associated with a given set of crime scene behaviors) is that it has no practical meaning. Investigators do not randomly select crime pairs for analysis that have been committed by the same offender versus different offenders. Alternative approaches will be discussed below when we describe ways to deal with these definitional challenges.

4.2 | Focusing on accuracy metrics, not decision outcomes

A related problem we have observed in the published crime linkage literature is that researchers will often focus on accuracy metrics—usually the AUC—rather than focusing on decision outcomes. The challenge with this is that most metrics, by themselves, do not tell us anything about the actual frequency of specific decision outcomes (e.g., the frequency of hits or false alarms), which is arguably what police practitioners will be most concerned with. This is especially problematic for low base rate decision tasks (e.g., where the event of interest rarely occurs), like the pairwise crime linkage task. In this task, crime pairs committed by the same offender will be relatively rare,

TABLE 1 Contingency table comparing predicted and true linkage status.

	True linkage status		
Predicted linkage status	Linked	Unlinked	Total
Linked	33 (a)	628 (b)	661
Unlinked	47 (c)	12,012 (d)	12,059
Total	80	12,640	12,720

Note: AUC = 0.81; threshold ≥0.02.

unless the base rate is artificially manipulated. For low base rate tasks, accuracy metrics such as the AUC can give the illusion that desirable decisions will typically be made when that is not actually the case (McClelland, 2011). We discuss this issue in more detail below in the section on precision-recall graphs.

To illustrate, consider how the data presented above would play out in a real-life situation. The base rate of crime pairs committed by the same burglar in a particular jurisdiction is about 0.63%, which is not unheard of in studies that have examined the pairwise crime linkage task. Researchers in this jurisdiction have produced a logistic regression model, which has a moderately high AUC of 0.81. When using a predicted probability threshold of ≥0.02 the ROC curve associated with this regression model suggests that a moderate hit rate (41%) and a very low false alarm rate (5%) will be achieved. A particular pair of crimes comes to the attention of an investigator in this jurisdiction. They assess the across-crime similarity for the crime pair and subjects the resulting similarity score to the regression model. Based on the predicted probability produced by the model, they decide that the two crimes are likely to be committed by the same offender (i.e., the probability exceeded ≥0.02) and assumes their decision is likely to be accurate given the accuracy metric attached to the decision model (AUC = 0.81). However, what is the probability that the crimes they linked are actually the work of the same offender?

To answer this question, we can use the data from the Finnish dataset again. Recall the data from Table 1 where a threshold of ≥0.02 was applied to the predicted probabilities produced by the logistic regression model described above. The AUC for this data is 0.81, the base rate is 0.60% (80/12720), the hit rate is 41% (33/80), and the false alarm rate is 5% (628/12640)—the exact values from the previous paragraph. Now, consider the 661 linked decisions. Of those decisions, only 33 (5%) crime pairs are actually committed by the same offender. Thus, the likelihood that the linked crime pair in question in the previous paragraph (or any other linked crime pair) actually represents crimes committed by the same offender is 5%.

This low likelihood value is the direct result of the extremely low base rate (0.63%) in this sample of data; essentially, the regression model being used here is given many more opportunities (12,640 vs. 80) to make a false alarm than to make a miss. This example demonstrates that, despite having access to a highly accurate linkage approach with a reasonably high AUC and a moderate hit to false alarm ratio, many more false alarms than hits will be made when trying to detect low base-rate events (like crimes pairs committed by the same offender). To convey this information to practitioners so they can make more informed decisions, crime linkage researchers cannot simply rely on accuracy metrics. They must provide adequate information about how frequently hits, false alarms, misses, and correct rejections will be made.

4.3 | Providing AUC values, not full ROC curves

Another problem that frequently occurs in the crime linkage literature, including in our own research (e.g., Bennell & Jones, 2005; Ellingwood et al., 2013; Tonkin et al., 2017), is when AUC values are provided for a particular linking approach, but not the actual ROC curves that gave rise to those AUCs. Studies that only report an AUC (or some other, overall performance metric) are not providing a complete picture of the model's ability to accurately predict the status of crime pairs (e.g., Tonkin & Woodhams, 2017; Tonkin, Santtila, & Bull, 2012; Tonkin, Woodhams, et al., 2012).[vii] Indeed, models with ROC curves that differ in size and shape can produce similar AUC values (Dwyer, 1996) and therefore it is critical that the actual ROC curves are provided so that appropriate decisions can be made about how to approach the linkage task.

Consider Figure 3. These are two ROC curves generated from the Finnish serial burglary dataset. One of the curves represents the performance achieved when using across-crime similarity scores based on internal behaviors (e.g., how the offender searched the premises; AUC = 0.76, SE = 0.03, 95% CI [0.71–0.81]). The other curve represents the performance achieved when using across-crime similarity scores based on entry behaviors (e.g., how the offender accessed the premises; AUC = 0.72, SE = 0.03, 95% CI [0.66–0.79]). Based on the AUCs alone, one would presumably recommend that practitioners rely on internal behaviors to discriminate between crimes committed by the same offender versus different offenders, if only one of these behavioral domains can be relied on.

However, what if the desire is to keep the false alarm rate to a minimum when making linking decisions? Say, for example, that investigators do not want to exceed the *p*FA value of 0.05 discussed earlier (the dashed vertical line). In this case, one would likely change the recommendation, because under these conditions, the hit rate to false alarm rate ratio is better for entry behaviors. This occurs because the two ROC curves overlap, meaning that desirable decisions

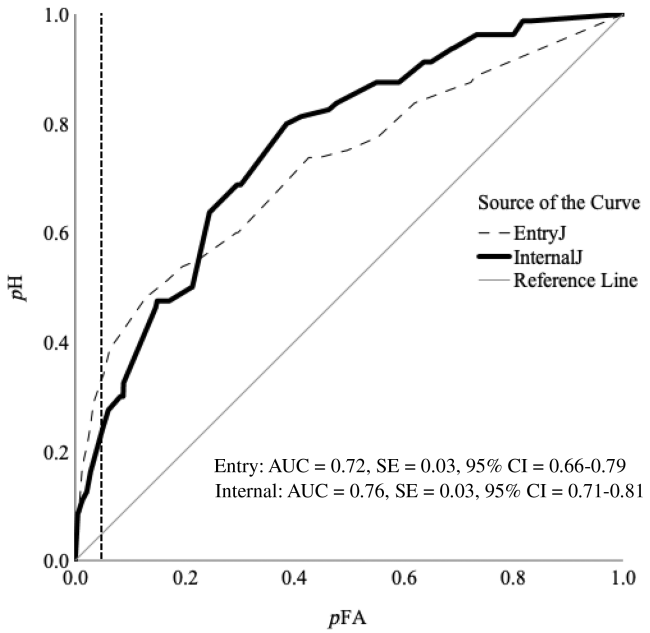

FIGURE 3 A receiver operating characteristic graph comparing the degree of linking accuracy for entry and internal search behaviors.

(i.e., making more hits than false alarms) shifts between the two behavioral domains depending on the specific conditions the practitioner wants to meet when making linkage decisions. This fact remains hidden if we only provide practitioners with the AUCs associated with the two ROC curves.

4.4 | Setting undesirable decision thresholds

The final problem we will discuss is the way in which decision thresholds are set by researchers. Again, this threshold is critically important because it determines when actual linkage decisions will be made (i.e., how similar two crimes need to be to one another to consider them "linked"). Currently, the most common method for establishing this threshold in the crime linkage context appears to involve methods for determining the point along the ROC curve that maximizes the probability of hits while minimizing the probability of false alarms (Bennell et al., 2009). This is often estimated by calculating Youden's index (i.e., sensitivity + specificity − 1; Youden, 1950; e.g., Pakkanen et al., 2020; Slater et al., 2015; Winter et al., 2013). While this is a rational approach, it is unlikely to be "optimal" in most investigative settings.

This is because the optimal decision threshold depends on several factors that are not captured using Youden's index, such as the base rate of the event of interest (i.e., crime pairs committed by the same offender) and the benefits and costs of the relevant decision outcomes (hits, misses, correct rejections, false alarms; Swets et al., 2000). When the rates of the diagnostic alternatives are equal, and the benefits and costs associated with the decision outcomes are the same, Youden's index will identify a threshold that is optimal. However, when these values deviate from this highly unlikely scenario, the optimal threshold will fall somewhere else along the ROC curve.

To illustrate a situation where base rate and cost/benefit considerations might matter, consider an approach to linking that involves two distinct stages, with each stage associated with their own ROC curve (Figure 4). The first stage consists of a crime analyst using an algorithm to search through a large database of crimes to identify potential linkages, like the Violent Crime Linkage Analysis System (ViCLAS; Collins et al., 1998). If potential linkages are found, the second stage involves the analyst informing the relevant investigators that they may have a serial offender operating in their jurisdictions, and that they should attempt to determine, using various investigative techniques, if the crimes in questions are indeed the work of the same offender.

Using this two-stage linking approach, the benefits and costs associated with decision outcomes might vary by stage. For example, the first stage might be treated like a sort of screening stage, where the goal is to ensure that any crime pairs committed by the same offender are captured, even if this means potentially capturing crime pairs committed by different offenders (much like a cancer screening test; see Carter et al., 2016). In this case, the benefit of making hits might outweigh the costs of making false alarms, and the appropriate decision threshold might be a relatively liberal one, as indicated in Figure 4. In contrast, greater caution may be warranted during the second stage, given that this stage will ultimately determine if an actual linkage decision is made, which in turn may influence if people will be arrested and charged. Here, the cost of a false alarm might be much greater than it was in stage one, and even greater than the benefit of a hit. This would suggest that a much stricter threshold, which reflects these facts, is appropriate. Methods for calibrating thresholds so they align with these sorts of considerations will be discussed below.

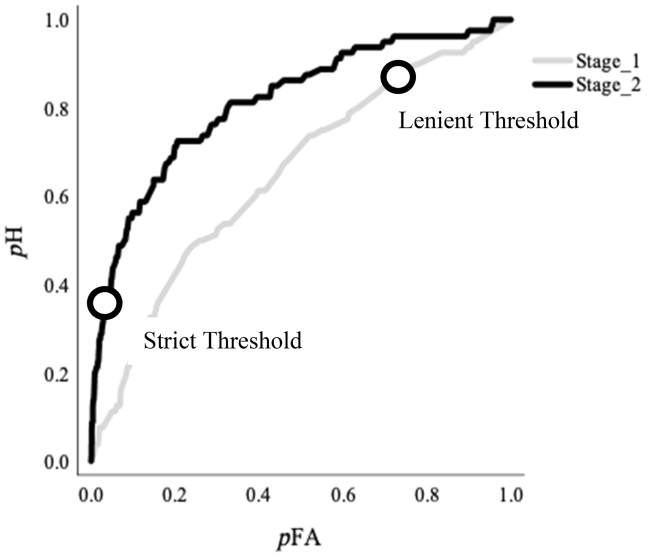

FIGURE 4 Hypothetical receiver operating characteristic curves for the two-stage linking approach, illustrating a lenient and strict decision threshold.

5 | METHODS FOR MINIMIZING THESE PROBLEMS

At best, the problems described above limit the value of crime linkage research that relies on ROC analysis. At worst, the problems mislead practitioners who consume this research in the hopes that it will improve their performance when they attempt to link crimes. Given this, it is important to consider how these problems can be eliminated, or at least minimized. Below, we describe several mitigation strategies. It should be noted that some of these strategies have been used by researchers already; we provide citations to some of this research in the sections below. However, in our view, the strategies are not being adopted as broadly as they should. We recommend that all researchers consider adopting these strategies when examining the accuracy of decisions in the pairwise crime linkage task.

5.1 | Define precisely what the AUC means and how it can be used

As described above, in the context of the pairwise crime linkage task, the AUC can be interpreted as the probability that a randomly

selected pair of crimes committed by the same offender will have a higher similarity score (i.e., be "more similar") than a randomly selected pair of crimes committed by different offenders (Bennell et al., 2014). In other words, if 100 randomly selected crime pairs committed by the same offender are compared to 100 randomly selected crime pairs committed by different offenders, an AUC of 0.81 would indicate that in 81% of those comparisons, the crime pair committed by the same offender will have a higher across-crime similarity score than the crime pairs committed by different offenders. To avoid confusion as to what the AUC actually means in crime linkage research, this definition should be provided anytime AUCs are presented, something that is rarely done at the moment. We believe that an explicit statement should also be included in these papers that the AUC is not equivalent to percent correct (i.e., the AUC *does not* indicate how many times a correct decision will be made when determining if a crime pair is or is not the work of the same offender).

It is also important to reinforce what the AUC can be used for. To us, the real value of the AUC is as a *relative* (as opposed to *absolute*) measure of linkage accuracy. Relative accuracy is important in the crime linkage context, and the AUC provides a useful metric for comparative purposes. For example, when studying linkage methods for serial burglary, it is important to know whether some types of behaviors (e.g., inter-crime distances) are more accurate predictors of linkage status than others (e.g., similarity of entry behaviors). The AUC can provide an answer to this question. It is also important to determine if some types of crime (serial burglary) can be linked more accurately than other types (serial rape). Again, the AUC can provide an answer to this question. Based on relative AUCs, theories of crime linkage can also be developed and tested, such as the idea that personal versus situational influences should be a key consideration when attempting to link crimes (e.g., behaviors that are personally controlled, such as inter-crime distance, might be more accurate predictors of linkage status than those that are situationally determined, such as property stolen).

5.2 | Present contingency tables

Given that the appropriate interpretation of the AUC may have little meaning in practical contexts, and that the AUC in isolation tells us little about the frequency of decision outcomes in the crime linkage task, researchers should provide contingency tables for specific decision thresholds that compare linkage predictions to actual linkage status (such as Table 1). Some recent crime linkage papers have included contingency tables, although only for one specific threshold (e.g., Tonkin et al., 2017; Woodhams et al., 2019). Providing contingency tables for multiple thresholds, for example a lenient, mid-range, and strict threshold, could provide a more complete and accurate picture of how a linkage method might actually perform under varying decision-making conditions if it were implemented in the field. This will be particularly important in cases where the base rate of crimes committed by the same offender is low.

5.3 | Provide full ROC curves, not just AUCs

Researchers also need to provide graphical representations of full ROC curves, rather than simply reporting the AUC. Many published crime linkage studies do this (e.g., Slater et al., 2015; Winter et al., 2013; Woodhams et al., 2019), but not all. As mentioned above, providing a full ROC curve will allow readers to gain a deeper understanding of how the linkage method under investigation performs across various decisions thresholds. This will prevent situations like those illustrated in Figure 3, where preference may be given to a particular linkage method due to it having a higher overall AUC, when in fact another linkage method may be preferable (i.e., results in a better ratio of hits to false alarms) for a range of decision thresholds that are more practically relevant.[viii]

5.4 | Identify limits of threshold setting methods and use valid approaches when possible

Researchers need to be more cautious about using threshold setting methods that make invalid assumptions about base rates and the benefits and costs associated with decision outcomes (e.g., using Youden's index or simply choosing the point on the ROC curve that falls closest to the upper left corner of the ROC graph). It is unlikely, given the realities within investigative contexts, that such approaches result in "optimal" decision thresholds and researchers need to clearly articulate the limits of their threshold setting methods when speaking about these issues. We also recommend that researchers: (1) not refer to thresholds as optimal (a mistake we have made in our own research; for example, Bennell & Jones, 2005) unless the work has been done to ensure this is the case, (2) highlight the issues that were unable to be considered when establishing decision thresholds (e.g., base rates, benefits/costs of decision outcomes), and (3) speak directly to the implications of these omissions (e.g., that if the benefits/costs are not the same across various decision outcomes, the threshold will produce an undesirable ratio of such outcomes).

Effort should also be invested over the longer term to study various other threshold setting approaches that might be useful in the crime linkage context. Ideally, researchers would be able to determine the precise base rates they are working with (i.e., the probability of encountering crime pairs committed by the same offender and different offenders), and the costs and benefits associated with the various decision outcomes. If this can be done, methods exist to combine these values to identify desirable thresholds (see Swets et al., 2000). We provide the formula here for the crime linkage scenario:

$$\text{Desirable threshold} = \frac{p(\text{crime pairs committed by different offenders})}{p(\text{crime pairs committed by the same offender})} \times \frac{\text{benefit}(\text{correct rejection}) + \text{cost}(\text{false alarm})}{\text{benefit}(\text{hit}) + \text{cost}(\text{miss})}.$$

The challenge with this approach of course is that these estimates are incredibly difficult to calculate (Swets, 1992). For example, the fact that many crimes go unsolved and unreported will obviously complicate any attempts to establish accurate base rates, and while costs

and benefits related to crime linkage decisions may be easier to estimate for some variables (e.g., costs and benefits related to finances), other costs and benefits will be very challenging to quantify (e.g., costs and benefits related to human suffering).

A slightly more realistic approach may be to consider ratios for these parameters instead of precise estimates (Swets et al., 2000). For example, it may not be possible to determine precisely what the costs and benefits are of the various types of crime linkage decisions, but it may be possible to determine (e.g., through carefully run focus groups) that investigators are twice as interested in being correct when crimes have been committed by the same offender then when crimes have been committed by different offenders. In this case, we could use the following formula:

$$\text{Optimal threshold} = \frac{0.50}{0.50} \times \frac{1}{2}.$$

It is also possible to establish limits on *p*H or *p*FA so that a minimum *p*H is achieved or a maximum *p*FA is not surpassed, and to set thresholds that meet these pre-determined limits. This may be the simplest approach of those that have been discussed. An example of this was discussed above, where a limit of *p*FA ≤ 0.05 was set. Tonkin et al. (2017) also adopted this approach in their study comparing the ability of different statistical approaches for linking stranger sexual assaults. In that research, the crime linkage practitioners involved in the study set the threshold themselves at *p*FA ≤ 0.15.

Providing practical advice on how to proceed with this topic goes beyond the scope of this article. Interested readers will want to review Swets (1992) where he thoroughly discusses the issue of establishing decision thresholds to improve the utility of decisions in high-stakes diagnostics. In that work, Swets also cites multiple examples from fields where serious attempts have been made to navigate the challenges encountered when setting sensible thresholds.

5.5 | Provide other metrics of performance to complement the AUC

We also recommend that crime linkage researchers consider presenting other performance metrics, in addition to the AUC and the frequency of decision outcomes, to provide readers with more information about the performance of the crime linkage method(s) under investigation. The metrics we propose are all commonly used in other fields and can be calculated easily from a contingency table. To illustrate calculations in the following sections, we rely on the contingency table in Table 1, particularly the notations *a*, *b*, *c*, and *d*, which will be referred to in formulae we outline.

5.5.1 | Diagnostic effectiveness and misclassification rate

Perhaps most obviously, diagnostic effectiveness (DE) and the misclassification rate (MR) can be provided to supplement AUCs. In the crime linkage context, DE is simply the proportion of cases that are correctly categorized by the linkage method under investigation when using a particular decision threshold (i.e., the proportion of crime pairs that were correctly predicted as being linked or unlinked). DE can be calculated from Table 1 as $(a + d)/(a + b + c + d)$. The MR is the inverse of DE, or the proportion of cases that are incorrectly predicted. The MR can be calculated from Table 1 as $(b + c)/(a + b + c + d)$. Like predictive values, which we describe next, the DE and MR are directly affected by the base rate of the event in question (Shaikh, 2011). They are therefore less valuable when the aim is to compare accuracy in settings where base rates vary (e.g., comparing linkage accuracy across different crime types or police jurisdictions).

Using Table 1, the DE is 0.94 and the MR is 0.05. In other words, using the particular threshold examined here, the linkage method under investigation correctly classified the status of crime pairs 94% of the time, and failed to correctly classify the status of crime pairs 5% of the time. Like the AUC, however, the DE and MR do not distinguish between correctly predicted crime pairs that are actually the work of the same offender versus correctly predicted crime pairs that are actually the work of different offenders. While still providing valuable information about the accuracy of crime linkage decisions, other metrics that provide this information, such as positive predictive value and negative predictive value, should also be reported.

5.5.2 | Positive and negative predictive values

Positive and negative predictive values (PPV and NPV, respectively) can be useful in the crime linkage context, although they are not relied on in this literature yet. In the crime linkage context, the PPV would indicate the percentage of crime pairs predicted to be linked that have in fact been committed by the same offender (Altman & Bland, 1994). In other words, the PPV provides information about the likelihood that a positive prediction truly means the crimes are the work of the same person (Chu, 1999). With reference to Table 1, researchers can simply calculate the PPV as $a/(a + b)$ (Blakely & Salmond, 2002). In comparison, the NPV indicates the percentage of crime pairs predicted to be unlinked that are indeed committed by different offenders. Again, the calculation for the NPV is quite straightforward: $d/(c + d)$.

As indicated above, unlike the AUC, predictive values incorporate a particular sample's base rates into their calculation (Shaikh, 2011). When the event of interest is relatively rare (as will be the case when carrying out pairwise crime linkage), the PPV will always remain low, even if the AUC itself is high (Altman & Bland, 1994). Thus, just like with DE and MR, because of their dependence on base rates, predictive values calculated from a particular sample should not be compared across samples (Parikh et al., 2008). However, predictive values provide specific information about a model's performance on a particular sample, and they include information that is absent when only presenting the AUC (Akobeng, 2007).

For example, a low PPV in the crime linkage context would indicate a high number of false alarms. Drawing on the data from Table 1, the PPV can be calculated as 0.05, suggesting that among crimes predicted to be linked, the probability of the crimes actually being linked is less than 5%. In comparison, the NPV for the model that produced the data included in Table 1 is 0.99, which indicates that among crimes predicted to be unlinked, there is a 99% likelihood of them being committed by different offenders. Again, this is information that is simply unavailable from the AUC that this data produced (0.81) and provides valuable information (for researchers and practitioners alike) about the performance of the model, especially if the goal is to correctly link crimes committed by the same offender.[ix]

5.5.3 | Likelihood ratios

Along with PPV and NPV, likelihood ratios (LRs) can also be used to provide important information about the methods used to predict whether crimes pairs have been committed by the same offender or different offenders. In the context of crime linkage analysis, the LR indicates the likelihood that a given "test result" (e.g., a predicted probability from a logistic regression model that exceeds a specific threshold, triggering a "linked" decision) would be expected for a crime pair that was committed by the same offender compared to the likelihood of that the same test result would be expected for a crime pair that was committed by different offenders.

Likelihood ratios can be presented for a positive test result (LR+) and a negative test result (LR−). LR+ is equal to the probability that a crimes pair committed by the same offender receives a positive test score (i.e., a probability score that triggers a "linked" decision) over the probability that a crime pair committed by different offenders receives a positive test score (i.e., LR + =pH/pFA). Using the data from Table 1, LR+ is calculated as $[a/(a + c)]/[b/(b + d)]$. LR− is equal to the probability that a crime pair committed by the same offender receives a negative test score (i.e., a probability score that triggers an "unlinked" decision) over the probability that a crime pair committed by different offenders receives a negative test score (i.e., LR− = pM/pCR). As per Table 1, LR− is calculated as $[c/(a + c)]/[d/(b + d)]$.

Using Table 1 from the Finnish dataset, LR+ is 8.3, while LR− is 0.62. What do these values mean? It is generally understood that a predictive test with a LR of 1.0 is not useful: there is no difference in the probability of a given "test result" between crimes pairs that have been committed by the same offender versus different offenders (i.e., the linkage method under investigation has no discriminatory value). Essentially, linkage methods with very high LR+ values and very low LR− values have the greatest discriminatory value. In medical settings, LR+ values above 10 and LR− values below 0.1 generally considered "good" (Grimes & Schulz, 2005). Thus, according to those cut-offs, the LR values are not as high (or as low) as one would hope.[x] That being said, these clinical guidelines might not be relevant in contexts where crime linkage is conducted, for example where crime linkage is used to informally guide a police investigator.

5.6 | Use precision-recall graphs

A slightly more radical proposal is for crime linkage researchers to supplement ROC graphs with precision-recall (PR) graphs, something that has never been done in this context as far as we are aware, despite such graphs being commonly used in other settings (e.g., information retrieval) and being easy to construct using widely available statistical analysis software like SPSS and R. These graphs are likely to be more informative in cases where extremely low base rates are an issue—in other words, when there is a significant *imbalance* between the diagnostic alternatives (Saito & Rehmsmeier, 2015)—and are therefore well-suited to the pairwise crime linkage task. In this task, researchers will always encounter many more crime pairs committed by different offenders compared to the same offender, and good performance will always be attributable largely to the high number of correct rejections that will be made.

Instead of plotting the false alarm rates on the x-axis and hit rates on the y-axis, like we do with ROC analysis, PR graphs plot hit rates (recall) on the x-axis and PPV (precision) on the y-axis.[xi] Referring back to the formula for these parameters, one can see that the (inevitably large number of) correct rejections made are not considered at all when calculating the values plotted on a PR graph, but they are when constructing a ROC graph. Another key difference between the two types of graphs is the reference point that is used to judge performance. Whereas the positive diagonal (AUC = 0.50) is used in ROC analysis to assess the degree to which the performance of a classification model varies from chance (random) performance, random performance in PR graphs is determined by the base rate in the dataset under examination (i.e., positive cases/[positive cases+ negative cases]) and is represented as a horizontal line at that point on the graph. Just like in ROC analysis, the ideal classifier will be represented by a curve on a PR graph that is higher than this reference line, but instead of the top left corner representing optimal performance, which was the case with ROC analysis (hit rate = 1, false alarm rate = 0), the optimal operating point on the PR graph is the top right corner (precision and recall = 1). Basically, a PR curve that falls along the upper and right axis of the graph indicates perfect performance.

For illustrative purposes, we provide a PR plot in Figure 5 that was generated from the logistic regression model developed for the Finnish burglary dataset we described above (predicting linkage status from a combined across-crime similarity score using all behaviors in the dataset). Recall is plotted on the x-axis and precision on the y-axis, across various decision thresholds. The base rate of 0.6% is reflected on the graph by the horizontal line. A number of interesting things can be inferred from this graph. First, despite the high AUC associated with this logistic regression model (AUC = 0.81), the PR curve clearly does not reflect near perfect performance. This suggests that while the logistic regression model can generally classify crimes committed by the same offender versus different offenders into linked and unlinked crime categories reasonably well (not a surprise when nearly all of the data are crime pairs committed by different offenders), high

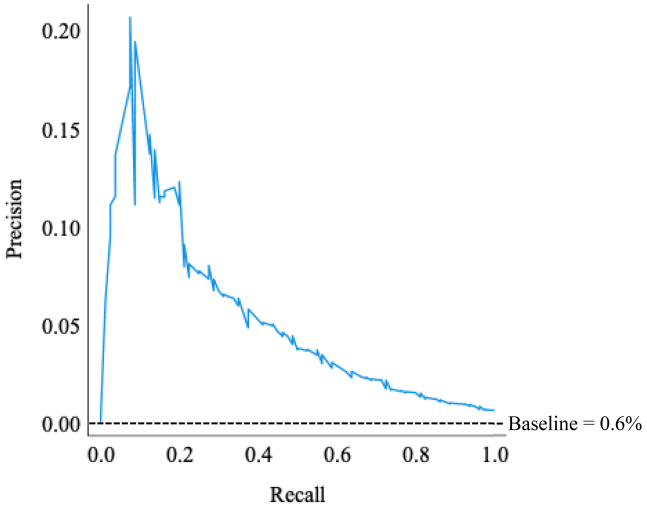

FIGURE 5 An example precision-recall plot from a logistic regression model generated from the Finnish serial burglary Dataset.

similarity scores do not correlate well with crimes committed by the same offender. Second, despite not performing very well, the logistic regression model is clearly not performing at chance levels with respect to recall and precision. Recall that while the performance floor is AUC = 0.50 for ROC graphs, the performance floor is the positive base rate in our dataset for the PR graph (0.6%), and the model is obviously exceeding that. Finally, we can draw specific conclusions about recall and precision at various decision thresholds. For example, the model can achieve relatively high levels of recall, but precision levels remain consistently low regardless of what threshold we use. This suggests that we need to be cautious when the regression model is classifying cases as linked because the false alarm rate is higher than we would like (low precision). However, depending on the threshold we use, we can be fairly confident that we are not missing many linkages when crimes have in fact been committed by the same offender (high recall).

6 | CONCLUSION

Crime linkage decisions are important. To advance knowledge in this area we need a method for assessing the accuracy of linkage decisions. ROC analysis, and the AUC specifically, has been used for this purpose in many studies. The AUC is associated with numerous benefits, but it also has limitations. Researchers need to recognize and mitigate these limitations. We hope that the recommendations outlined above help researchers do this. We believe that if the recommendations are followed, it will make the resulting research more useful and will provide practitioners with the sort of information they need to make more informed decisions when they attempt to link serial crimes. Given that ROC analysis is used in many areas of psychology to examine a wide range of prediction tasks, we hope that our recommendations benefit those outside of the crime linkage context as well.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Research data are not able to be shared.

ORCID

Logan Ewanation https://orcid.org/0000-0002-4174-9099

ENDNOTES

[i] Of course, we recognize that there can be delays with this form of crime linkage analysis as well. For example, there can be substantial delays in analysts receiving case files from investigators and inputting data from those files into databases that are used to assist with the crime linkage task.

[ii] When calculating across-crime similarity for a pair of crimes, Jaccard's coefficient ranges from 0 to 1 and is calculated as $a/(a + b + c)$, where "*a*" refers to the frequency of behaviours present in both crimes, and "*b*" and "*c*" refer to the frequency of behaviours present in one crime but absent in the other. Joint non-occurrences in behaviour are not taken into account when using Jaccard's coefficient, which may be appropriate given that joint non-occurrences in behaviour may not imply higher levels of behavioural similarity. This is because behaviours may be absent from two crimes committed by the same offender for various reasons, only one of which is the offender deciding not to exhibit the behaviours (e.g., victims failing to report the presence of such behaviours).

[iii] Throughout this paper, we have opted to use the terms hits, correct rejections, false alarms, and misses rather than the synonymous terms true positives, true negatives, false positives, and false negatives, which are often used in contexts where ROC analysis is conducted. This is because the former terms are relied on almost exclusively in the crime linkage literature. For the same reason, when referring to the axes of ROC graphs, we use the terms hit rate (pH; for the *y*-axis) and false alarm rate (*p*FA; for the *x*-axis) rather than the synonymous terms sensitivity and 1-specificity, which are sometimes used by other researchers.

[iv] The probability of misses, *p*M, and correction rejections, *p*CR, are also illustrated on ROC graphs, but usually not labelled. They fall on the opposite axes to *p*H and *p*FA, respectively.

[v] Other guidelines also exist. For instance, Hosmer and Lemeshow (2000) suggest that an AUC between 0.70 and 0.80 indicates acceptable discrimination, AUCs between 0.80 and 0.90 indicate excellent discrimination, and AUCs above 0.90 indicate outstanding discrimination.

[vi] Note that a number of crime linkage studies using ROC analysis have been published since this time (e.g., Pakkanen et al., 2020; Tonkin et al., 2017; Woodhams et al., 2019), but the range reported by Bennell et al. (2014) appears to still be accurate.

[vii] It should be noted that sometimes this may not have been the fault of the authors. For instance, some journals may prohibit the inclusion of several figures in a single article. The use of supplemental and online appendices will help with this issue.

[viii] This situation may be more common under certain conditions, such as when drawing on small samples or using data that generates only a limited number of across-crime similarity scores (e.g., Jaccard coefficient values). Under these circumstances, ROC curves are unlikely to be smooth because there are less points on the curve to connect. The jagged nature of these curves can result in increased overlap between the curves.

[ix] Of course, if the goal is something other than making binary yes–no type crime linkage decisions, such as using a statistical algorithm to generate a prioritized list of potentially linked crimes for investigators to investigate more thoroughly, then this argument may not apply.

[x] An LR+ value of 8.3 means there is an 8.3-fold increase in the odds of a crime pair committed by the same offender having a positive "test result". An LR value of 0.62 means there is s 1.61-fold decrease (1/0.62) in the odds of a crime pair committed by the same offender having a negative "test result".

[xi] Recall reflects the ability to detect positive cases; in other words, it reflects the ability of a model/decision-maker to accurately link crimes committed by the same offender. Precision reflects the credibility of a claim that a case is positive; in other words, it reflects the degree to which we should have confidence that a crime pair is actually committed by the same offender when a model/decision-maker suggests it should be linked.

REFERENCES

Akobeng, A. K. (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatrica*, *96*(5), 644–647. https://doi.org/10.1111/j.1651-2227.2006.00178.x

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal*, *308*(6943), 1552.

Bennell, C. (2002). *Behavioural consistency and discrimination in serial burglary (unpublished doctoral thesis)*. University of Liverpool.

Bennell, C. (2005). Improving police decision making: General principles and practical applications of receiver operating characteristic analysis. *Applied Cognitive Psychology*, *19*(9), 1157–1175. https://doi.org/10.1002/acp.1152

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. *Science & Justice*, *42*(3), 153–164. https://doi.org/10.1016/S1355-0306(02)71820-0

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. https://doi.org/10.1002/jip.21

Bennell, C., Jones, N. J., & Melnyk, T. (2009). Addressing problems with traditional crime linking methods using receiver operating characteristic analysis. *Legal and Criminological Psychology*, *14*(2), 293–310. https://doi.org/10.1348/135532508X349336

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*(1), 29–56. https://doi.org/10.1002/jip.1395

Bennell, C., Mugford, R., Woodhams, J., Beauregard, E., & Blaskovits, B. (2021). Linking serial sex offences using standard, iterative, and multiple classification trees. *Journal of Police and Criminal Psychology*, *36*, 691–705. https://doi.org/10.1007/s11896-021-09483-6

Blakely, T., & Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology*, *31*(6), 1246–1252. https://doi.org/10.1093/ije/31.6.1246

Canter, D., Heritage, R., Wilson, M., Davies, A., Kirby, S., Holden, R., McGinley, J., Hughes, H., Larkin, P., Martin, L., Tsang, E., Vaughan, G., & Donald, I. (1991). *A facet approach to offender profiling* (Vol. 1). University of Surrey.

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, *159*(6), 1638–1645. https://doi.org/10.1016/j.surg.2015.12.029

Chu, K. (1999). An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine*, *11*(3), 175–181.

Collins, P. I., Johnson, G. F., Choy, A., Davidson, K. T., & Mackay, R. E. (1998). Advances in violent crime analysis and law enforcement: The Canadian violent crime linkage analysis system. *Journal of Government Information*, *25*(3), 277–284. https://doi.org/10.1016/S1352-0237(98)00008-2

Davidson, S., & Petherick, W. (2020). Case linkage in Australian serial stranger rape. *Journal of Criminological Research, Policy and Practice*, *7*(1), 4–17. https://doi.org/10.1108/JCRPP-01-2020-0016

Davies, A. (1991). The use of DNA profiling and behavioural science in the investigation of sexual offences. *Medicine, Science and the Law*, *31*(2), 95–101. https://doi.org/10.1177/0025802491031002

Davies, K., & Woodhams, J. (2019). The practice of crime linkage: A review of the literature. *Journal of Investigative Psychology and Offender Profiling*, *16*(3), 169–200.

Douglas, K. S., Otto, R. K., Desmarais, S. L., & Borum, R. (2012). Clinical forensic psychology. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology, volume 2: Research methods in psychology* (pp. 213–244). John Wiley & Sons.

Dwyer, A. J. (1996). In pursuit of a piece of the ROC. *Radiology*, *201*(3), 621–625. https://doi.org/10.1148/radiology.201.3.8939207

Ellingwood, H., Mugford, R., Bennell, C., Melnyk, T., & Fritzon, K. (2013). Examining the role of similarity coefficients and the value of behavioural themes in attempts to link serial arson offences. *Journal of Investigative Psychology and Offender Profiling*, *10*(1), 1–27. https://doi.org/10.1002/jip.1364

Fox, B., & Farrington, D. P. (2018). What have we learned from offender profiling? A systematic review and meta-analysis of 40 years of research. *Psychological Bulletin*, *144*(12), 1247–1274. https://doi.org/10.1037/bul0000170

Grimes, D. A., & Schulz, K. F. (2005). Refining clinical diagnosis with likelihood ratios. *The Lancet*, *365*(9469), 1500–1505. https://doi.org/10.1016/S0140-6736(05)66422-7

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

Halligan, S., Altman, D. G., & Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology*, *25*(4), 932–939. https://doi.org/10.1007/s00330-014-3487-0

Hazelwood, R. R., & Warren, J. I. (2003). Linkage analysis: Modus operandi, ritual, and signature in serial sexual crime. *Aggression and Violent Behavior*, *8*(6), 587–598. https://doi.org/10.1016/S1359-1789(02)00106-4

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley.

Labuschagne, G. N. (2015). The use of linkage analysis evidence in serial offense trials. In J. Woodhams & C. Bennell (Eds.), *Crime linkage: Theory, research, and practice* (pp. 197–224). CRC Press.

McClelland, G. (2011). Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence analysis: Behavioral and social scientific oundations* (pp. 83–100). National Academies Press.

Melnyk, T., Bennell, C., Gauthier, D. J., & Gauthier, D. (2011). Another look at across-crime similarity coefficients for use in behavioural linkage analysis: An attempt to replicate Woodhams, Grant, and Price (2007). *Psychology, Crime & Law*, *17*(4), 359–380. https://doi.org/10.1080/10683160903273188

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*(4), 783–792. https://doi.org/10.1037/0022-006X.62.4.783

Munro, E. (2004). A simpler way to understand the results of risk assessment instruments. *Children and Youth Services Review*, *26*(9), 873–883. https://doi.org/10.1016/j.childyouth.2004.02.026

Pakkanen, T., Sirén, J., Zappalà, A., Jern, P., Bosco, D., Berti, A., & Santtila, P. (2020). Linking serial homicide—Towards an ecologically valid application. *Journal of Criminological Research, Policy and Practice*, *7*(1), 18–33. https://doi.org/10.1108/JCRPP-01-2020-0018

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50. https://doi.org/10.4103/0301-4738.37595

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*(5), 737–748. https://doi.org/10.1037/0022-006X.63.5.737

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Santtila, P., Fritzon, K., & Tamelander, A. L. (2004). Linking arson incidents on the basis of crime scene behavior. *Journal of Police and Criminal Psychology*, *19*(1), 1–16. https://doi.org/10.1007/BF02802570

Santtila, P., Junkkila, J., & Sandnabba, N. K. (2005). Behavioural linking of stranger rapes. *Journal of Investigative Psychology and Offender Profiling*, *2*(2), 87–103. https://doi.org/10.1002/jip.26

Santtila, P., Pakkanen, T., Zappala, A., Bosco, D., Valkama, M., & Mokros, A. (2008). Behavioural crime linking in serial homicide. *Psychology, Crime & Law*, *14*(3), 245–265. https://doi.org/10.1080/10683160701739679

Shaikh, S. A. (2011). Measures derived from a 2×2 table for an accuracy of a diagnostic test. *Journal of Biometrics and Biostatistics*, *2*(128), 1–4.

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law*, *31*(1), 55–73. https://doi.org/10.1002/bsl.2053

Slater, C., Woodhams, J., & Hamilton-Giachritsis, C. (2015). Testing the assumptions of crime linkage with stranger sex offenses: A more ecologically-valid study. *Journal of Police and Criminal Psychology*, *30*(4), 261–273. https://doi.org/10.1007/s11896-014-9160-3

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

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*(4), 522–532. https://doi.org/10.1037/0003-066X.47.4.522

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. https://doi.org/10.1111/1529-1006.001

Szmukler, G., Everitt, B., & Leese, M. (2012). Risk assessment and receiver operating characteristic curves. *Psychological Medicine*, *42*(5), 895–898. https://doi.org/10.1017/S003329171100208X

Tonkin, M., Lemeire, J., Santtila, P., & Winter, J. M. (2019). Linking property crime using offender crime scene behaviour: A comparison of methods. *Journal of Investigative Psychology and Offender Profiling*, *16*(2), 75–90. https://doi.org/10.1002/jip.1525

Tonkin, M., Pakkanen, T., Sirén, J., Bennell, C., Woodhams, J., Burrell, A., Imre, H., Winter, J. M., Lam, E., ten Brinke, G., Webb, M., Labuschagne, G. N., Ashmore-Hills, L., van der Kemp, J. J., Lipponen, S., Rainbow, L., Salfati, C. G., & 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. https://doi.org/10.1016/j.jcrimjus.2017.04.002

Tonkin, M., Santtila, P., & Bull, R. (2012). The linking of burglary crimes using offender behaviour: Testing research cross-nationally and exploring methodology. *Legal and Criminological Psychology*, *17*(2), 276–293. https://doi.org/10.1111/j.2044-8333.2010.02007.x

Tonkin, M., & Woodhams, J. (2017). The feasibility of using crime scene behaviour to detect versatile serial offenders: An empirical test of behavioural consistency, distinctiveness, and discrimination accuracy. *Legal and Criminological Psychology*, *22*(1), 99–115. https://doi.org/10.1111/lcrp.12085

Tonkin, M., Woodhams, J., Bull, R., & Bond, J. W. (2012). Behavioural case linkage with solved and unsolved crimes. *Forensic Science International*, *222*(1–3), 146–153. https://doi.org/10.1016/j.forsciint.2012.05.017

Tonkin, M., Woodhams, J., Bull, R., Bond, J. W., & Palmer, E. J. (2011). Linking different types of crime using geographical and temporal proximity. *Criminal Justice and Behavior*, *38*(11), 1069–1088. https://doi.org/10.1177/0093854811418599

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*(1), 28–56. https://doi.org/10.1002/jip.1372

Winter, J. M., & Rossi, G. (2021). Closer to reality? The application of sequence analysis in crime linkage. *Journal of Criminological Research, Policy and Practice*, *7*(1), 34–50. https://doi.org/10.1108/JCRPP-02-2020-0025

Woodhams, J., Hollin, C. R., & Bull, R. (2007). The psychology of linking crimes: A review of the evidence. *Legal and Criminological Psychology*, *12*(2), 233–249. https://doi.org/10.1348/135532506X118631

Woodhams, J., Tonkin, M., Burrell, A., Imre, H., Winter, J. M., Lam, E. K. M., Jan ten Brinke, G., Webb, M., Labuschagne, G., Bennell, C., Ashmore-Hills, L., van der Kemp, J., Lipponen, S., Pakkanen, T., Rainbow, L., Salfati, C. G., & 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. https://doi.org/10.1111/lcrp.12144

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

How to cite this article: Ewanation, L., Bennell, C., Tonkin, M., & Santtila, P. (2023). Receiver operating characteristic curves in the crime linkage context: Benefits, limitations, and recommendations. *Applied Cognitive Psychology*, 1–13. https://doi.org/10.1002/acp.4122