Domestic Abuse: Testing the RFGV algorithm

University College London

Belur J., Posch K., Hammock D., Davies K., Bradford B., and Myhill A.

Contact details:

Dr Jyoti Belur
Department of Security and Crime Science,
University College London,
35 Tavistock Square,
London, WC1H 9EZ.
Tel: 020 3108 3050

Email: j.belur@ucl.ac.uk

Executive Summary

Context and aims of the study

The use of algorithms to aid police decision-making is becoming more prevalent. One such algorithm to identify high harm perpetrators of Domestic Abuse is the RFGV (Recency, Frequency, Gravity and Victimisation) which was adapted from a similar version used by the Scottish Police to prioritise high risk offenders. Recently, the College of Policing has called for an evaluation of the RFGV, as part of evidence-based practice, prior to potentially standardising it for a national roll out. Consequently, this research, funded by the Home Office has two aims:

1. To develop an understanding of, as well as, examine and test the RFGV algorithm.
2. To assess the existing evidence on the use of algorithms in policing for crime prevention or harm reduction.

Methods

Three police forces. Northumbria Police, West Midlands Police and the Metropolitan Police Service voluntarily partnered with the research team at UCL and the College of Policing to participate in the research. The research adopted a mixed methods approach that included conducting interviews with key stakeholders in the services to understand the decision making involved in the design and construction of the algorithm in conjunction with analysing relevant data to test whether the algorithm is optimum.

Key Findings

From further examination of the algorithm in the context of three forces

- The RFGV is not being used as a risk assessment tool to predict risk of reoffending but is primarily functioning as a prioritisation tool to identify high harm perpetrators (based on their offending history) in order to engage with them to desist from reoffending.
- There is significant inconsistency, even among the three forces, as to how the algorithm is constructed and used.
- Decisions about how much data to incorporate and what time periods to include in the running of the algorithm were made more on the basis of capacity of police systems to process data and of the offender management teams to work with offenders, rather than theoretical considerations.

- There is considerable subjectivity in the construction of the Gravity scores and there is concern that the algorithm does not pay sufficient attention to the impact of high harm but non-crime coercive and controlling behaviour in DA situations.
- Practitioners are very satisfied with the tool as it helps them rank and prioritise DA perpetrators with some degree of perceived objectivity and minimal effort.

Findings from the REA on the use of algorithms for offender management

- The definition of what constitutes an algorithm in the context of offender management in the CJS was varied.
- A majority of studies included in the REA evaluated algorithms related to those predicting risk of recidivism and a few were used for prioritising offenders.
- The lack of access to data, the lack of transparency about what processes the data was subjected to, and the lack of information about how the algorithm outputs were used to manage offenders makes it very difficult to evaluate the accuracy, reliability and validity of the algorithm
- The evidence indicates that although algorithms are more accurate at predicting recidivism than humans, especially if there is a large amount of data to process, the level of predictive accuracy is around 66 to 70% which is just below or at an acceptable level.
- Algorithms were shown to have biased outcomes for some ethnicities and for women and were only as good as the quality of data that was input.

Key Recommendations

Key recommendations for practice

- Since the scoring and ranking produced by the algorithm is subject to further triaging by practitioners before choosing which perpetrators to focus on, it is worth considering whether the triaging process can be refined and made more systematic and transparent so as to produce equitable and fair results.
- The analysis identified the need to consider a larger time-window for data filtering to improve accuracy.
- There is a need for a national conversation around how Gravity should be understood and scored in the context of DA, in order to ensure consistency and fairness.
- The impact and relevance of Recency scores in the analysis were minimal, thus indicating a need to modify the initial data filtering window to ensure greater accuracy.

Key recommendations for future research

- Better understanding of how the algorithm is constructed and how decisions are made in the MPS.
- Lead a national level discussion to devise an acceptable new scale to score DA offences on harm that is not related to potential sentencing (crime harm index) or 'relative harm' of an offence (crime severity score).
- Conduct a detailed evaluation study of how the triaging process works and decisions are made to ensure that fair and equitable results are produced.
- Test whether information gathered on alternative variables by the forces have stronger predictive power in identifying high harm perpetrators of DA who are at risk of reoffending.
- Test whether the algorithm produces fair and equitable outcomes for all offenders.

Testing the RFGV Algorithm

Introduction

The RFGV[1] algorithm is a prominent and widespread iteration of an algorithm designed to rank domestic abuse suspects who are deemed more likely to reoffend thus, could be considered as 'high risk'. Originally a commercial product[2] adapted by Strathclyde Police (now Police Scotland), the RFG algorithm in its principal form looked at the recency and frequency of incidents and crimes as well as the severity, or gravity, of each suspect's reported incidents to produce a final score (named 'offender risk score') between 0 and 100. This score can then be used by offender management teams to target individuals on the upper end of the spectrum. It is another tool in the array of offender risk assessment and management arrangements currently in use to identify high risk perpetrators, such as the DASH (Domestic Abuse Stalking and Honour Based Violence)[3] risk assessment tool, as well as the MARAC (Multi-Agency Risk Assessment Conference)[4] and MAPPA (Multi Agency Public Protection Arrangements)[5] processes for risk management that are currently in place in almost all police services in England and Wales.

Northumbria Police evolved RFG into RFGV by incorporating an additional component that considers the number of different domestic abuse victims an offender has interacted with. The intuition behind this was to help identify domestic abuse offenders with multiple victims which can (1) identify actors who tend to harm more people and (2) can be indicative of seriality. This algorithm has since been adopted by many forces, including West Midlands Police (WMP). The RFGV also inspired other forces to devise modified algorithms, such as the Metropolitan Police Service (MPS), which came up with an alternative, RFHA. More recently, the College of Policing (2021) has called for an evaluation of the usage of the RFGV and a potential standardisation of the algorithm for use across English and Welsh forces to ensure that

[1] RFGV here stands for Recency Frequency Gravity and Victimisation.

[2] The variables included Recency, Frequency and Monetary/Consumer Behaviour to predict consumer purchasing habits.

[3] The DASH tool (2009) is a checklist for identifying, assessing and managing risk implemented across all police services in the UK since 2009. (https://www.dashriskchecklist.co.uk/)

[4] Multi Agency Risk Assessment Conference is a regular risk assessment meeting in which stakeholders from relevant agencies share information on high-risk domestic abuse cases and put risk management plans in place. (Safe Lives 2014, https://safelives.org.uk/sites/default/files/resources/MARAC%20FAQs%20General%20FINAL.pdf)

[5] Multi Agency Public Protection Arrangements to assess and manage risks posed by sexual and violence offenders (https://www.gov.uk/government/publications/multi-agency-public-protection-arrangements-mappa-guidance)

the "approach is having the intended effect and the right people are being identified"6. Consequently, this exploratory study was funded by the Home Office aimed at better understanding how and why the algorithm evolved into its current version, how are its outputs being used in practice in various forces, and whether the algorithm is optimal or can be further refined.

Aim of the study

This study aims to contribute to evidence based practice through two key objectives:

1. To develop an understanding of, as well as, examine and test the RFGV algorithm.
2. To assess the existing evidence on the use of algorithms in policing for crime prevention or harm reduction.

Three pieces of work were undertaken by a team of researchers at the University College London and the College of Policing, in partnership with Northumbria Police, WMP and the MPS.

1. Acquiring a qualitative understanding of the construct of the algorithm based on interviews with key informants responsible for running the algorithm and working with the outputs generated thereby.
2. Testing the algorithm in and across forces to assess the contribution of each of the RFGV elements to identifying the most prolific perpetrators of DA.
3. Conducting a Rapid Evidence Assessment on the use of algorithms in offender risk assessment for crime prevention or harm reduction.

The report consists of two parts in pursuance of the two objectives of the research. The first part of the report is divided into three main sections. First, we provide an overview of the RFGV algorithm, its components, and review how this nominally same algorithm was adopted by the Northumbria and West-Midlands police forces, and to what extent minor differences had an impact on the people who were identified by the algorithm. Second, we will turn to our assessment of the RFGV algorithm, including its measurement model and whether it can be used as a tool to track the potential escalation of domestic assault incidents. Finally, we describe an alternative algorithm used by the Metropolitan Police Force: the RFHA, and discuss some of the potential advantages and disadvantages of this alternative proposal

In the second part of the report, we present the findings of the Rapid Evidence Assessment of the existing evidence on the use of algorithms in offender management that are currently being used.

[6] https://committees.parliament.uk/writtenevidence/36621/html/

Methods

Interviews with key stakeholders in two forces:

A total of 5 interviews were conducted with 8 people in the two forces with the aim of understanding the underpinnings of the algorithm from the perspective of those responsible for running it in force, and the perspective of practitioners working with DA perpetrators identified by the algorithm. Interviewees included a mix of police staff and officers involved in the day to day running of the algorithm and those involved in working with perpetrators identified by the algorithm in both forces. Interviews were carried out via Microsoft Teams during the month of August 2021. Every effort has been made to anonymise the participants and maintain confidentiality as appropriate. Ethical permission for the study was granted by UCL's Department of Security and Crime Science Ethics Committee.

Data analysis:

Obtaining the data from both the Northumbria and West Midlands police (WMP) forces required the completion of a project proposal document as well as separate data sharing agreements that complied with each forces data protection policies and procedures. Once these had been agreed on by both parties the already anonymised data was exchanged using appropriate levels of encryption.

The West Midlands Police Force supplied us with three datasets (victims, perpetrators, and the relationship between them) that were split across six files. The constituent files were combined, with the different datasets joined using the crime number, to produce a single dataset of 459,661 crime and non-crime incidents which is the input for the WMP variant of the RFGV algorithm. In contrast, the Northumbria Police Force supplied a single file containing two datasets pertaining to domestic abuse crime and non-crime incidents. These were imported as is, albeit with missing character values replaced with "NA", to provide two distinct datasets of 359,168 and 166,185 domestic abuse incidents and crimes respectively which is the input for the Northumbria variant of the algorithm. No additional pre-processing of the data was undertaken in order to mimic the operational pipeline used by practitioners. However, the algorithm has been designed to deal with any missing data by imputing 0 scores.

In order to utilise the WMP data with the Northumbria RFGV variant, the dataset was split using the "NON CRIME" descriptor used in the offence grouping to produce a dataset of incidents (inclusive of crimes) and a dataset of crimes. By comparison, merging the two Northumbria datasets into a single dataset for use in the WMP algorithm was not possible based on the fields supplied. A minor adaptation to the input aspect of the algorithm was therefore written to select the appropriate dataset for each of the algorithm components in order to achieve a working algorithm.

Findings:

DA related set up in the forces

The composition and make-up of the teams working with DA perpetrators and the RFGV algorithm were very different in the two forces. In Northumbria the RFGV feeds into the MATAC unit (Multi Agency Tasking and Co-ordination) which is headed by senior police staff and aided by an analyst. The whole unit comprises of seven police staff, including three analysts and two Domestic Abuse and Criminal Justice Workers who work with the identified perpetrators. No police officers work in the unit. The DA workers support all high-risk offenders identified by the algorithm who are currently not in any other kind of statutory management or in prison.

The DA management process is slightly different in the West Midlands. WMP is made up of several local policing units (LPUs), and there is a DA team embedded in each LPU that comprises of police officers who are offender managers and supervisors. Most of these teams consist of around six or seven officers, all of whom are offender managers. The algorithm is run centrally, but such that perpetrators are scored and ranked for each LPU and each DA team manager decides what percentage of the top ranking individuals would be managed by their team. At any given point offender managers manage a load of about 12 to 15 perpetrators on average,. The DA teams are closely linked to and feed into the MARAC process. Perpetrators of concern, who are not in the top 5 or 10 % of the list produced by the algorithm, are passed on to the local neighbourhood teams to supervise or support.

Aim of the algorithm:

Interviews were conducted with stakeholders who are involved in the day-to-day construction and running of the algorithm in both forces. The aim of the algorithm in both forces seems to be to process offender data in order to identify or prioritise offenders according to the highest score based on their criminal record and conduct, which comprises of four variables R (recency), F (frequency), G (gravity) and V (victimisation [number of]). It was clear that the purpose of the algorithm, as it is being used at present in both the forces, is to identify serial perpetrators – defined as those who offend against two or more victims.

> “The very, very high-risk perpetrators are probably already on the radar because they will either potentially be in the MAPPA process or the MARAC process, but the ones that seem to be kind of, be under the radar are those ones that we're moving from victim to victim to victim, where the victim potentially wasn't really being identified as high risk. So that these serial perpetrators were causing a lot of harm to a lot of people at the same time, so that came out really as the problem.” (Staff, Northumbria)

The definition of a serial perpetrator was further clarified by one interviewee:

> “They need to have offended against at least two persons and it has to be a partner or very close family member. So if you've offended against your mum and your brother, you would be a serial DA offender and there is no relationship there, you know, as in a sexual relationship, but there is personal relationship. If you've offended against two girlfriends, then you would be right. If you have offended against a girlfriend and a brother then that also would be. But if you've offended against a random member of in the community, like a shopkeeper, and then you've offended against girlfriend, you're not serial.” (Police officer, West Midlands)

Interviewees in Northumbria were careful to say that the algorithm was focused on identifying the most harmful and serial perpetrators, while defining harm very broadly. They preferred to use the term ‘harm’ instead of ‘risk’, which was considered to be a problematic term and can often be confused with the kind of risk assessments done by the DASH (domestic abuse, stalking, and harassment) tool; as one interviewee said,

> "We try and steer clear of using the risk terminology" (Staff, Northumbria).

The algorithm, first adopted by Northumbria, is a modification of the original algorithm based on RFG (recency, frequency, gravity) scoring that was being used by Police Scotland to identify high risk offenders for different crime types.

The team at Northumbria added number of victims to the original criteria of recency, frequency and gravity, and used the algorithm to identify the highest scoring DA perpetrators. The algorithm is used by the Multi-agency Tasking and Co-ordination unit set up in November 2015 and focuses exclusively on working with DA perpetrators. According the MATAC manager, the motivation for using the algorithm in the first place emerged on the back of an HMIC inspection in 2014 which concluded that no force really had a good DA perpetrator management process. One of the recommendations was that police forces look to the RFG algorithm used by Police Scotland, who had been using the algorithm to work with other types of offenders. A problem profile on DA done by the Northumbria police identified that there was a problem with serial DA perpetrators who moved from one victim to another causing harm to a large number of victims.

In WMP the algorithm has been in use for about three years. One of the interviewees said that they had previously done some work using algorithms to prioritise cohorts of outstanding suspects to help officers in their decision-making. They had constructed their own algorithm around the concept of harm to assist officers with their decision-making. Consequently, many within the force had some experience of working with algorithms and so there was little resistance internally when they linked in with Northumbria and the College of Policing to adopt the RFGV in principle and practice. WMP interviewees said that they adopted the Northumbria's algorithm faithfully, and although they made minor tweaks in terms of what data sources would feed into it, they did not fundamentally change the algorithm.

The impetus for using the algorithm came from some national conversations that their Public Protection Units were involved with and there was a lot of interest from the College of Policing who were pushing the DA agenda. The decision to adopt and adapt the algorithm was taken at the level of the Performance Manager for the force who was involved in developing the work, there was a lot of strategic governance and oversight, but the interviewees said that they was not specifically tasked strategically to use it. Instead, they were proactive in adopting it. Interviewees explained that there was quick acceptance of the algorithm because not only were practitioners in the force used to working with algorithms, but that they found this algorithm to be helpful, as compared to the system in place before, which was just a risk assessment form based on subjective reviews by the offender managers. It was felt that the algorithm provided police officers with a more objective aid to decision-making.

Components of the RFGV algorithm

In this section, we discuss each of the components of the RFGV algorithm: recency (R), frequency (F), gravity (G), and serial victimisation (V). Our overview entails what the components are capturing from a substantive perspective, how the practitioners understood the components based on the interviews, and how the components are calculated for the algorithm.

In both forces, the RFGV algorithm is run on a database of reported DA offences that consists of two elements - recorded crime and reported DA related incidents. Both forces were keenly aware of the importance of reported incidents – as opposed to simply recorded crime – in understanding the degree of harm and risk for the victim and therefore included both recorded crimes and reported incidents that were DA related only in their algorithms. Consequently, interviewees recognised that the identified perpetrators may not have been convicted or even prosecuted, and interviewees therefore stated that ‘perpetrator’ was a more appropriate term to use rather than ‘offender’. However, our analysis suggests an individual is unlikely to score highly unless they were involved in an actual crime (see below). Hence, individuals with high scores, for purpose of this report, can rightly be called ‘offenders’. We will use therefore use the terms ‘perpetrators’, ‘suspects’ and ‘offenders’ interchangeably as, for the purposes of the algorithms discussed here, this distinction is not of major relevance.

Recency (R)

The recency component looks at all of the reported domestic abuse incidents (including crimes) for every offender with a domestic abuse flag on their record. Using the date when each incident first occurred, and the date of analysis (the date on which the algorithm is run), the number of days which have passed since each event occurred are calculated. This information is then transformed into a score for each offender, marginally varying based on each force's implementation of the algorithm, with both WMP and Northumbria Police taking the average (mean) of the set of individual event scores to produce a final score. This component is designed to assess whether the offender is still actively committing domestic abuse and bring the more recent offenders to an offender management team's attention.

To be clear, when calculating R both forces use the average recency score instead of the most recent incident. The logic for this, according to interviewees from Northumbria, was that the two-three year time window they tend to consider is so large that averaging the recency scores will eliminate bias towards only the most recent offences.

> “We don’t want to just include people who committed an offence last night” (Staff, Northumbria).

Two more justifications were provided for averaging the recency scores. One was that if a perpetrator has offended multiple times over the time period being considered, there would be multiple scores; and secondly that, focusing on only the most recent score would not capture the fact that the perpetrator might have offended multiple times against multiple victims over the period under consideration (which in this case was two years). However, taking the average does little to address this point. Indeed, as we will demonstrate in our assessment of the algorithm, taking the mean will actually decrease the score of those who have committed multiple incidents, relative to others, because by definition it collapses multiple events into one number, potentially lowering the overall score of multiple offenders. Interviewees in West Midlands said that they adopted the principle of using average recency that Northumbria employed simply because they didn’t have a specific reason for making any changes, as it seemed to be working well enough. However, there are slight differences between how recency scores are being calculated by each of the forces, which will be discussed in the ‘differences’ section.

Frequency (F)

The frequency component assesses the number of incidents that each domestic abuse offender has been involved in within the period of interest. The total number of incidents is then converted into a score. The frequency component is therefore designed to identify offenders who are involved in a significant number of domestic abuse incidents. Both Northumbria and West Midlands Police calculate frequency using the same scoring system, as shown in Table [1]. However, it should be noted that pre-analysis filtering (discussed later in section [x]) will have an effect on the number of incidents included in the counting process.

	x = Number of Incidents Involving Offender X	
Score	**Northumbria Police**	**West Midlands Police**
200	$32 \leq x$	$32 \leq x$
150	$21 \leq x \leq 31$	$21 \leq x \leq 31$
100	$17 \leq x \leq 20$	$17 \leq x \leq 20$
75	$13 \leq x \leq 16$	$13 \leq x \leq 16$
50	$9 \leq x \leq 12$	$9 \leq x \leq 12$
30	$5 \leq x \leq 8$	$5 \leq x \leq 8$
15	$x \leq 4$	$x \leq 4$

Table 1: Table demonstrating how the number of incidents involving an offender are converted into a score for the frequency component.

F (frequency) scores are calculated in both forces, according to interviewees, using both recorded crimes and reported DA related incidents. This was said to be in conscious recognition of the fact that low level DA incidents can have negative consequences for the victim. Thus, all crimes, as well as reported incidents against all victims of a perpetrator are added to the calculation of the F score.

Gravity (G)

The gravity component assesses the severity of crimes committed by each offender. The calculation of the gravity scores developed organically in Northumbria. According to interviewees from Northumbria, in the early days they had used the Home Office Classification Code to assign gravity scores to offences7, but had quickly found them to be unsuitable when applied to DA cases. Reasons included the fact that the Index consists of over 7000 offences which were considered to be too many and prone to being changed from time to time.

Furthermore, the gravity scores often did not reflect the true gravity of DA offences. For example, the scores for murder and offences involving physical violence or injury are higher as compared to coercive control. However, it is well known that although verbal and psychological abuse, breach of restraining orders, and harassment can cause great harm and damage to the victim in DA cases, they have low sentences and are therefore low scoring. Thus, interviewees said that based on their experience they selected a list of 100 most common or recurring offence types based on Home Office Severity Scores, gave them scores based on perceived severity as related to DA cases. This was done by staff involved in the MATAC unit intuitively, based on their experience and understanding, but interviewees felt that these have subsequently been found to be quite satisfactory.

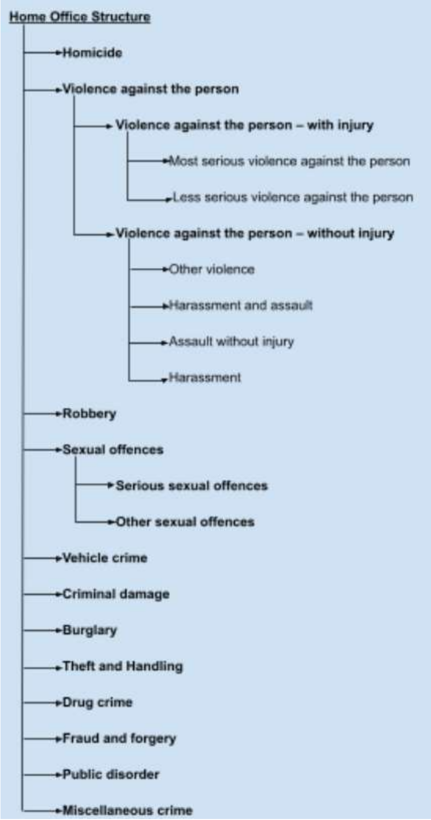


Figure X Home Office Severity Structure

[7] The Home Office Offence Classification Code assigned to each crime is used in conjunction with a lookup table to provide a gravity score for each recorded crime.

Interviewees from Northumbria admitted that, due to the changes initiated by them, the gravity aspect of the algorithm was the most subjective and therefore was probably the area most open to challenge. Despite their tweaks, the Home Office severity structure was largely retained, as depicted by Figure X.

For any given incident, an offender is given a gravity score equal to the maximum score assigned to each of their individual offences during that incident. To put it differently, they will only get the highest ('maximum') score of the most severe incident they perpetrated. This gravity score varies between 100 for the most severe offences to 0 for perpetrators who have not committed any crimes or been involved in any non-crime incidents deemed of importance by the respective police force. The inclusion of the gravity component is designed to raise the profile of domestic abuse offenders who are generating the most harm to their victims. As discussed in the upcoming section on differences, Northumbria and WMP use slightly different numbers due to later revisions and updates carried out by Northumbria.

Victim (V)

Since the main aim of the algorithm, according to the interviewees, was to identify serial perpetrators, the V (victimisation) component is important for capturing DA abuse and is the addition made by Northumbria (and subsequently adopted by WMP) to the original RFG algorithm of the Scottish police.

The victim component counts the number of distinct victims of the perpetrator within the period of interest (i.e. the number of victims affected by their actions). Interviewees clarified that if an offender had offended multiple times against the same victim, the V score would be calculated for a single victim however their repeat offending (against the same victim) would be captured by the F (frequency) counts

The victim component was incorporated into the algorithm in an attempt to aid with identifying the serial offenders who are committing harm toward a larger number of individuals.

The victim component scoring system used by the Northumbria and WMP are identical. The more victims affected by a given suspect the higher the score (as shown in Table 2), with a maximum score of 200 for incidents involving more than 8 different victims and a minimum score of 15 for incidents involving only a single victim within the period of analysis.

	x = Number of Victims Associated with Suspect Y	
Score	**Northumbria Police**	**West Midlands Police**
200	8 ≤ x	8 ≤ x
150	6 ≤ x ≤ 7	6 ≤ x ≤ 7
100	x = 5	x = 5

75	x = 4	x = 4
50	x = 3	x = 3
30	x = 2	x = 2
15	x = 1	x = 1

Table 2: Table conveying the conversion of the number of distinct victims involved with a given suspect into a suspect's Victim score.

Final Score & Component Weightings

To derive the final score for each perpetrator, the scores for the four components are calculated and entered into a single spreadsheet. These four components are then added up and then divided by six to produce a final score which will take values between 0 and 100. This can be expressed by the following algorithm:

$$Final\ Score\ = \frac{1}{6}(Recency\ +\ Frequency\ +\ Gravity\ +\ Victim)$$

It should be noted that the components are not weighted equally. The frequency and victim components are provided a weighting double that of the recency and gravity components: the former being measured from 0 to 200 and the latter measured from 0 to 100. This extra weight for frequency and victimisation reflects their increased importance in the eyes of Northumbria and WMP. Simply put, the algorithm considers frequency and the number of victims twice as important compared to recency or gravity. This, however, also means that by simply adding up the four components, one could score between 0 and 600. Hence, to ‘normalise’ the result of the final score to a number between 0 and 100, the sum of the four components needs to be divided by six.

When interviewees were asked why the component was weighted differently – i.e. frequency and victimisation were given double weighting as compared to recency and gravity, they responded by saying that this was to identify those serial perpetrators who were very harmful because they were offending against several victims. They did not believe that this way of weighting the variables was biased because:

> “I think it's pretty rare where we would have somebody with a high frequency and a huge number of victims as well” (Staff, Northumbria).

In the interviewees' experience, often serial offenders perpetrate one or two lower-level offences but against a number of victims, so they believed that the current method of weighting captured the nuances of serial offending better.

Interviewees from West Midlands said that they retained the weightings given by Northumbria. Interviewees mentioned that some recent research [cite] into the weightings used by West Midlands RFGV algorithm concluded that they are using the best combination of rating and thus, said that

> "We feel reasonably confident that the core construct of the algorithm is quite solid." (Police officer, West Midlands).

After deriving the final scores for each individual in the dataset, all perpetrators are ranked. The expectation is that those ranking highest are the serial domestic abuse offenders who are committing the most harm and need to be dealt with by the relevant force's offender management teams. As will be shown in the following section, both forces use the output from this algorithm as a prioritisation tool, where officers and other people involved in offender management can select from the high-ranking individuals the ones they want to enrol for future intervention(s).

Differences Between Implementation in WMP and Northumbria Implementations

While the components of the RFGV algorithms are the same for the Northumbria and West Midlands police forces, their implementation is slightly different mainly due to the varying data infrastructure used by each of the forces the preliminary data filtering is distinct.

Preliminary Data Filtering

Prior to running the RFGV algorithm, the relevant crime and incident data must be extracted from the respective forces database systems. The methodology used by the Northumbria and West Midlands forces differs significantly. We have separated the data selection part of each force's implementation of the RFGV algorithm and the algorithm itself by calling it the filtering stage.

Whilst the original RFG algorithm used by the Scottish police only used 6 months of data, it was recognised early on that DA is a different kind of crime, especially since relationships can last over a much longer period of time. Consequently, analysts in Northumbria decided to increase the time period under consideration to two years. Thus, the Northumbria filter takes two years' worth of domestic abuse incidents and any associated crime records dating back from the date of analysis. By contrast, the West Midlands filter identifies all domestic abuse offenders *active* within the last 3 months. This list of "active offenders" is then used to extract 3 years' worth of crime and non-crime incidents from the date of analysis for these offenders only. It is clear from the above difference that (a) the Northumbria method identifies more potential offenders in a two-year period, making it easier to track them over time, although many single, inactive suspects are also included while (b) the West Midlands approach identifies fewer offenders for each run but for them it considers a longer period (3 years). We will revisit the question of preliminary data filtering and the advantages/disadvantages of each approach in our recommendations section.

The reason for the particular time period chosen by Northumbria, according to interviewees was partly theoretical (in that DA offences can often last for longer than six months), and largely practical, based on the capacity of their systems to cope with the running of the algorithm. The justification for the choice of filtering method in West Midlands was similar, in that, the time period was determined by their system's capacity to cope. Given that the number of DA offences in West Midlands is much larger, their systems were able to cope with three months' worth of DA data to identify active offenders as it was considered to be more reasonable as the list would not be constantly changing and would give them some time to put in some interventions in place. Further, we were told that the justification for using three years of perpetrators' criminal history to feed into the algorithm was more in line with the national definition of repeat victimisation according to which the cutoff for being considered a repeat victim is three years between victimisation incidents.

Recency (R)

The way in which the scores for recency are calculated from the number of days varies slightly between the two forces, with the West Midlands Police force incorporating a zero score category, assigning the value '0' to incidents which transpired 731 or more days ago. Further separating the way the algorithms worked was the score 5 band used by Northumbria. These differences are highlighted in Table [3], and mean, that the final score from the two algorithms will not be identical.

	Days Since Event (x)	
Score	**Northumbria Police**	**West Midlands Police**
100	$x \leq 56$	$x \leq 56$
75	$57 \leq x \leq 120$	$57 \leq x \leq 120$
50	$121 \leq x \leq 240$	$121 \leq x \leq 240$
30	$241 \leq x \leq 360$	$241 \leq x \leq 360$
20	$361 \leq x \leq 480$	$361 \leq x \leq 480$
10	$481 \leq x \leq 600$	$481 \leq x \leq 600$
5	$601 \leq x$	$601 \leq x \leq 730$

0	N/A	731 ≤ x

Table 3: Table showing the conversions from days since an incident into a frequency score for the Northumbria and West Midlands Police Forces.

Gravity (G)

The gravity scores used by West Midlands Police resemble an earlier version of those used by Northumbria. However, Northumbria Police constantly review the score assigned to each Home Office Offence Classification Code in line with an ever-changing criminal landscape. As these differences are very specific and, in the case of Northumbria, ever-changing, we will not go into further detail about them. For the purposes of calculating the final scores by the algorithms, these minor differences will also result in change scores. For reference, the gravity lookup tables used by each of the two forces are available in Appendix [x].

Data Analysis:

Following the interviews with stakeholders which permitted a clearer understanding of the implementation of the RFGV in the two forces, we proceeded with the analytical portion of this research. We performed two different types of analyses; the first, focusing on the raw data to test whether the various algorithm components and the composite appeared suitable; and second consisted of utilising the scores generated from the algorithm to assess how informative each of the components were to the overall score.

Replicating the RFGV algorithm on the datasets of the other forces

Due to the considerable similarity between the Northumbria and West-Midland algorithms, running one's algorithm on the other's data could be implemented. To reiterate, the main differences between the two algorithms were (1) the preliminary data filtering (i.e. the time-window considered by each of the forces for the calculation of the scores), (2) a slight difference in how recency was estimated, and (3) again, some minor differences in the estimation of the gravity score. Apart from these discrepancies, the two algorithms were very similar.

For all subsequent analysis discussed in this section, we only considered data from November 2016 and November 2020 from each of the forces. This was done as this was deemed as a sufficiently long reporting period and also due to the relatively limited change to reporting practices (i.e. the year on year growth in reporting DA incidents was considerably larger in the previous years, which could have reasonably affected the results). To produce the scores, each algorithm was run once for each month, largely in line with the current practices of the two police forces.

A general overview of the scores derived by the two algorithms

First, it is worth considering the shape of the scores derived by the respective algorithms. To exemplify this, we chose scores for a single month, the March of 2019, and ran both algorithms on the West Midlands dataset. Figure 1 shows the final scores Northumbria police would have received had they run their algorithm, and Figure 2 represents the final scores WMP received.

Due to the differences in data filtering and the minor distinctions in how the scores are calculated, the scores are not directly comparable. Yet, the shape of the histograms depicted by Figures 1 and 2 are instructive in at least three ways.

- First, both histograms have the vast majority of their cases on the left side. This is usually called a *leftward skew*, which implies that most DA cases do not score high (i.e. they are likely to be less severe, singular, incidents, against one person).
- Second, and by contrast, high scoring incidents appear to be quite rare, with the distribution having a *long right tail*. This means, that only a fairly limited number of DA cases are considered to be of high severity, with multiple instances and against multiple people.
- Finally, it is worth juxtaposing the two histograms. Mainly due to the preliminary data filtering, Figure 1 for the Northumbria police have a much thicker left tail, with many cases scoring close to zero. Conversely, Figure 2 for West Midlands police have fewer cases close to the minimum. This represents an important difference in approach. While Northumbria police consider all perpetrators in the last two years, West Midlands only focus on the last three months (i.e. 'active suspects') and include their history for the past three years.

We will consider each of these strategies and their potential implications in the recommendations section.

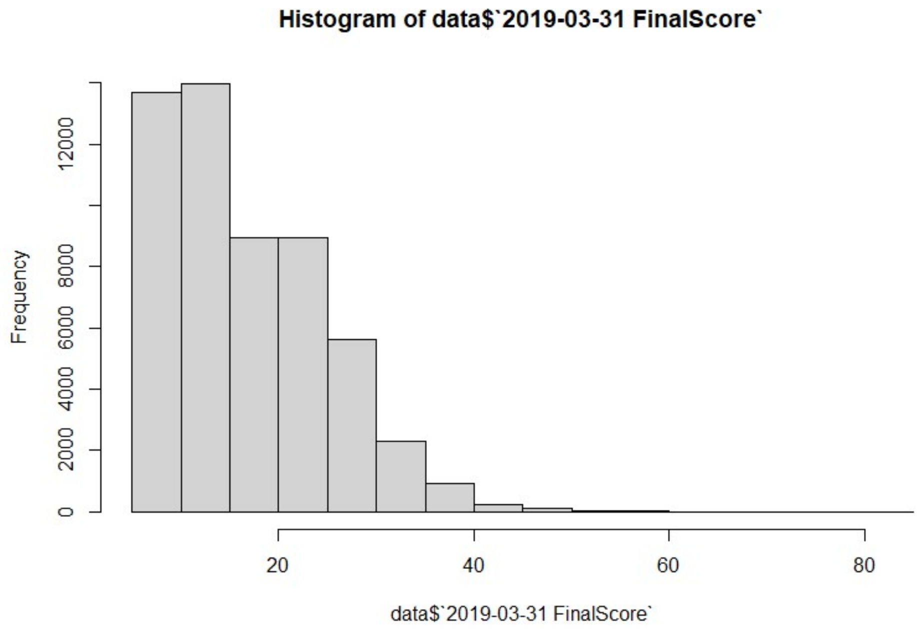


Figure 1 Final scores derived using the Northumbria algorithm on the data provided by WMP (March 2019)

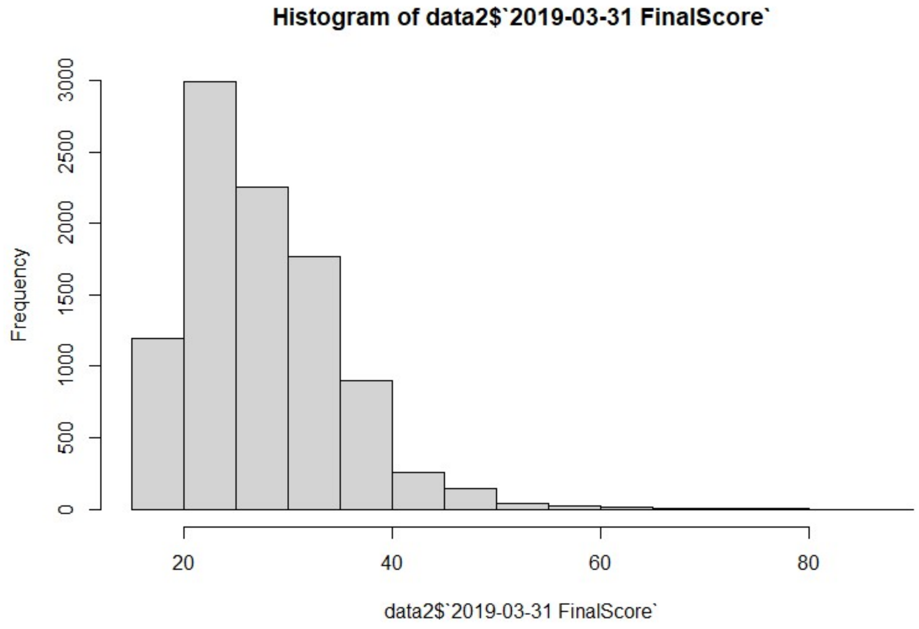


Figure 2 Final scores derived using the WMP algorithm on the data provided by WMP (March 2019)

Similarity between the results of the algorithms

The first question we sought to answer was how similar the results of these algorithms were. In other words, how similar the results would have been, had the algorithm used by one force been adopted by the other force? To put it another way, how much difference using one RFGV algorithm instead of another makes when it comes to identifying DA offenders?

Due to the slight differences between the two algorithms noted above, the scores derived by the algorithms are not directly comparable. Thus, to address this issue of comparability, it is worth taking an alternative approach, namely, how a certain individual would have been ranked by each of the algorithms. In this case, the highest scoring individual(s) would (all) receive rank number 1, the second highest scoring individual(s) rank number 2, and so on (i.e. all individuals tied with the same score would receive the same rank). This way, instead of inquiring about the score each individual would have received by the two algorithms, the question becomes whether the individual would have received *the same rank* had the other algorithm been used by a particular police force.

Recall that the distribution of DA incidents and crimes have a decidedly leftward tilt, with the vast majority of cases being singular non-severe events against a single victim. In recognition of this, our analysis focused on the right tail of the distribution (i.e. the most serious offences). In particular, we assessed the overlap between the two algorithms by quantifying the proportion of the individuals who would have been selected by both algorithms. Each algorithm was run on both police forces' datasets, resulting in four runs altogether: (1) Northumbria algorithm on Northumbria dataset, (2) WMP algorithm on Northumbria dataset, (3) WMP algorithm on WMP dataset, and (4) Northumbria algorithm on WMP dataset..

Percentile	Overlap for the West-Midlands dataset (%)	Overlap for the Northumbria dataset (%)
Top 10%	99.1% [98.1%-99.8%]	98.7% [97.2-99.3%]
Top 5%	97.8% [95.3-99.1%]	97.1% [94.1-98.8%]
Top 1%	94.8% [90.1-97.3%]	93.2% [87.1-95.6%]

Table 4: Degree of concordance between the rankings of the two algorithms with the range of possible values in square brackets

Table 4 summarises the results of this analysis. These results indicate that there is a large level of agreement between the two algorithms, with up to an average of 99.1% agreement when it comes to the top 10% of ranked offenders. However, this table also implies that closer to the top, the overlap becomes slightly smaller, with only an average of 94.8% overlap for the top 1% in the case of the West-Midlands data and 93.2% overlap in the case of the Northumbria data. This indicates, that closer to the top, the two algorithms are more likely to disagree with each other.

The type of data collected by each force also appears to play a role, with a slightly lower level of overlap in the case of the Northumbria dataset compared to the WMP dataset. This could occur for several reasons: differences in reporting practices by each force, the varying composition of DA offences and offenders, and so on.

What do the above results indicate? For the top 10% of the perpetrators, it seems that these algorithms were very closely aligned, with only very minor differences in the people selected by each. However, closer to the top, the differences become larger. On average, one in twenty individuals flagged by one algorithm for the top 1% of the perpetrators is not flagged by the other, and vice versa. It is hard to give a precise number on how many people are likely to be affected by this, as the top 1% of perpetrators fluctuates based on the dataset of each force, by each month, and year. Yet, generally speaking, it entails around 600-900 people who are ranked 1-300 by the algorithm – in other words, somewhere between 30 and 45 will (not) be listed by one of the algorithms. As indicated by the interviews, the top few hundred people are most likely to come under scrutiny by the police forces, although there is a certain level of persistence in the people who will take on the highest scores.

Assessment of potential improvements to the RFGV algorithm

The generous support of the Northumbria and West-Midlands police forces meant that we had access to multiple years' worth of data from both forces. With these datasets we were able to carry out analysis to (1) evaluate how the various components of the RFGV algorithm are related to each other and (2) to explore potential ways to improve the algorithm.

With regards to improving the algorithm, we are, however, constrained by the lack of information on what happens to high scoring individuals. We do not know which individuals in the dataset were arrested, were being managed, and so on. In other words, we cannot carry out a proper evaluation of the performance of the algorithm and can only speculate about why certain patterns emerged in the results. Nevertheless, we believe that our results are still indicative of how the RFGV algorithm could be potentially improved.

Relationship between the various components of the RFGV algorithm

As the first step, we were interested in how the various components of the RFGV algorithm are related to each other. Therefore, we estimated the association between the various components of the algorithm: recency, frequency, gravity, and serial victimisation. All analysis was done using the unweighted scores for both iterations of the algorithm on monthly estimates (using weighted estimates produces almost identical results).

Correlation and Measurement Models

First, we considered the association between the four components using correlation analysis. As before, we considered data from four runs of the algorithms, with Northumbria's algorithm run on the Northumbria and WMP dataset, and the WMP's algorithm run on the WMP and Northumbria dataset. In the case of correlation analysis, estimated values span between -1 and 1. Negative values indicate a negative linear relationship, positive values a positive one, whilst values closer to -1 or 1 are stronger than the ones close to 0. In fact, values around 0 imply no association.

From the four components, the association between frequency and serial victims was the strongest, taking on statistically significant values of 0.38-0.44 (WMP dataset) and 0.36-0.42 (Northumbria dataset). The relationship between frequency and gravity (WMP: 0.23-0.27; Northumbria: 0.21-0.25) and serial victims and gravity (WMP: 0.20-0.24; Northumbria: 0.17-0.21) was weaker, but still statistically significant. However, recency was unrelated to the other three variables in either of the datasets, with non-significant values of -0.01 and 0.01. In other words, the recency of an event appeared to be independent of the other components of the algorithm.

We carried out more complex measurement models using principal component and factor analysis. These techniques reinforced the observations outlined above: while the frequency, serial victimisation, and gravity components produced acceptable measurement models, recency appeared to be unrelated to the other components in the algorithm. In other words, and purely from an empirical point of view, frequency, serial victimisation, and gravity appear to measure something similar, whilst recency seems to be extraneous to them.

Pragmatic and Theoretical Concerns Regarding the Recency Component

Provided that the recency component appears to be unrelated to the other three elements of the RFGV, it is worth discussing how to interpret these findings. On its own, the uniqueness of one of the components does not necessarily give reason for concern, as long as its inclusion makes sense from an empirical and theoretical point of view. However, we believe that there are at least four reasons which make dropping the recency component compelling.

First, and as discussed earlier, each of the forces use preliminary data filtering to only consider more recent cases. This means, that for West Midlands, people who reach less than 75 points for recency (i.e. were involved in an incident more than three months ago) would not even appear. In other words, by modifying the preliminary data filtering to only include more recent cases, the recency score becomes rather redundant in the case of WMP.

Second, the recency score is prone to create a 'calendar lottery', where ending up on one or the other end of a certain cut-off point can boost/reduce someone's score. For instance, if the algorithm is run on the fourth day of each month (and not considering February), someone committing a crime between the 5th and 8th will see their points drop within two months, as their incident happened 57 or more days ago. Conversely, committing a crime on the 9th would have seen their score unchanged. This 'calendar lottery' can give certain people more/less favourable scores just by chance.

Third, it is unclear why the algorithm would be designed to prioritise more recent events. As indicated earlier, the preliminary data filtering on its own makes certain events 'obsolete' by not including them in the analysis any longer. Should a pre-defined sensible time-window be identified, including recency as a component hardly has any added value.

Fourth, taking the average recency also appears to be counter-intuitive and counter-productive. By way of example, a person in Northumbria who committed a crime in the last 56 days would receive 100 points. By contrast, another person who also committed a crime in the last 56 days, would only receive a value of half that (50), provided that they also were involved in an incident 601 or more days ago. Although this second person would have a higher score in frequency, this is unlikely to be able to compensate for the 50 point difference between a first time and a multiple offender. In effect, the current algorithm would bring more attention to a first-time offender compared to a serial offender.

Escalation: WMP Case Study

In this section, we focus specifically on the West Midlands Police dataset, empirically analysing the occurrence of DA events to gain a better understanding on whether the data lends itself to the algorithm currently employed by the various police forces or other future algorithms which might be capable of preventing future DA-related incidents and harm. For this analysis, we use all available data from the West-Midlands police.

Descriptive Overview

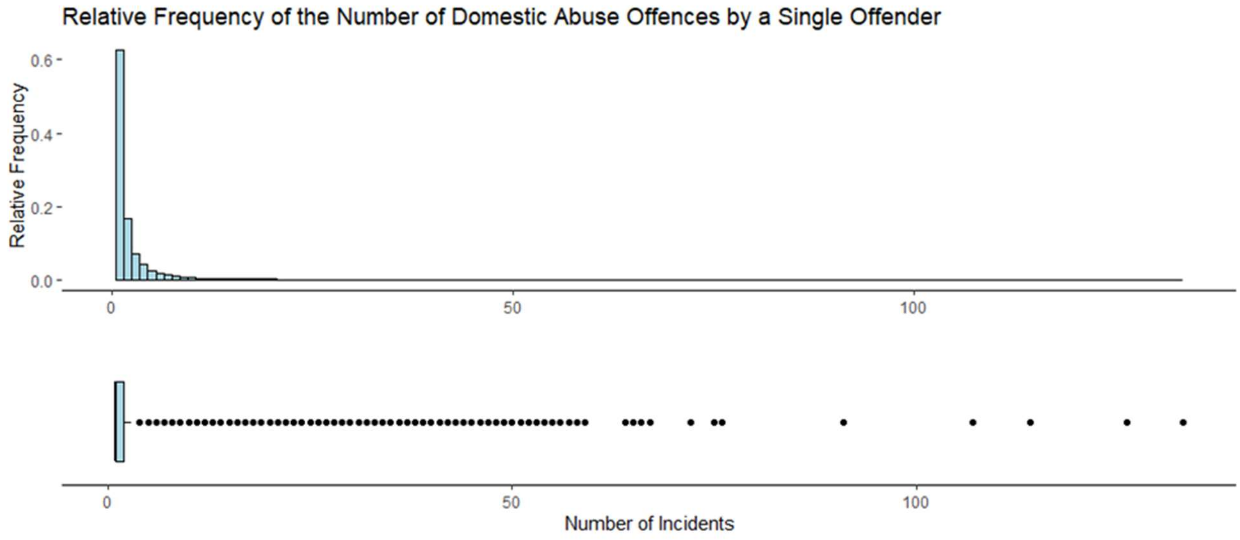


Figure [3]:

Min	Lower Quartile	Median	Mean	Upper Quartile	Max
1.000	1.000	1.000	2.204	2.000	133.000

Table[5]:

Figure [3] shows us that the number of incidents committed by a single offender is heavily skewed. We can see that over 60% (130,733) of domestic abuse perpetrators are only involved in a single incident across the span of the dataset with serial offenders (defined as someone who commits more than two domestic abuse offences) only representing 20.81% of all offenders. It should however be noted that these figures are artificially boosted, as single domestic abuse incidents involving multiple victims are counted as separate incidents in the dataset. This means that the number of serial offenders will be even smaller. In other words, most offenders only appear in the dataset once, and are very unlikely to reoffend, which is also indicated by the median, which is 1. The mean is 2.2, also implying that most offenders either only involved in one or two DA incidents.

Removing the uppermost outliers by selecting only the data that sits within the 97.5th percentile (replotted in figure [X]) shows the extent of the skew. We can see that with the outliers removed the greatest number of incidents committed by an offender is 9. This suggests that it is unlikely, in the vast majority of cases, that offenders would go on to commit more than 9 domestic abuse offences. This could be due to offender management or other interventions currently employed by the police force.

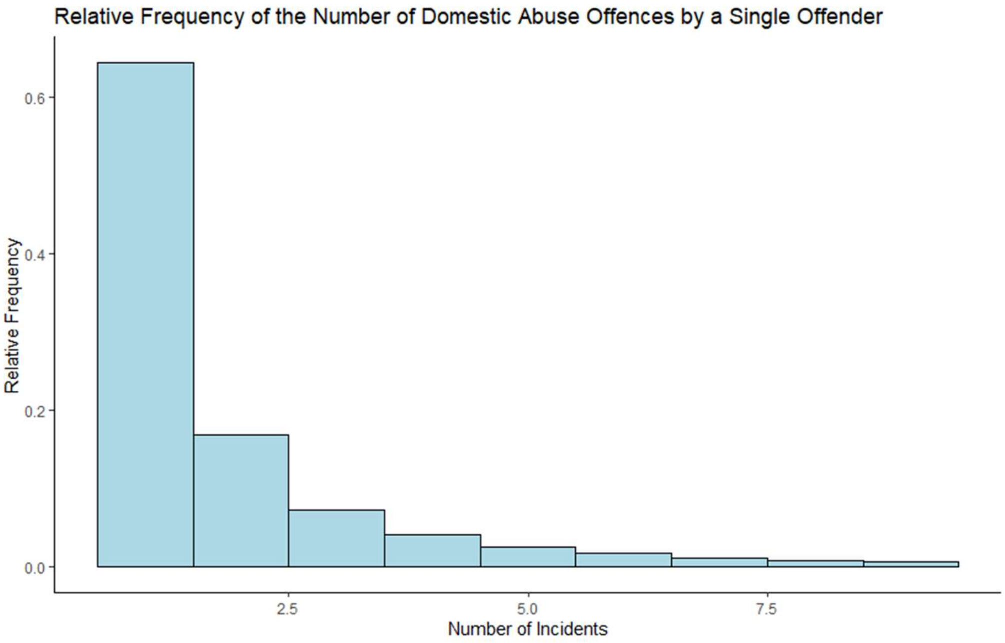
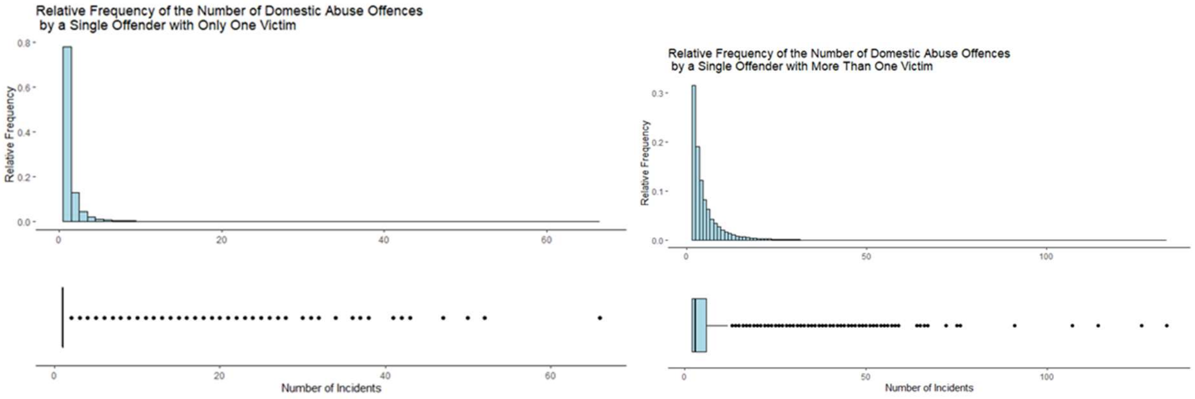


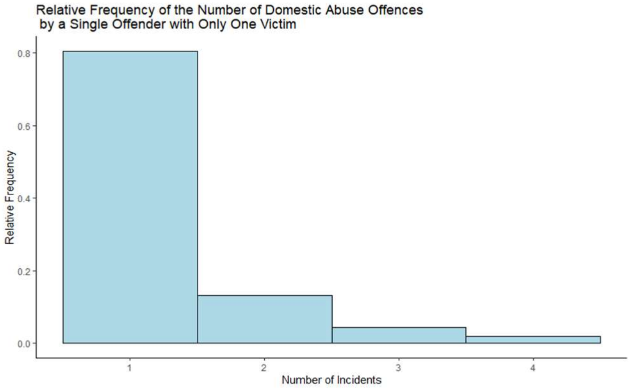
Figure [4]:

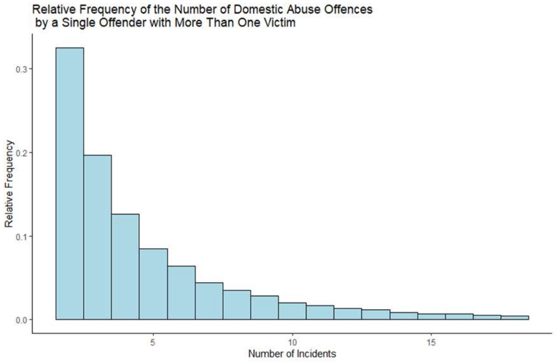
Multi-Victim Offenders

Splitting the dataset to analyse offenders with single and multiple victims separately, for offenders associated with only one victim, nearly 80% are recorded as being involved in a single incident only ,as shown in figure [4], whereas for offenders associated with more than one victim, approximately 70% will be involved in more than one incident. Repeating the outlier removal techniques utilised above results in figure [4] we can see that the range of incidents is greater for offenders with multiple victims (16) compared to that of single victim offenders (3). This combined with the relative frequency highlights how the number of victims affected by a given offender could be a useful factor in informing future abusive criminality.



Min	Lower Quartile	Median	Mean	Upper Quartile	Max		Min	Lower Quartile	Median	Mean	Upper Quartile	Max
1.0000	1.0000	1.0000	1.1.452	1.0000	66.0000		2.0000	2.0000	3.0000	5.2360	6.0000	133.0000





Min	Lower Quartile	Median	Mean	Upper Quartile	Max		Min	Lower Quartile	Median	Mean	Upper Quartile	Max
1.0000	1.0000	1.0000	1.2780	1.0000	4.0000		2.0000	2.0000	3.0000	4.6160	6.0000	18.0000

Relationship Between Number of Victims and Frequency of Incidents

Figure [5] shows a scatter plot of the number of victims and the number of incidents for each offender as well as the correlation coefficient between the two variables. A significant strong positive correlation of 0.68 is observed, implying that as the number of victims increases the number of incidents will also increase. This further suggests that the number of victims is an informative variable for predicting future domestic abuse incidents and justifies the additional attention (and weighting) received by this component.

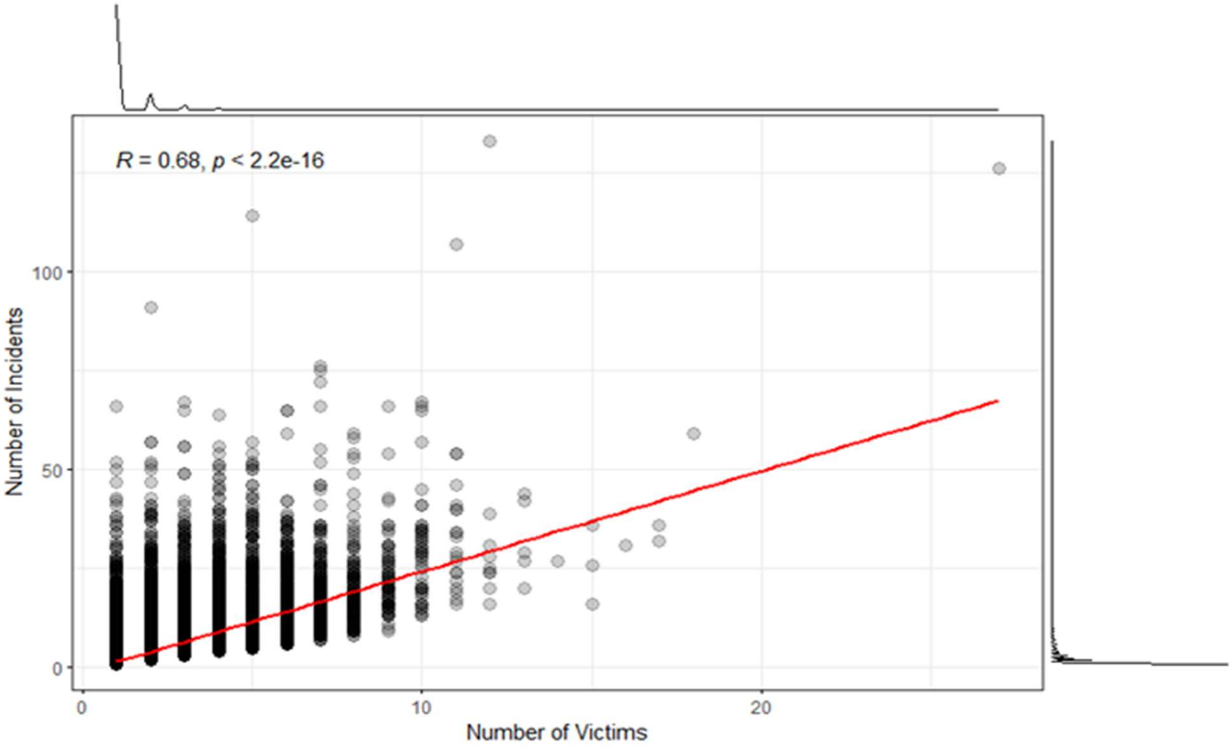


Figure [5]:

Time Interval Between Offenders' Incidents

The purpose of the RFGV algorithm is to identify the so-called high harm offenders that are likely to offend again as well as those who are likely to escalate in their actions. However, as discussed in section below, the implementation of the RFGV algorithm used by the various forces is dependent on the pre-filtering of the data prior to the analysis. This is problematic as it could lead to offenders disappearing off the algorithms radar before committing another domestic abuse. It is therefore pertinent to analyse the time between incidents to gauge whether the algorithm is able to identify high-harm offenders in a timely manner.

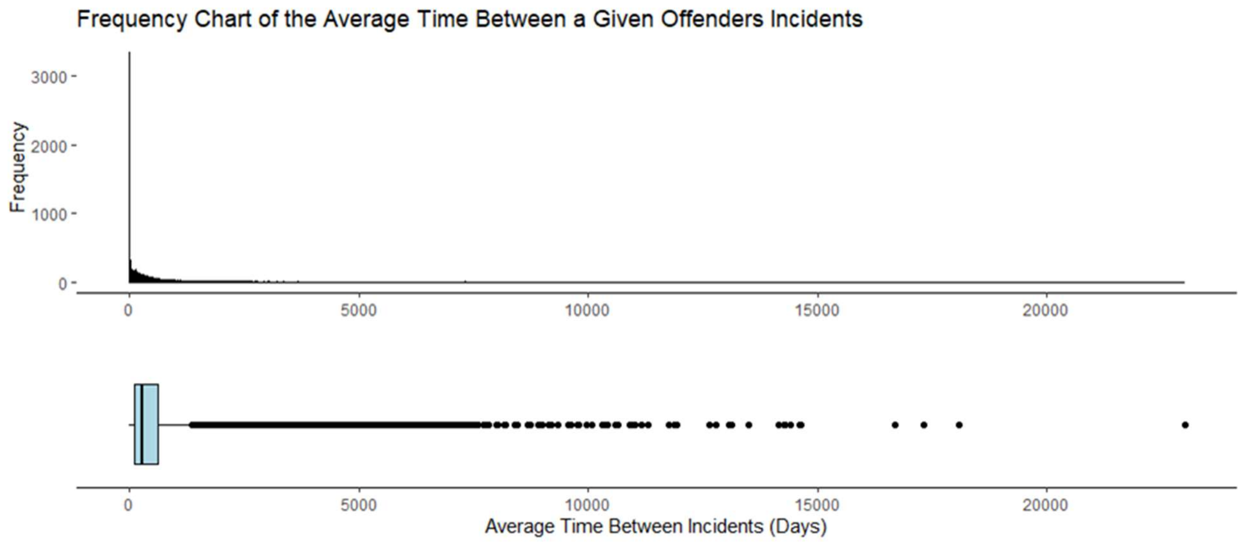


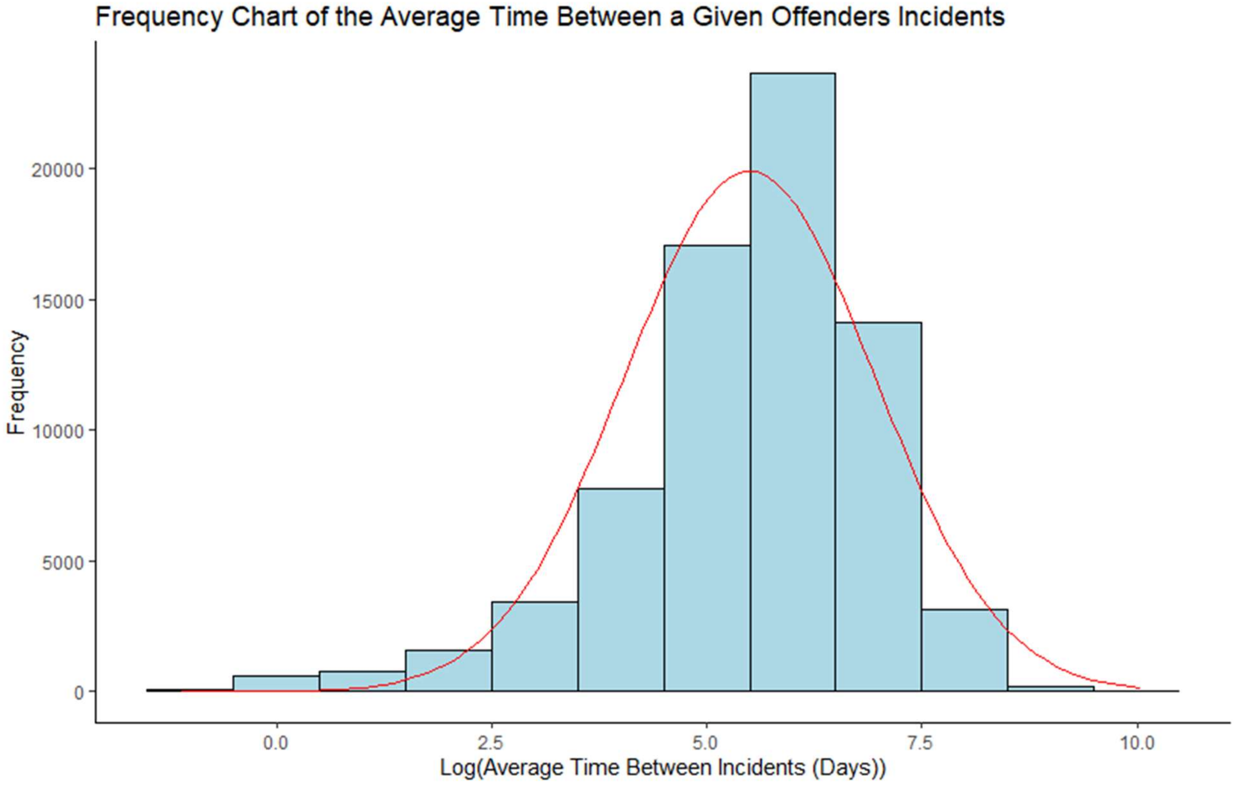
Figure [6]: ……………………. Note that there are 0 days between incidents as incidents involving multiple victims are counted separately.

Min	**Lower Quartile**	**Median**	**Mean**	**Upper Quartile**	**Max**
0.0	99.5	269.0	483.0	604.5	23011.0

Table[6]:

Plotting the frequency of the average time between incidents in days we can see that the median time between incidents is 269 days (note that the mean will be more heavily affected by outliers) with 50% of the data sitting between 99.5 and 604.5 days. The WMP algorithm looks for offenders active within the past 3-months or approximately 90 days. This means that approximately 75% of offenders brought to attention by the algorithm are not likely to offend again until they are no longer considered relevant by the algorithm.

Furthermore, due to the fact that single incidents with multiple victims are recorded as separate records we consequently observe the spurious 0 day average between incidents. It is safe to assume that all offenders with an average of 0 days between domestic abuse incidents are the result of a single incident involving multiple victims. These can therefore be excluded from the dataset since these are skewing the data with a false time interval. Removing also the outliers that lie outside of the 97.5th percentile and performing a logarithmic transformation results in figure [6], which shows that the average time between incidents is approximately normally distributed. Normally distributed data (after log-transformation) is indicative that we discovered some kind of natural feature of DA incidents instead of random noise in the data. This grants some credence to the data collected by the police as it reduces the chances that the data collection was biased.



Relationship Between the Average Time Between and the Number of Incidents

It can be assumed that the more incidents there are in a fixed time frame, the smaller the interval between them will be. However, we may find that serial offenders go on a "spree" before disappearing due to effective offender management. It remains an unanswered question whether the RFGV algorithm can actually provide an operational response to serial offenders or their serial offending occurring over a short time period.

We produced a scatter and correlation plot (Figure 7) of the average time between and the number of incidents to explore the association between the two. The wavy lines in the figure represent marginal distributions of each variable which were also overlaid to provide an informative overview of each respective distribution due to the number of coincident points. As a preliminary step, we removed any outliers and then employed a logarithmic transformation on both variables as the correlation coefficient requires normality. The correlation coefficient was less than -0.01, indicating that there was no significant association between the two variables. In other words, the number of incidents and the gap between those incidents seem to be unrelated to each other. Therefore, people involved in high number of incidents are not more or less likely to wait longer/shorter periods of time before reoffending.

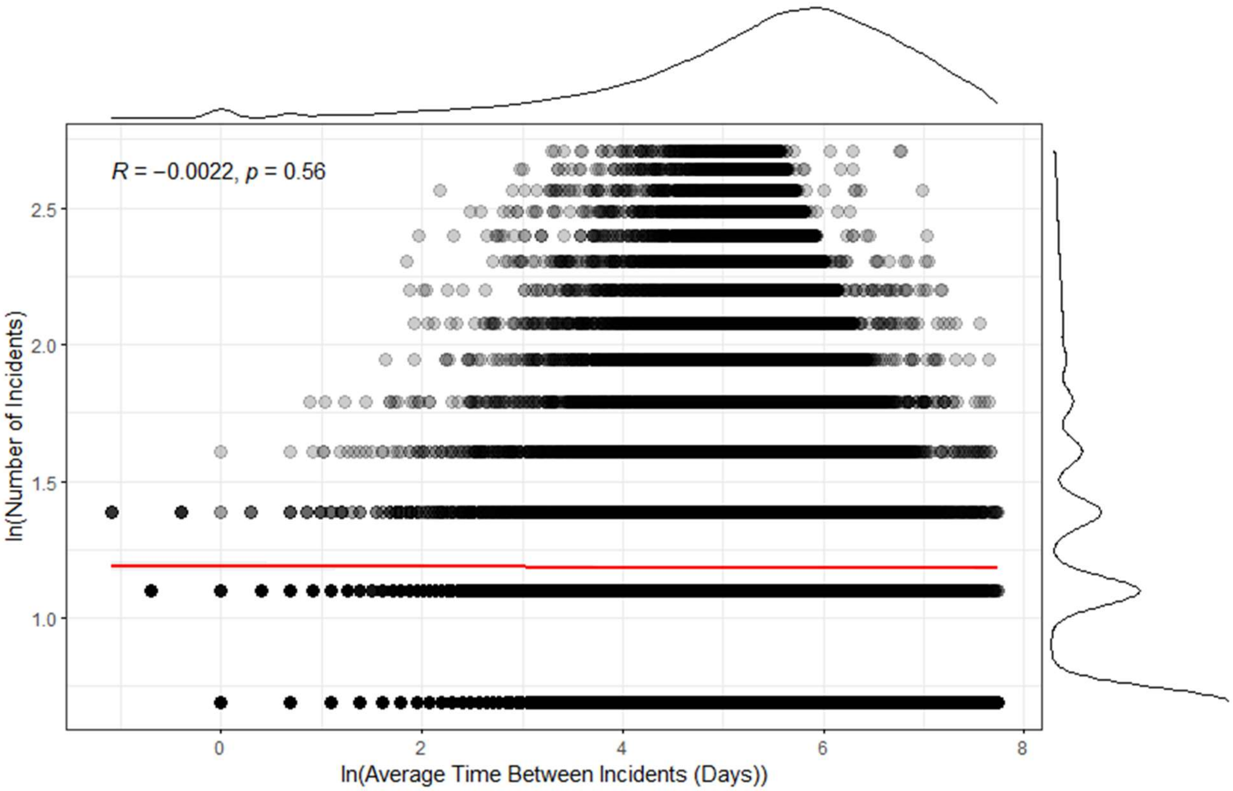


Figure 7: Scatter plot of time between and number of incidents

Although no linear correlation is observed it may be due to the fact that there is an underlying pattern in the occurrence of incidents, for example a drop off corresponding to effective offender management by the police. We can analyse any such pattern by considering an offender's first domestic abuse incident to occur on day 0 and then by taking each subsequent incident in relation to the first to produce an offender's domestic abuse timeline. This was completed for all offenders in the dataset who were involved in more than 1 incident producing Figure [8]. The figure shows that incidents are more likely to occur in the nearby temporal vicinity of the first incident and that the likelihood of an additional incident decreases over time. However, as these time windows are really small, and because the algorithms are not meant to be 'nowcasting' tools that would advise police action in real time, they will not be able to effectively prevent most of these incidents. It follows, that a modified version of the 'recency' component is also unlikely to improve police practice.

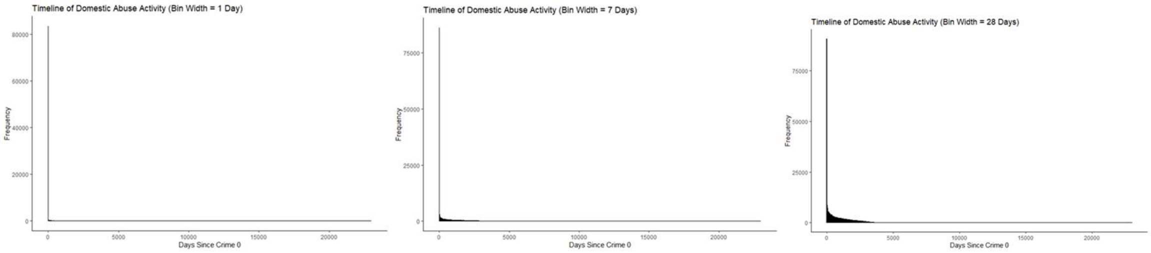
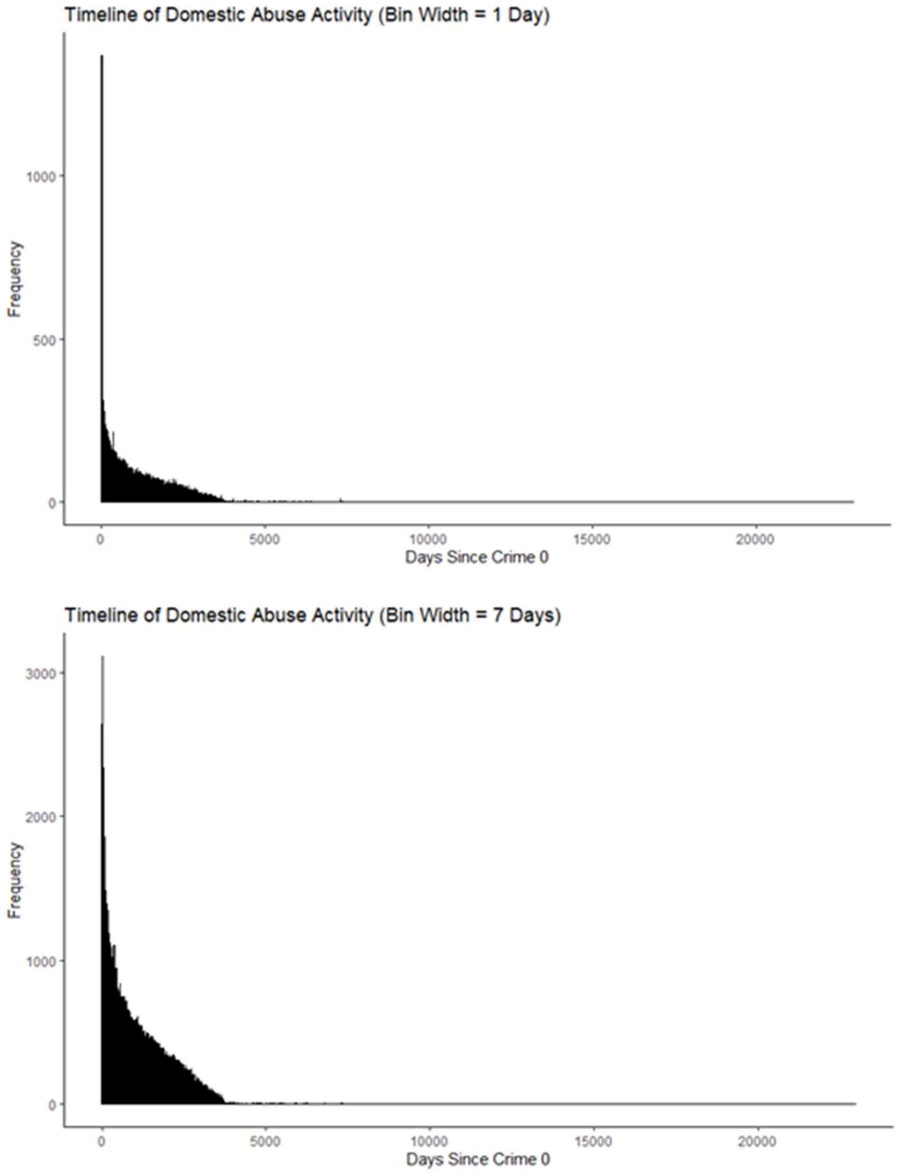
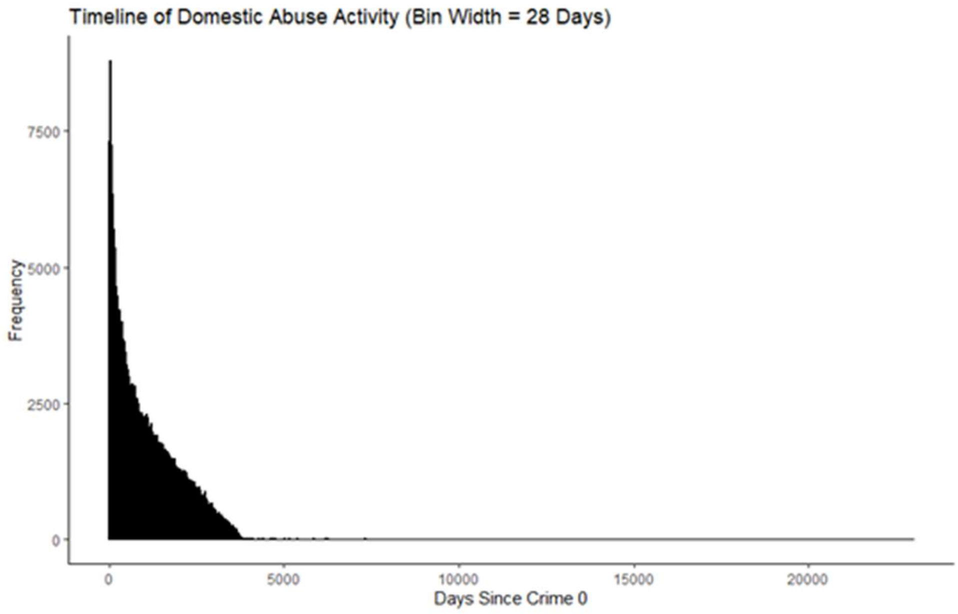


Figure [8]:



Timeline of Domestic Abuse Activity (Bin Width = 1 Day)

Timeline of Domestic Abuse Activity (Bin Width = 7 Days)



For the following four sections we return to the interviews and explore from the perspective of stakeholders how the outputs of the algorithm are used, the extent of human inputs to the algorithm, how success of the algorithm is measured, and if the algorithm is helpful to practitioners.

Use of outputs: Triaging and Offender Management

The process of triaging and interventions offered to identified perpetrators was quite different in the two forces.

Offender Management was very different in both forces. In Northumbria the intervention was provided by police staff called Domestic Abuse workers and the focus was on serial perpetrators who were not being otherwise statutorily managed either by probation or other agencies. Thus, most intervention was dependent on perpetrators 'voluntarily engaging' with the DA workers. These perpetrators are approached with an offer of help or support to change their behaviour but are told that if they don't then the police will disrupt their behaviour in any legal way possible. This 'carrot and stick' approach adopted by police staff was considered by interviewees as having been very successful, especially the fact that DA workers were emphatic, and identified themselves to perpetrators as staff and not police officers, which was considered to encourage much greater engagement.

In Northumbria, then, the DA workers were police staff who worked with identified perpetrators and offered them support to help them to reduce their reoffending. The algorithm provides scoring and ranking for perpetrators for each of the six geographic areas. On an average the MATAC unit reviews the top 100 or so offenders in each area and works with as many as their DA workers have capacity. At this point the analysts do some triaging by going through the list of the top 100 or so offenders and conducting what they call a 'quality assurance process' whereby the analyst goes through all their data and remove those who are in prison or have moved away or are under some other kind of statutory management, before presenting a short list of 10 or so candidates for each area to the MATAC manager. At this point further analysis of individual cases by the manager allows them to pick one or two individuals who would be considered suitable for MATAC intervention. These are taken to the biweekly MATAC meetings, which are separate from the MARAC process, and discussed in this multi-agency forum. There are two DA workers to then engage with offenders for the entire force are. At any given point in time, each of the two DA workers for the force are engaging with 15 offenders. They also take in referrals from other agencies who might be aware of a particular perpetrator of concern based on their professional judgement, but who might have escaped the notice of the police because of low RFGV scores.

In West Midlands, which is divided into several areas with a DA Team for each area, the goal is to manage the top 5 % of the top ranking perpetrators. However, since the numbers are so large, some areas manage to only focus on the top 3%. The DA teams are very well integrated into the MARAC process, and these meetings are held every day. The team manager is able to access the output of the RFGV algorithm for their area and they are responsible for managing the top 3 or 5% of their ranked perpetrators. They also have to provide justifications for why they choose not to manage any particular offender in that list. One of the main driving forces for triaging is capacity to take on more cases,

> “It's all down to capacity as well because we have teams and each member of my team will have a limit to the number of offenders they can manage. So my team carry a cohort of 15 offenders.” (Police officer, West Midlands)

In case of those perpetrators who were subject to some kind of statutory management, the OMs would liaise with the agency responsible for managing that offender. One interviewee articulated a three stage policy, what they called “a sliding gauge” for dealing with identified perpetrators: change, control and disrupt. The interviewee explained that the aim is to help support the offender change their behaviour by assisting them in whatever way; if they refuse to respond, then the intervention is ramped up to find ways of controlling their offending behaviour through civil injunctions or warnings. In the worst case, the force would issue a Community Protection Warning, described as:

> “It's just a piece of paper so we can put a CPW on somebody, which means that we could say – ‘you're not allowed to go to this address’, and it can be their partner's address, so you know we're talking about, quite, you know, infringing on human rights a little bit here. But because we can say your behaviour is causing a disturbance in the area, we can put on a restriction, so perhaps they can't go into a certain area, or they can't go to a certain house or and will put very basic stuff on there as well that you cannot drink in a public place and you know we can put some really obscure things on a CPW which you know try to try and get some control when they breach a CPW.” (Police officer, West Midlands).

Finally, as a last step the police adopt whatever means possible to disrupt recalcitrant perpetrators, for example, for arresting them for anti-social behaviour, driving a vehicle without insurance or without a valid MOT for roadworthiness, or enforcing bail conditions very stringently to try and arrest them for something.

A perpetrator is managed or supported for as long as necessary, but on average the intervention and contact lasts for 3 to 6 in Northumbria; whereas in West Midlands, it can last between 2 weeks to 6 months. The difference is because one force works with perpetrators who are not under statutory supervision or not in prison or otherwise irrelevant, but the policy in the latter is to manages the top 5% or so of all perpetrators identified by the algorithm.

The interviewees unanimously agreed that there are always far more perpetrators in need of intervention than they have capacity for – but that the RFGV provides a good tool to help them prioritise their cases.

Human algorithm interaction

Human decision making is involved at both ends of the algorithm – firstly in construction of the variables that feed the algorithm (especially the subjectively created weighting scores for DA related offences in Northumbria); in making decisions about how much data to use as also how often to run the algorithm. As mentioned above, decisions about gravity scores had a subjective element; decisions about how much data (time period) to include in the algorithm was determined by the analyst and officers involved based on what they thought was a reasonable time frame and taking into consideration the data load their systems were capable of handling; finally, the decision about how often to run the algorithm (monthly in both forces) was made in consultation with offender managers who found this was the most effective time period for them to get everything in place and was dependent on the volume of work involved. Similarly, at the other end of the algorithm human decision-making was at the forefront, in deciding and prioritising which perpetrators would be offered some support or otherwise managed (see above).

Further, in Northumbria interventions are not offered to perpetrators purely on the basis of the RFGV scores. When presented with the list of ranked perpetrators, relevant stakeholders make a conscious decision whether or not to select the perpetrator for intervention and would not include a person just because their scores were inflated :

> “And I might look at them and say I'm not going to take them because the score is inflated purely and simply because of the frequency. If there's nothing else around that I would probably not take it.” (Staff, Northumbria).

Interviewees agreed that the weighting given to the frequency scores could be revisited, especially if it were possible for the algorithm to somehow distinguish between frequency of offending against one victim as against multiple victims. The interviewees went on to say that there is a subjective element of assessing whether there were particularly concerning issues like a sudden escalation of frequency or seriousness of offences or other factors and then decide how many and which perpetrators would be supported or managed.

The decision-making process was somewhat different in West Midlands, where lists were generated by the algorithm specific to each of the various force areas and DA teams. Offender managers were responsible for managing the top 3 or 5% (whatever has been agreed as the target for that area), regardless of whether these this included perpetrators who might be in prison, had moved away or were otherwise statutorily managed. The DA team manager made decisions about whether they had the capacity to focus on the top few or all perpetrators in the top 5% (depending upon the number of offenders and resource capacity of the unit) and when it was acceptable to deselect a perpetrator for management. Managers had to record their decisions if they decide to stop managing an offender or wanted to make a case for managing someone who is not in the top 5%. Given the fact that West Midlands focused on active perpetrators in the preceding three months, in effect the number of overall perpetrators they manage is around the top 1% identified by the Northumbria algorithm. Thus, even if the stated policy is to manage the top 5% of perpetrators, if compared with Northumbria on similar data filtering terms, we could say that both forces manage approximately the top 1% of offenders identified by the algorithm.

Thus, in West Midlands it seems as if the cut off points where a perpetrator has to be managed become significant, thus making it more important for the algorithm to be accurate in its calculation of those who are causing most harm. But interviewees said that they were less concerned about any possible ethical issues because,

> "The algorithm is not trying to predict risk, there is distinct difference between what we know and this is what our systems and data are telling us, we are simply reporting back on what has happened and we know rather than trying to predict risk and harm. We cannot offender manage thousands of people only manage a certain percentage. And this way we can be fairly confident that we are working with those around whom there is signification harm. There is a reality to it, we have to have a cut-off point at some point, there is a reality to that" (Police officer, West Midlands)

Given the extent of discretion used to finalise which offenders would ultimately receive an intervention in both forces, interviewees were not very concerned about the finer nuances of the calculation used by the algorithm.

Measuring success

When asked what the measures of the success for the RFGV algorithm as well as offender management efforts were, interviewees in Northumbria said that their measure of success was the fact the nearly 80% of the perpetrators selected for intervention voluntarily engaged with the DA workers and there was a decrease in reoffending. Interviewees suggested that the RFGV scores of those perpetrators selected by the MATAC process and who engaged with the DA workers fell in the 6 months following the intervention.

In West Midlands, the measure of success was expressed by one interviewee thus:

> "So every month I have to do some data for my supervising officer, so I will say how many people we have? How many people we're currently managing? How many people I've deselected this month? How many new people we've taken on? How many people have reoffended? And what is happened to those people that have reoffended? And any good news stories? Because there's been no further offending then have we got controls in place? Any good control, new control we've put in place? So it's kind of measured by success, either in non-offending, or by putting that control in place to prevent offending" (Police officer, West Midlands)

One interviewee admitted that the algorithm has a blind spot in that it is not able to incorporate information such as if the perpetrator is in prison or is not offending for any other reason, so whether the algorithm is working or not is difficult to evaluate.

It seems as if the measures of success were more in terms of actions taken rather than outcomes achieved. It might be the case that there aren't adequate resources to conduct a proper evaluation of the impact of the algorithm and offender management in these forces, which begs the question of how successful is the algorithm in preventing serious DA offending? This might be a moot point if the aim of the algorithm is merely to prioritise use of resources rather than predicting and preventing high risk perpetrators. As one interviewee commented,

> "I don't view it as a job done, but as a journey. Have we got the perfect algorithm? Probably not. What we had before, it was almost a random process of making decisions on subjective decision-making - this is better. It can always be improved, and I am open to questioning it and be open to changing it." (Police officer, West Midlands).

Support for the algorithm

Reportedly, there was a great deal of support for the use of the algorithm in both forces, particularly from practitioners responsible for DA and working with offenders. One interviewee said,

> "For me to research one individual would take a very long time because we have so many different police systems and they don't all connect to each other… But then we've still got intelligence logs on one computer system. We've got a back load of history on an old computer system. We've just got so many computer systems they don't talk to each other that to research somebody fully, uh, it just takes so, so long. So having the algorithm is a very quick and easy way of me going, 'Great, they are serial [perpetrators] and they fall in my 5%'" (Police officer, West Midlands)

Another argument in favour of the algorithm was the protection it offered by way of a justification for why particular perpetrators (who for example, go on to commit a domestic homicide) were not identified by the DA teams earlier, especially if they were not serial offenders.

> "You know we need something… It helps us. It saves time, it gives us that protection as well when we are decision-making." (Police officer, West Midlands)

In both forces interviewees had few concerns about how fine-tuned the algorithm ought to be and whether it was able to actually identify perpetrators causing most concern. The aims and use of the algorithm were more modest – to identify serial perpetrators based on retrospective data and provide support or interventions to encourage them to desist from reoffending. Interviewees were therefore content with the fact that the algorithm was saving them the effort of trawling through many databases and trying to rank or score perpetrators and was providing them with a 'transparent' and ostensibly objective way of justifying their focus on a few individuals that they had the capacity to engage with.

Conclusions

Our overarching conclusion is that the RFGV algorithm is neither considered nor being used as a risk assessment tool; nor is it purporting to make predictions of who is most likely to reoffend. Instead, the primary purpose of the algorithm is as a ranking tool in order to prioritise police resources to provide interventions in the form of support or disruption to serial DA offenders who are not otherwise under any mandatory supervision. It is to identify not the most violent or most harmful offenders but "those who are flying slightly under the radar" (staff Northumbria). Consequently, here is less concern with whether potential offenders were missed out. It seems acceptable that the algorithm provides a justification that is objective, transparent, less cumbersome, and a defensible process for identifying DA perpetrators the police should be working with. Whether the algorithm is actually efficient in predicting those most likely to reoffend does not seem to be a priority for the forces.

It is also clear that even with a sample of three forces there is significant inconsistency in how the algorithm is constructed and is being used. Indeed, the MPS algorithm (as shown in Appendix 1 below) is so different to those used in Northumbria and WMP it cannot really be considered the same thing at all. It seems highly likely this variability would be replicated, and probably increase, in a wider sample of forces. It is beyond the scope of this report to comment on whether this variability is a problem. One could make the argument, however, that what matters is that the algorithm enhances the effectiveness and efficiency of police work, and that a wide variety of local solutions may be appropriate in achieving this end. It is also the case that while a national solution might have the appeal of consistency, it seems that at present this could only be achieved through restricting some forces to the lowest denominator of computing power, which would seem counter-productive to say the least.

On the other hand, we found that the setup of the preliminary filtering was the key difference between the forces' use of the algorithm. Based on the qualitative interviews, it seems clear that two factors determined this process. First, each force wanted an *efficient* system, with a sufficiently small time-window for extracting the relevant data and running the algorithm. Second, they were also mindful of *limiting the number offenders* considered to a manageable level. Notably, both of these factors were decided based on the grounds of feasibility: efficiency was limited by the available computing power; whilst limiting the number of offenders was necessary due to resource constraints at each of the forces (i.e. the number of officers who can be enrolled into various interventions). While we understand and appreciate these limitations, we note that the practice of shaping an algorithm based on the computational restraints is problematic, as this can lead to erroneous results. Similarly, the number of offenders considered should be driven by an evidence-based approach, instead of resource limitations. Determined by the local context of the forces involved, it is distinctly possible that these inconsistencies on the input side produced inconsistencies, and therefore injustice, on the output side (i.e., offenders and victims with comparable behaviours and experiences are being treated differently in the two forces).

The ability of the algorithms, as currently constituted, to deal with coercive control is also open to some question. This is most obvious when it comes to the calculation of harm. Whilst the respective forces have adjusted the gravity scores (with WMP also introducing non-crime scoring) these are primarily based on the average sentence length – and thus likely to significantly under-estimate the harm caused by controlling behaviour that on its own constitutes only a minor crime, or indeed no crime at all – but it is well know that there is a systematic bias in police systems and behaviour toward identifying and recording physical assaults and against identifying and recording coercive control (REF?). Currently, a perpetrator is much more likely to be flagged by the algorithm if they use physical violence than if they employ non-physical coercion, because much controlling and coercive behaviour remains as a 'non-crime incident' and is downplayed in the calculation of harm.

Given that it is considered mainly to be a tool to help police officers make decisions, rather than being a decision-making tool, practitioners are very satisfied with the algorithm as it helps them prioritise their resources. There is an awareness that the tool might not be perfect and there is a willingness to work towards refining the tool. With that as a starting point, we offer some tentative conclusions and recommendations based on our study. We divide our conclusions and recommendations into two distinct parts: those pertinent to practitioners; and those related to directions of future research:

Preliminary conclusions for practitioners:

Based on the above analysis we can draw some tentative conclusions about how the algorithm is currently being used in forces:

1. Firstly, the list of perpetrators scored and ranked by the algorithm does not directly determine who will be the recipient of offender management. It is a sorting exercise used as a basis for further triaging by humans to identify potential candidates suitable for intervention.

If possible, it is worth considering whether the triaging process can be refined and made more systematic and transparent so as to produce equitable and fair results.

2. Secondly, decisions made about cut off points for the time period for which data will be considered for running the algorithm, for shortlisting who is considered potentially suitable for intervention, and for actually providing an intervention, is very much determined by computing capacity and resource constraints.

If possible, it is worth considering a bigger time-window during the data filtering. The two-year time-window used by Northumbria is likely capture most of the repeat offenders[8], as a 730-day window captures more than 80% of the data points for these offenders.

3. Thirdly, there are concerns about how gravity is measured and whether there is a better method for weighting it which has been raised in the interviews.

[8] We recognise that this recommendation might be limited by the size of the dataset and the capacity of force systems to cope with the added demands.

We recommend that is worth having a national conversation on how DA offences should be scored on harm rather than relying on the HOI or the Cambridge CHI. Furthermore, it is also unclear whether taking the maximum severity/harm into consideration is a better practice, or taking the sum of them (as one by the Metropolitan Police).

4. Fourthly, concerns about the impact and relevance of the recency scores were revealed in the analysis.

Our finding supports removing recency from the algorithm, for multiple reason. Recency scores are heavily influenced by the preliminary data filtering, a 'calendar lottery', while potentially giving counter-intuitively large focus on most recent new offenders and lower weight on repeat offenders. As implied by the subsequent analysis of the time gap between multiple offences, it appears unlikely that a more recent DA incident would lead to subsequent ones in a time window that could be effectively captured by the algorithm. We believe that modifying the preliminary data filtering window should be an effective tool to take into consideration the relevant cases.

5. Finally, current evaluation of whether the algorithm 'works' (i.e. leads to a reduction in recidivism) is focused more on the number of perpetrators selected and 'successfully managed' or restricted to checking whether the RFGV scores of the perpetrators who were provided with an intervention dropped in the subsequent 6 months following the intervention.

Nevertheless, these conclusions are only tentative, as we were unable to factor in the impact (if any) of any applied methods of offender management or other policing interventions offered by the forces. Without this additional information we cannot exclude the possibility that the above interruptions in reoffending behaviour were at least partially due to some intervention measures put in place by police services.

Recommendation for future research:

- Carrying out similar interviews and analysis as presented in this report with the practitioners at the MPS in order to better understand the implementation of the algorithm and the rationale behind some of the decisions made in its construction and use. The quantitative part of this endeavour would have a special focus on (1) whether using harm scores instead of severity scores improved the algorithm and (2) how the current risk scores are decided by officers (i.e. text analysis of their responses to the survey).

- Leading a national level discussion with experts and relevant stakeholders from the police, government, NGOs, and academia, to devise an acceptable new scale to score DA offences on harm that is not related to potential sentencing (crime harm index) or 'relative harm' of an offence (crime severity score). This new score could replace the 'gravity' component of the

RFGV, and the ‘harm’ component of the RFHA algorithm. The potential impact of any change due to the implementation of this new scale could be estimated and demonstrated on past data.

- Conducting a detailed evaluation study of how the triaging process carried out by each force works, does it produce equitable and ethically defensible outcomes and whether a more objective process can be introduced to aid decision-making. This would be coupled with a quantitative assessment of people who are being managed by the police compared to those who are not thus, evaluating the impact of the intervention[9].

- Assessing whether alternative variables gathered by the forces have stronger predictive power of involvement in future DA-related incidents. Using past data, this research could help refining the existing algorithms and offer a better prioritisation tool for the forces. The proposed analysis would incorporate all available information about the individual, how/whether the police managed them, and their prospective outcome. Being mindful that adding personal information into an algorithm could create biased policing outcomes, any new algorithm would have to be weighted to produce fair results.

[9] This is not to replicate previous evaluations of multi-agency approaches to manage DA perpetrators, for e.g. the evaluation of MATAC (Davies, P., and Biddle, P. 2017. *Multi Agency Tasking and Co-ordination (MATAC): Tackling perpetrators of domestic abuse. Evaluation Report*. Newcastle-upon-Tyne: Northumbria University) but to focus on offender management approaches and assess their impact on re-offending behaviour.

Appendix 1

Development of the RFHA Algorithm by the Metropolitan Police Service

Unlike the RFGV algorithm used by Northumbria and WMP, the MPS devised its own algorithm: the RFHA. In this section, we will introduce the components of the RFHA and highlight the differences with the RFGV.

Due to time and resource constraints at the MPS they were not able to supply us a dataset large enough to allow us to carry out robust assessment. However, due to the peculiarity of the RFHA and the RFGV, direct comparison would have been unlikely to be possible.

The development of the RFHA algorithm

Initially, the Metropolitan Police Force had been using a variation of the original RFG algorithm developed by Strathclyde Police with an alteration to the scoring feature to include the DASH risk assessments, as opposed to the serial victimisation used by the Northumbria and West Midlands Police Forces. Our conversations with a data analyst at the Metropolitan Police revealed that their data infrastructure does not record the number of victims of each of the perpetrators thus, there was no way for them to upgrade the original RFG algorithm to an RFGV algorithm. Hence, their original RFG algorithm had four components: recency (R), frequency (F), gravity (G), and the current risk assessment of the individual (based on the DASH risk assessment, but not included in the acronym).

Recently, the Metropolitan Police Force modified the RFG algorithm by replacing the gravity component (which originated from the Crime Severity Score) with a harm score based on the Cambridge Crime Harm Index leading to what can be coined the RFHA algorithm. The difference between the two is that the crime harm index uses sentencing guidelines to estimate the harm caused by each crime whereas the crime severity score uses the average sentence imposed on offenders for each crime category. Whilst from afar these differences may seem nominal, research has shown that choosing one over the other can have “major influences on the conclusions drawn” (Ashby, 2017). A further difference with the RFHA is that instead of only considering the most serious offence of an individual (i.e. the ‘maximum’ score) like Northumbria and West Midlands do, the harm score for an individual adds up the harm caused by each of the previous offences (i.e. they consider the sum of the harm). The RFHA algorithm still considers the recency (R), frequency (F), harm (HA), and the current risk assessment (still missing from the acronym).

Preliminary data filtering for and components of the RFHA

The details of the algorithm were disclosed to us mainly through various process documents and a single conversation we had with the person responsible for deriving the RFHA scores for the Metropolitan Police. Based on our understanding of the provided information, the following steps are taken to estimate the scores:

- The preliminary data filtering takes 10-months' worth of domestic abuse offence and incident data, but only 6-months of nominal relationship data, any offences or incidents without a linked nominal on joining are then removed.

- For the incidents, all relevant pieces of information are recorded in a different way compared to the other forces. For each incident, the Met records both the 'perpetrator' and the 'victim' as victims, which is being done for legal considerations, as the Met had been advised that they cannot hold the name of suspects without charging them and sharing their information with the Home Office. The RFHA still considers incident-level information accrued by adding the information of both the victim and perpetrator to the dataset (thus 'flipping' their status from victims to perpetrators).

- The recency (R) component is calculated differently from Northumbria and West-Midlands, taking the most recent event into consideration with the assigned score conversion shown in Table [x] (instead of the average recency used by the other forces).

- The frequency (F) component differs in that it measures the average time in days between an offender's recorded incidents in comparison to the RFGV algorithms which calculate the number of events within the period of interest, [Scoring details available in Table X]. Each person in the database is considered an offender.

- The harm (HA) component is a standardised [0-100] version of the Cambridge Crime Harm Index (CCHI) obtained by looking up the Home Office Code and taking the CCHI divided by 54.75. It should also be noted that if a single record has multiple crime classifications, then the score may be greater than 100 as each of these is included and counted (i.e. summed up).

- The risk assessment component score is based on the DASH risk assessment result which can be either S, M, or H resulting in a component score of 25, 50, and 100 respectively for 'small', 'moderate', and 'high' risk individuals. Any non-classified records receive a 0 score. Furthermore, in the instance a case is flagged as strangulation, the risk assessment score is boosted to 100 (i.e. the maximum possible score). Crucially, only a subset of the DA suspects go through the DASH risk assessment (a survey containing 19 questions). This means that the limited data provided to us indicated a high proportion of missing values (in the dataset received by us, 79.25% did not receive a risk assessment score). The algorithm assigns 0 to all missing values.

- Each of the four components is treated with equal weighting for a given offender.

- As a final step, after having run the algorithm, they only consider the top 10% of those offenders who scored at least one standard deviation away from the mean (i.e. the top 1.358% of the sample). They run the algorithm every fortnight. They could not provide any information on how they use the scores after being delivered.

Score	**Days Since Event (x)**
100	x ≤ 14
75	15 ≤ x ≤ 30
50	31 ≤ x ≤ 60
30	61 ≤ x ≤ 90
20	91 ≤ x ≤ 120
10	121 ≤ x ≤ 150
5	151 ≤ x ≤ 180
0	181 ≤ x

Table 1A: Table Detailing the Metropolitan Police Recency Component Score Calculation Based on the Number of Days Since the Most Recent Event.

Score	**Average Days Between Incidents (x)**

100	$10 \leq x$
75	$5 \leq x \leq 9$
50	$x = 4$
38	$x = 3$
25	$x = 2$
10	$x = 1$
0	$x = 0$

Table 2A: Table Detailing the Metropolitan Police Frequency Component Score Calculation Based on the Average Number of Days Between an Offenders Incidents

[1] https://committees.parliament.uk/writtenevidence/36621/html/

Comparing the RFG, RFGV, and the RFHA

As indicated by the above description, the RFG, RFGV, and the RFHA are not comparable for multiple reasons below:

First, and foremost, the way the Metropolitan police process their data means that, on the incident level, perpetrators and victims cannot be distinguished from each other. This means that RFHA scores are being calculated for individuals who were victims of a given incident, not only the perpetrators. Although it is extremely unlikely that these victims would end up being flagged by the system (after all, only the top 1.358% are actively being considered by the police for further action), this still means that even if all the variables needed were present, a direct comparison with the RFGV would be impossible.

Second, the way frequency is measured by the RFG and RFHA is also quite different, as not the number of DA-related incidents/crimes are being considered, instead the number of these activities during a certain period.

Third, and along similar lines, the fourth component of the RFGV, serial victimisation is not being recorded by the current data infrastructure of the MPS, which means that this variable is essentially missing from their database.

Fourth, the RFG and the RFHA consider the sum of the harms done by an individual, instead of only the maximum value (i.e. the most serious one). While it would be possible to translate the gravity/harm component of the RFG or RFHA to the gravity component of the RFGV, this would not be possible the other way around, as the RFGV only records the most serious (i.e. most severe) instance, which can mask repeated violations, especially if they are lower level.

Finally, the risk assessment scores are not being used by the other forces. This is a subjective element as the assessment depends on an officer's personal deliberation, although, in the case of strangulation, a crime will be automatically considered 'high risk'. What remains unclear (without the ability to carry out further interviews) is when and why a certain individual receives a current risk score. Notably, the other forces used to rely solely on officer risk assessment and moved away from it thanks to the emergence of the algorithm. By contrast, the Met incorporated the risk assessment into their newly adopted algorithm thus, preserving the main element their initial practice.

A rapid evidence assessment of the research evaluating the use of offender-centric algorithms in law enforcement

Introduction: The use of algorithms in law enforcement

The use of algorithms in law enforcement has become increasingly common with the advancement of computer science, including large-scale data extraction and machine learning techniques (Brayne & Christin, 2021). They are perhaps most commonly associated with the prioritisation of police resources on the basis of geographical locations; using geographic information systems (GIS) to predict the temporal and geographical locations of future crime (Jefferson, 2018).

As well as being used to prioritise geographical areas, however, algorithms can also be offender-centric, used to analyse and make judgements about individual offenders (Oswald & Babuta, 2019). Some of the most widely used algorithms are used in a range of circumstances to try and predict an offender's risk; that is, their risk of recidivism (reoffending). These types of algorithms can be used throughout the criminal justice system (CJS) process; for instance, an offender's risk of recidivism will be relevant during a police investigation, during decisions being made about awarding bail once an offender has been apprehended, and during parole hearings after an offender's incarceration (e.g. Davies et al., 2021; Hamilton, 2019a). Algorithms have, therefore, the potential to play a part in the decisions made about offenders at several stages of the CJS process and could be used by a number of different law enforcement agencies (Kehl et al., 2017).

What is meant by an algorithm?

The simplest definition of an algorithm is 'a finite list of instructions, most often used in solving problems or performing tasks' (Kosek, 2018); an equation which is calculated based on set, standardised criteria. When we think about something like the different types of risk assessments that are conducted with offenders in prison settings, for example, the actuarial tools used (e.g. the Sex Offender Risk Appraisal Guide, or SORAG; Rice & Harris, 2016) could be defined as algorithms in the way that they used standardised data input in the same way to produce a result, whereas the use of structured clinical judgement tools to assess and manage risk (e.g. the HCR-20, or Historical Clinical Risk Management-20; e.g. Murray & Thomson, 2010) might be more considered more subjective in that professional judgement plays a role in influencing the way scores are calculated and personalised risk assessment cannot be considered algorithmic.

When discussing the use of offender-centric algorithms within a law enforcement setting, however, there are two additional criteria which seem to be assumed within the literature, and are thus added here into our definition of an algorithm:

1) The algorithm has, at least in part, some automated aspect (Huq, 2019); and
2) The data used in the algorithm are, again at least in part, taken automatically from a database or databases that were not designed for the purpose of running the algorithm (e.g. police incident databases).

The literature discussing the use of these tools draws attention to the fact that there is an element of harvesting data from CJS databases and there is an element of computerisation and automation involved in the process. What makes it an algorithm is that all the data (regardless of whether harvested from systems or based on professional judgement) is subject to the same uniform process to arrive at an assessment for all offenders under consideration.

The potential benefits of using offender-centric algorithms

There are several potential benefits to the use of offender centric-algorithms. As with other computerised tools, these types of algorithms remove the possibility of human error in calculation, although it does not account for errors in transposing data which is often done by humans. This includes both errors of data transposition when transferring data from one database to another, as well as errors made during the calculation process. Second, they may be able to eliminate elements of human bias that could exist in the entry or interpretation of data. Third, they are able to cope with interpreting large levels of data that would be impossible for humans to do. With law enforcement having to process large amounts of information, and with research suggesting that bias can exist within law enforcement officials towards certain groups of people, these are benefits that may mean decisions made about offenders are conducted more fairly.

The potential issues with using offender-centric algorithms

There are, however, several potential issues with using offender-centric algorithms. While it could be argued that algorithms could eliminate bias, it could equally be argued that they may perpetuate the systematic bias seen in human decision making (Alikhademi et al., 2021). While it is generally agreed that it would be incorrect and unethical to include race as a factor in such algorithms, there is no reason why factors that could act as proxies for race (those that are highly correlated with race) couldn't be used (Kehl et al., 2017). In this way, the inequalities seen in our CJS could be played out in offender-centric algorithms that perpetuate these biases towards certain groups of people.

One of the other issues with many of the algorithms used in the CJS is that they are produced by private companies (COMPAS, for instance, was developed by a company called Northpointe; Hamilton, 2019), who are often unwilling to share the mechanisms of the algorithm on the basis that it is proprietary information which they have a vested interest in keeping confidential (Wexler, 2018). When the details of such algorithms are kept confidential, this makes it very difficult to independently evaluate their validity and reliability. Further, it also makes it difficult for individuals to challenge decisions made by these algorithms. The State versus Loomis (2016) in the US is a well cited example of where the use of COMPAS was challenged by a defendant, on the basis that the judge's use of an algorithm-generated risk assessment score violated his right to due process (Kehl et al., 2017).

The current study

Despite the fast-growing body of literature on the use of algorithms in law enforcement contexts, a review has not yet been conducted which looks specifically at the literature evaluating how well these algorithms are working. The current study was a rapid evidence assessment of the published literature which evaluates the efficacy of offender-centric algorithms currently being used by CJS agencies. The purpose of the REA was to assess the number and types of offender-centric algorithms being used, the countries and type of law enforcement agencies that are using these types of algorithms, and how able they are to fulfil their purpose. We are also interested to explore whether any of the potential benefits and challenges that are outlined in the more general literature on algorithms were specifically addressed within the context of these evaluations.

Choosing an REA as a method of review

A rapid evidence assessment is a way of conducting a quick-time systematic review of the available evidence on a topic, which is limited in its scope or depth because of the rapidity with which the review is conducted but that nevertheless aims to present a balanced view of the available literature (Barends, Rousseau, & Briner, 2017; Berry, Briggs, Erol, & van Staden, 2011).

The search parameters

The search terms were established through consultation with the research team, as well as some initial searches run to establish the number of documents that were returned. The search terms included various combinations of and synonyms for three terms: criminal justice, risk assessment and algorithms. Four databases were searched using the searches constructed above for relevant material which included ProQuest, Science Direct, Scopus and PsycINFO. These databases were chosen for the breath of material that they cover including several different disciplines10.

The following inclusion and exclusion criteria were established, which formed the basis of the screening process as outlined below:

- INCLUDE – in order to be included studies had to fulfil the following requirements:

[10] See Appendix A for further details on search terms and for small amendments made to the searches during the search process.

- They had to be empirical studies evaluating actuarial tools where at least some automation of the process exists AND
- The tools had to be offender focused AND
- Tools had to be in use by one or more criminal justice agencies
- Had to be in English
- Had to be published post 2000[11]

- EXCLUDE – the criteria for exclusion were:
  - Studies that were reviews, commentary pieces, descriptive pieces, or were not evaluating any specific tool
  - Studies that were not peer reviewed – this included grey literature.

The search procedure

All of the searches were run on 5th May 2021. Appendix B outlines the number of articles that were returned in each of the searches. Additionally, a limited number of hand searches were conducted on relevant websites in order to supplement the information found in the constructed searches.

An assessment of the studies returned by our search terms indicated that a number of tools were called algorithms if subjective risk assessment scores made by practitioners were reduced to a number or category (such as low, medium or high risk) and formed part or whole of the risk assessment tool. Since our accepted definition of 'algorithms' was focused around those tools that keep subjectivity to a minimum when calculating the outcome (i.e. a set of fixed rules are used to combine a set of fixed input variables to produce a risk score), we have included only those studies where it was indicative that the bulk of the automated process recovered data from information systems or risk assessment instruments, but where subjectivity was kept to a minimum[12] in the actual calculation of the risk score.

Once the searches had been compiled, they were uploaded to EPPI Reviewer. EPPI Reviewer is a data management software[13], which provides researchers support, guidance, and training for those looking for specific research evidence and is specifically designed to aid the conduct of systematic reviews. Once in EPPI reviewer, duplicates were removed automatically. The remaining documents were then

[11] The decision to exclude grey literature, studies published in other languages and prior to 2000 were the result of limitations arising from project timescales, language skills of the research team, and to ensure that technologically relevant studies were included.

[12] We do acknowledge that this process is imperfect as it is reliant on information provided in the study, which often did not explain the antecedants of the variables in detail.

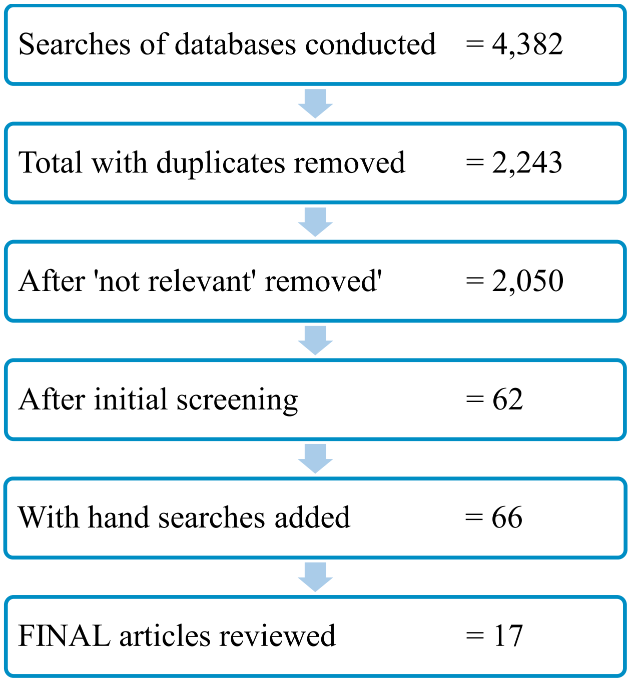
[13] The software is managed by the EPPI Centre https://eppi.ioe.ac.uk/cms/

subjected to screening on title and abstract. The included studies were then screened on the full document. Screening and coding was conducted by two researchers and was based on a code book. At the initial screening on title and abstract stage, a preliminary Interrater Reliability test was conducted on a random sample of 100 studies with two aims: to check that the researchers had understood and were applying the inclusion criteria appropriately; and in order to refine the codebook. There was 78% agreement in the scores in the first instance and any disagreements were resolved through discussion. This exercise was useful in clarifying the inclusion criteria as well as what we would consider as the definition of an algorithm for the purposes of this REA. This was admittedly the most difficult part of the screening process.

Full texts were screened to assess which were relevant for final inclusion. The final list of included studies were coded by two researchers who met frequently to resolve any questions or doubts through discussion. The coding strategy for these documents can be seen in Appendix C.

The search process is presented visually in Figure 1. All texts deemed relevant to the REA were included for further analysis, without discrimination on the basis of quality assessment. This was a conscious decision given the small number of studies that met our inclusion criteria. Furthermore, most of the studies did not really contain adequate information to be able to judge them on quality and would therefore return very poor quality assessment scores.

Figure 1


Searches of databases conducted = 4,382
Total with duplicates removed = 2,243
After 'not relevant' removed' = 2,050
After initial screening = 62
With hand searches added = 66
FINAL articles reviewed = 17

Findings

A total of 17 studies fulfilled our inclusion criteria including 16 primary studies and one systematic review. Of these, 13 studies evaluated algorithms being used in the United States, with the remaining studies looked at algorithms in use in Japan, Spain, Sweden, and the United Kingdom. The agencies using these algorithms included probation, police, and prison and corrections agencies.

The algorithms evaluated in the reviewed literature were overwhelmingly risk assessment tools for the prediction of recidivism, with 15 out of 17 articles pertaining to some sort of recidivism.

As mentioned, one of the 15 studies was a systematic review of the predictive accuracy and validity of various risk assessment tools, both actuarial and those based on clinical judgement (Singh et al., 2011). A total of 10 studies were focused on evaluating some aspect of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm (Brackey, 2019; Dietrich et al., 2013; Dressel & Farid, 2018; EPIC, 2014; Farabee et al., 2010; Hamilton 2019a; 2019b; Hartmann & Wenzelburger, 2021; Lin et al., 2020; and Tan, 2019). Five studies looked at other risk assessment instruments such as PACT (Positive Achievement Change Tool; Baglivio, 2009), MnSTARR (Minnesota Screening Tool Assessing Recidivism Risk; Duwe & Rocque, 2017), RisCanvi (Karimi-Haghighi & Castillo, 2021), and Automated Telephony (Vasiljevic et al., 2007). One of the remaining studies looked at an algorithm aimed to predict harm caused by offenders as well as to prioritise high-harm offenders for management (Davies et al., 2021). The final study focused on an algorithm used by investigators to prioritise offenders (burglary mainly) who might fit the criminal profile based on their criminal *modus operandi* (Yokota & Watanabe, 2002).

Thus, most of the tools included in this study were recidivism prediction tools, with some indication algorithms are also being used for the purpose of prioritisation of offenders for resource allocation purposes. Recidivism was generally defined as reoffending in some capacity, with some specificity as to the type of reoffending, such as violent reoffending or sexual offending. These tools were predicting risk and were used to guide either offender sentencing or management decisions. Only one tool was used to predict harm caused by offenders and was used to prioritise resources based on the notion of harm[14]. In terms of the prediction of harm, this is potentially much more difficult to measure, given that the notion of harm is much more subjective than recidivism and therefore harder to define and to measure. While there are tools designed to quantify harm (e.g. the Cambridge Crime Harm Index; Sherman, Neyroud, & Neyroud, 2016), they are not without their limitations, and therefore how harm is defined and measured may have a great deal of impact on the perceived efficacy of an algorithm.

Given the difficulty in synthesizing the findings of the studies, we make a few observations on the content of the included studies and the state of the existing evidence base:

[14] Davies et al 2021

Study aims

The studies were very disparate in their aims and therefore it was difficult to conduct any meaningful synthesis. For example, some studies were focused in testing the predictive accuracy of an existing algorithm by itself (e.g. Baglivio, 2009; Brackey, 2019; Dietrich et al., 2013; Hartmann, 2021; Vasiljevich et al., 2017; Yokota & Watanabe, 2002); or in comparison with another existing tool (e,g, Farabee et al., 2010); or with an algorithm created by the researchers (e.g. Karimi-Haghighi & Castillo, 2021; Lin et al., 2020); or as against risk assessment done by humans (e.g. Dressel & Farid, 2018; Duwe & Rocque, 2017; Lin et al., 2020); or compared actuarial tools with tools based on clinical judgement (Singh et al., 2011); or against a hybrid model of human algorithm assessment (Tan, 2019).

Some studies were exclusively focused on testing whether the predicted results were fair and unbiased (see e.g. Hamilton 2019a; 2019b; Brackey, 2019). Finally, one study was focused on evaluating the core composite of the norm group for COMPAS in relation to one specific jurisdiction (EPIC, 2014). Thus, given the wide array of study aims, this REA only draws very high level findings from the evidence presented.

Algorithm accuracy

Overall, the studies indicated that for the most part algorithmic predictive accuracy was, at best, as good as human risk assessment when the amount of information was limited but was better than human ability to predict accurately when a large amount of data had to be processed (Lin et al., 2020). A number of studies concluded that although the algorithm was able to predict more accurately than humans, the predictive accuracy was just within the acceptable range (which was around 66%), for example, the predictive accuracy of the COMPAS algorithm was found to be between 66 to 70% by a number of studies (Brackey, 2019; Dietrich et al., 2013; Dressel & Farid, 2018; Farabee et al., 2010; Lin et al., 2020) and was far from perfect for the MnSTARR (Duwe & Rocque, 2017). One study found that actuarial tools of risk assessment were more accurate for particular kinds of offenders rather than general offenders (for e.g. sexual offenders) and was more accurate for predicting violent offending as compared to general offending (Singh et al., 2011).

Furthermore, two studies found that humans were able to predict recidivism with similar accuracy as COMPAS with lesser information. However, as the size of data, (i.e. items of information that fed into the risk assessment of individual offenders, as well as the number of offenders being assessed), increased, the algorithm's performance was better than that of humans (Dressel, 2018; Lin et al., 2020). Duwe and Rocque (2017) concluded that automation (MnSTARR) can improve predictive accuracy as compared to risk assessments made by prison caseworkers.

Predicting and Prioritising

Whilst most studies included in the REA evaluated tools that were predictive risk assessment tools, two studies looked at tools used for prioritisation purposes (Davies et al., 2021; Yokota & Watanabe, 2002). As mentioned above although the algorithm evaluated by Davies et al. (2021) was used to predict harm, it was also used to prioritise offenders for offender management purposes (Davies et al., 2021). This demonstrates that algorithms do not necessarily have to be used for prediction; rather they can be used as a tool to assist law enforcement in the identification of offenders who require urgent attention. Prioritisation tools also require evaluation, though, for instance, the algorithm used by Surrey Police was designed to identify high-harm offenders, and is reliant on using a measure of harm that accurately encapsulates what the force's definition of harmful is. This is important because actions designed to manage offenders identified by these tools may not reduce reoffending if the identification was inaccurate.

Another algorithm was also used to prioritise offenders in police investigations, based on the identification of similar *modus operandi* behaviours in burglary (Yokota & Watanabe, 2002). This is, however, perhaps more of an offence prioritisation tool, as it is possible that two crimes may be suggested as linked. Thus, officers are identifying potential series of offences through which greater efforts could then be focused on apprehending supposed repeat offenders. Evaluation indicated the tool had an accuracy rate of 20% underscoring the need to evaluate accuracy of prioritisation tools, especially if the resulting processes produce unfair outcomes for offenders based on race, gender or other characteristics.

Advantages of using algorithms

The studies indicated that there were a number of advantages in using algorithms as they were more accurate than humans in predicting risk (Lin et al., 2020), were capable of processing more data, quickly, and efficiently (Dressel & Farid, 2018), thus saving costs and staff time (Duwe & Rocque, 2017). Studies mentioned that one of the main advantages for practitioners of using algorithms was that since they were perceived as being objective and unbiased, practitioners found them useful, especially because it enabled them to shift responsibility for morally problematic outcomes on to the algorithm (Hartmann, 2021). The latter was considered an advantage by practitioners, although admittedly that is not an advantage from the viewpoint of offender management.

Challenges in using algorithms

One of the perceived benefits of using algorithms to predict risk is that it addresses issues of bias and errors inherent in human decision making. However, this notion was shown to be erroneous in the study conducted by Dressel and Farid (2018) which showed that humans showed greater accuracy and less bias when predicting recidivism risk using fewer data items as compared to COMPAS. This changed, however, when the data was enriched, as humans were less accurate when processing large amounts of data.

Some studies indicated that unless the composite norm group against which scales are created for an algorithm such as COMPAS is similar to the group of offenders to which the algorithm is applied, the results can be very skewed against certain groups (Brackey, 2019; EPIC, 2014). Hamilton's (2019a; 2019b) studies demonstrated that COMPAS overpredicts risk for Hispanics and women to confirm earlier studies that showed that the tool is biased against black and minority ethnic communities. Brackey (2019) concluded that COMPAS clearly fails to treat recidivists and non-recidivists equally and produces unfair outcomes based on gender and race.

Although studies such as Duwe and Roque (2017) asserted that the use of algorithms and machine learning saved a lot of money and resources in terms of staff time, which could be put towards providing interventions for the offenders instead of calculating risk - nevertheless, they acknowledge that there are initial costs involved improving IT capability, and we can add, in training staff or personnel in how to use the software and work with the algorithm.

Furthermore, Hartmann and Wenzelburger's (2021) study showed that although algorithms are used to underpin decision-making by judges and prison workers and are considered mainly to be a way of reducing uncertainty, evidence suggests that decision-makers fail to account for the fact that it is merely a tool that provides statistical probability. Instead, algorithms are considered as providing definitive answers - thus highlighting the dangers of basing sentencing or parole decisions purely on algorithms.

Challenges involved in evaluating algorithms

One of the biggest challenges in evaluating algorithms was the lack of transparency or any information about how the algorithm processed data, especially algorithms like COMPAS that were owned by private companies (Dressel, 2018).

The lack of access to data also meant that a number of studies evaluating COMPAS used the same data that was originally published in a study conducted by ProPublica. The biggest challenge, which was not articulated by any of the studies, but seems obvious is the lack of any information about what intervention (if any) was provided to the offender once they were predicted as being high risk and how that might have had an impact on testing the predictive accuracy. For example, if the usual practice was to increase sentence length for high risk individuals or to manage them in some way (provide support, employment opportunities etc) so as to move them away from reoffending, then these offenders are unlikely to reoffend – thus reducing the perceived accuracy of the tool. Unless there is information about what was done with the identified high risk offenders, any evaluation will be flawed.

Another issue with the studies reviewed here was the lack of information about control groups (if any) in many of them. There was also little or no contextual information about how decisions were made about what data would feed into the algorithm, and what weightage was given to different pieces of data, or how variables were combined, makes it difficult to evaluate the worth of the algorithm. Issues of confidentiality were also highlighted as making it difficult to evaluate algorithms (citation wanted here? Or errant bracket?!

Finally, studies indicated that despite the fact that algorithms were being used to guide decision-making by judges and officials, it did not mean human judgement and discretion was totally absent (Hartmann, 2021), this makes it difficult to evaluate the impact of the algorithm, given there is virtually no information about how this human judgement is exercised and what impact that has on the outcomes for the offenders.

Discussion

A review of the evidence on offender focused algorithms indicated that although there is a substantial body of literature on the use of algorithms by the criminal justice systems and on the ethics of using algorithms in law enforcement, there is little evaluation related evidence in this area. Furthermore, this evidence is quite varied given that the study aims were very varied as discussed above making the task of synthesizing the evidence very challenging.

We understand from the evidence reviewed that the use of algorithms to guide decision-making has strong support from practitioners mainly because it is quick, efficient, has the appearance of being unbiased, and absolves individuals of any responsibility for morally questionable outcomes. The ostensible objectivity and absence of bias in arriving at a risk assessment via an algorithm comes from the fact that all data is subject to the same process without regard to extraneous factors or individual judgement. It is important to acknowledge however, that any form of data has some subjectivity and element of human decision-making involved in its construction, for e.g. crime data is dependent on several subjective decisions made by practitioners about whether a reported incident is to be recorded as a crime or not, the specific offence classification, and the amount and quality of information to be entered etc. Furthermore, human decision-making is involved in the initial choice of processes and weightage given to various variables in the construction of the algorithm. Thus, we question claims that algorithms are neutral, but acknowledge the element of uniformity introduced by automating the process of calculating risk based on a large amount of data. It remains to be tested whether particular algorithms then produce just and equitable outcomes for all.

We modestly advance a few observations based on the findings:

- Most of the evidence reviewed relates to algorithms that are predicting risk of reoffending and are being used by criminal justice agencies to guide decision-making. Although there was evidence of fewer algorithms being used as prioritization tools, it is clear that both types should be evaluated in order to ensure they are achieving the requisite outcomes.
- The lack of access to data and the lack of transparency in how the algorithm calculates risk, makes the task of evaluation very challenging. It also makes it difficult to assess whether or how the algorithm could be improved.
- Lack of information about management of high-risk offenders identified by the algorithm (whether sent to prison, or moved away or stayed in the community) makes it difficult to assess the predictive accuracy of the algorithm in the follow-up period.
- Although algorithms are better than humans at processing large amounts of data and make more accurate predictions, studies indicated that those currently in use have a predictive accuracy at levels that are just below or in the acceptable range. However, there is wider evidence (not included in this study) to suggest that researchers have developed and conceived of smarter algorithms that used fewer variables and that theoretically can have a much higher level of predictive accuracy, that have as yet not been adopted by CJS agencies.

- The perceived neutrality and unbiased nature of algorithms was challenged by a few studies which demonstrated that algorithms can often produce unfair outcomes for particular groups, especially if the reference group used as the norm for developing an algorithm is different to the population to which it is now being applied.

Conclusion: Whither algorithms?

The review of existing evidence on algorithms indicated that there are gaps in the literature on the use of offender focused risk assessment algorithms that are currently being used by criminal justice agencies. It has indicated that there is a need for caution in the use of algorithms to guide decision-making, especially since they might produce unfair outcomes, might be inaccurate one out of three times, and their working is often opaque and hidden behind data protection laws. In conclusion, the review was able to shed light on what don't we know about offender focused algorithms: we don't know nearly enough about a) which algorithms are currently being used by the police, b) how they work, c) whether they are accurate, d) whether they produce fair and just outcomes for all, and e) how the results are interpreted and operationalized by practitioners.

Thus, we recommend that future research into the areas identified above focuses on the gaps identified in the review.

References

*(Studies marked with an * were included in the REA)*

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2021). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-021-09286-4

*Baglivio, M. T. (2009). The assessment of risk to recidivate among a juvenile offending population. *Journal of Criminal Justice*, *37*, 596-607. https://doi.org/10.1016/j.jcrimjus.2009.09.008

Barends, E., Rousseau, D. M., & Briner, R. B. (2017). *CEBMa guideline for rapid evidence assessments in management and organizations.* Retrieved from https://cebma.org/wp-content/uploads/CEBMa-REA-Guideline.pdf

Berry, G., Briggs, P., Erol, R., & van Staden, L. (2011). The effectiveness of partnership working in a crime and disorder context: A rapid evidence assessment. *H. Office (Ed.), Research Report*, *52*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/116548/horr52-summary.pdf

*Brackey, A. (2019). *Analysis of racial bias in Northpointe's COMPAS algorithm.* Doctoral dissertation, Tulane University School of Science and Engineering.

Brayne, S. & Christin, A. (2021). Technologies of crime prediction: The reception of algorithms in policing and criminal courts. *Social Problems*, *68*, 608-624. https://doi.org/10.1093/socpro/spaa004

*Davies, K., Woodhams, J., Abramovaite, J., Evans, E., Banerjee, A., & Bandyopadhyay, S. (2021). *Evaluating Surrey Police Force's High Harm Perpetrator Unit.* An official report submitted to the College of Policing, UK. Retrieved from https://paas-s3-broker-prod-lon-6453d964-1d1a-432a-9260-5e0ba7d2fc51.s3.eu-west-2.amazonaws.com/s3fs-public/2021-07/vvcp-evaluation-of-high-harm-perpetrator-unit.pdf

*Dieterich, W., Oliver, W., & Brennan, T. (2013). *Predictive validity of the COMPAS Reentry risk scales.* An outcomes study conducted for the Michigan Department of Corrections: Updated results on an expanded release sample. Retrieved from https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-MDOC_ReentryStudy082213.pdf

*Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*, 1-5. doi: 10.1126/sciadv.aao5580

*Duwe, G. & Rocque, M. (2017). Effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy*, *16*, 235-269. https://doi.org/10.1111/1745-9133.12270

*Electronic Privacy Information Center (2014). *COMPAS core norms for adult institutions.* Results from a psychometric study conducted for the Wisconsin Department of Corrections division of adult institutions. Research & Development Divison. Retrieved from https://epic.org/EPIC-19-11-08-NEDCS-FOIA-20191112-Northpointe-Self-Validation.pdf

*Farabee, D., Zhang, S., Roberts, R. E. L., & Yang, J. (2010). *COMPAS validation study: Final report.* Semel Institute for Neuroscience and Human Behavior. Retrieved from http://www.northpointeinc.com/downloads/research/COMPAS_Final_UCLA_08-11-10.pdf

*Hamilton, M. (2019a). The biased algorithm: Evidence of disparate impact on Hispanics. *American Criminal Law Review*, *56*, 1,553-1,581. Retrieved from https://sc.fd.org/sites/sc.fd.org/files/assets/training/Charleston_2019/Risk_Assessment_Materials_SC.pdf

*Hamilton, M. (2019b). The sexist algorithm. *Behavioral Sciences & the Law*, *37*, 145-157. https://doi.org/10.1002/bsl.2406

*Hartmann, K. & Wenzelburger, G. (2021). Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA. *Policy Sciences*, *54*, 269-287. https://doi.org/10.1007/s11077-020-09414-y

Huq, A. Z. (2019). Constitutional Rights in the Machine-Learning State. *Cornell Law Review*, *105*, 1,875-1,954. Retrieved from https://cornelllawreview.org/wp-content/uploads/2020/12/Huq-final.pdf

Jefferson, B. J. (2018). Predictable policing: Predictive crime mapping and geographies of policing and race. *Annals of the American Association of Geographers*, *108*, 1-16. https://doi.org/10.1080/24694452.2017.1293500

*Karimi-Haghighi, M. & Castillo, C. (2021). Efficiency and fairness in recurring data-driven risk assessments of violent recidivism. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (pp. 994-1002). https://doi.org/10.1145/3412841.3441975

Kehl, D., Guo, P., & Kessler, S. (2017). *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. Responsive communities initiative.* Berkman Klein Center for Internet & Society, Harvard Law School. Retrieved from https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf?

Kosek, P. (2018). *What is an algorithm? Definition & examples.* Retrieved from https://study.com/academy/lesson/what-is-an-algorithm-definition-examples.html

*Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*. doi: 10.1126/sciadv.aaz0652

Murray, J. & Thomson, M. E. (2010). Clinical judgement in violence risk assessment. *Europe's Journal of Psychology*, *6*, 128-149. https://doi.org/10.5964/ejop.v6i1.175

Oswald, M. & Babuta, A. (2019). *Data analytics and algorithmic bias in policing.* Briefing paper, Royal United Services Institute for Defence and Security Studies. Retrieved from https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/21729582/Babuta_Oswald_Data_Analytics_and_Algorithmic_Bias_in_Policing.pdf

Rice, M. E. & Harris, G. T. (2016). The sex offender risk appraisal guide. In A. Phenix & H. M. Hoberman (Eds.), *Sexual offending: Predisposing antecedents, assessments and management* (pp. 471-488). New York: Springer.

Sherman, L., Neyroud, P. W., & Neyroud, E. (2016). The Cambridge crime harm index: Measuring total harm from crime based on sentencing guidelines. *Policing: A Journal of Policy and Practice*, *10*, 171-183. https://doi.org/10.1093/police/paw003

*Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*, 499-513. https://doi.org/10.1016/j.cpr.2010.11.009

State v. Loomis, 881 N.W.2d 749, 774 (Wisc. 2016). Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjakIft-pLzAhUCZMAKHS_5C0QQFnoECAMQAQ&url=https%3A%2F%2Fwww.courts.ca.gov%2Fdocuments%2FBTB24-2L-3.pdf&usg=AOvVaw0VHzSCPPkWBFHMAeW4K6DV

*Tan, H. F. S. (2019). *Interpretable approaches to opening up black-box models*. Doctoral dissertation, Cornell University. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/67545/Tan_cornellgrad_0058F_11634.pdf?sequence=1

*Vasiljevic, Z., Berglund, M., Öjehagen, A., Höglund, P., & Andersson, C. (2017). Daily assessment of acute dynamic risk in paroled offenders: Prediction, predictive accuracy and intervention effect. *Psychiatry, Psychology and Law*, *24*, 715-729. https://doi.org/10.1080/13218719.2017.1308219

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal*, *17*, 131-160. Retrieved from https://ctlj.colorado.edu/wp-content/uploads/2019/03/4-Washington_3.18.19.pdf

*Yokota, K. & Watanabe, S. (2002). Computer-based retrieval of suspects using similarity of modus operandi. *International Journal of Police Science & Management*, *4*, 5-15. https://doi.org/10.1177/146135570200400102

Appendix A

List of search terms

- “criminal justice” AND “risk assessment” AND “algorithm”
- “criminal justice” AND “risk assessment” AND “automated”
- “criminal justice” AND “risk assessment” AND “artificial intelligence”
- “criminal justice” AND “risk assessment” AND “machine learning”
- “criminal justice” AND “decision making” AND “algorithm”
- “criminal justice” AND “decision making” AND “automated”
- “criminal justice” AND “decision making” AND “artificial intelligence”
- “criminal justice” AND “decision making” AND “machine learning”
- “criminal justice” AND “evidence based methods” AND “algorithm”
- “criminal justice” AND “evidence based methods” AND “automated”
- “criminal justice” AND “evidence based methods” AND “artificial intelligence”
- “criminal justice” AND “evidence based methods” AND “machine learning”
- “criminal justice” AND “policing” AND “algorithm”
- “criminal justice” AND “policing” AND “automated”

Additional search parameters

Scopus:	Article title, abstract, keywords searched
Scopus:	Parentheses not included
PsycINFO:	1967 – 2021 database searched
PsycINFO:	Parentheses not included
PsycINFO:	Abstract searched
Science Direct:	Parentheses included
ProQuest:	Filtered for peer reviewed documents
ProQuest:	All fields except for full text searched
ProQuest:	Parentheses not include.

Appendix B

Search Terms and Returns in individual databases

Search term	Pro Quest	Science Direct	Scopus	Psyc INFO
“criminal justice” AND “risk assessment” AND “algorithm”	43	195	36	5
“criminal justice” AND “risk assessment” AND “automated”	16	183	7	1
“criminal justice” AND “risk assessment” AND “artificial intelligence”	15	55	10	0
“criminal justice” AND “risk assessment” AND “machine learning”	25	50	20	4
“criminal justice” AND “decision making” AND “algorithm”	32	510	45	4
“criminal justice” AND “decision making” AND “automated”	14	466	8	2
“criminal justice” AND “decision making” AND “artificial intelligence”	27	148	22	0
“criminal justice” AND “decision making” AND “machine learning”	21	113	27	3
“criminal justice” AND “evidence based methods” AND “algorithm”	7	5	4	0
“criminal justice” AND “evidence based methods” AND “automated”	3	2	2	0
“criminal justice” AND “evidence based methods” AND “artificial intelligence”	4	2	2	0
“criminal justice” AND “evidence based methods” AND “machine learning”	6	0	2	0
“criminal justice” AND “policing” AND “algorithm”	17	712	14	0
“criminal justice” AND “policing” AND “automated”	9	871	8	0
“criminal justice” AND “policing” AND “artificial intelligence”	9	173	4	0
“criminal justice” AND “policing” AND “machine learning”	8	115	8	1
TOTAL	**256**	**3,600**	**219**	**20**

Appendix C

The final coding strategy for the REA included the following codes:

1. Context/Background
    a. Country
    b. Agency?
    c. Who is running the algorithm?
    d. Who owns the algorithm?
        i. Data
        ii. Software/equation
    e. Who is the target audience? (who is the offender)?
    f. Who is using the output?
    g. Aim of algorithm
    h. Description of algorithm
    i. Variables included
    j. Data used
2. Effect
    a. Outcomes measured
    b. Metrics used
    c. Finding
    d. Research design
3. Mechanism
    a. Justification for using the algorithm
4. Implementation
    a. Challenges and barriers
    b. Facilitators
    c. Ethical issues

Appendix D

A description of the 17 studies reviewed

	First author	Date	Country	Algorithm evaluated	Who used the algorithm	Algorithm purpose	Evaluation purpose
1	Baglivio	2009	United States	Positive Achievement Change Tool (PACT)	Florida Department of Juvenile Justice	Youth risk assessment	Algorithm efficacy (youth recidivism)
2	Brackey	2019	United States	COMPAS		Risk assessment (violent and non-violent); risk of failure to appear	Identification of any racial bias
3	Davies et al	2021	United Kingdom	High-harm algorithm	High Harm Perpetrator Unit, Surrey Police	Identification and prediction of high harm offenders	Algorithm efficacy (subsequent harm caused and offender prioritisation)
4	Dieterich et al	2013	United States	COMPAS	Michigan Department of Corrections	Risk assessment (violent and non-violent); risk of failure to appear	Algorithm efficacy (recidivism)
5	Dressel & Farid	2018	United States	COMPAS		Risk assessment (violent and non-violent); risk of failure to appear	Algorithm efficacy compared to human judgement
6	Duwe & Rocque	2017	United States	Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR)	Minnesota Department of Corrections	Risk assessment (non-violent, felony, non-sexual violent, first time sexual offending, repeat sexual offending	Interrater reliability and how this impacts on predictive validity

7	Electronic Privacy Information Centre (EPIC)	2014	United States	COMPAS	Wisconsin Department of Corrections	Risk assessment (violent and non-violent); risk of failure to appear	Evaluating the scale cutting points
8	Farabee et al	2010	United States	COMPAS	California Department of Corrections and Rehabilitation	Risk assessment (violent and non-violent); risk of failure to appear	Validation of risk scales
9	Hamilton	2019a	United States	COMPAS		Risk assessment (violent and non-violent); risk of failure to appear	Identification of any racial bias towards Hispanics
10	Hamilton	2019b	United States	COMPAS		Risk assessment (violent and non-violent); risk of failure to appear	Identification of any gender bias
11	Hartmann & Wenzelburger	2021	United States	COMPAS	On Eau Claire jail population, Wisconsin	Risk assessment (violent and non-violent); risk of failure to appear	How algorithm information is used by practitioners
12	Karimi-Haghighi & Castillo	2021	Spain	RisCanvi	Catalan prison service	Risk assessment for violence prevention in Catalan prison system	Algorithm efficacy (violent recidivism)
13	Lin et al	2020	United States	COMPAS		Risk assessment (violent and non-violent); risk of failure to appear	Algorithm efficacy compared to human judgement and other statistical models
14	Singh et al	2011		Systematic Review			Meta-analysis and comparison of algorithm efficacy of various RA tools

15	Tan	2019	United States	COMPAS		Risk assessment (violent and non-violent); risk of failure to appear	Algorithm efficacy compared to human judgement and other statistical models
16	Vasiljevic et al	2017	Sweden	Automated telephony	Swedish probation officers	Risk assessment – dynamic risk factors	Algorithm efficacy (recidivism)
17	Yokota & Watanabe	2002	Japan	Algorithm ranking offences by behavioural similarity	National Research Institute of Police Science	Prioritisation of suspects based on the behavioural similarity between offences	Effective prioritisation of offenders