Independent Report on the West Midlands Police RFSDi Harm Score and Integrated Offender Management Model

Alexander Babuta

November 2022

Contents

Executive Summary

This report presents the findings of an independent process evaluation of the West Midlands Police (WMP) RFSDi[1] harm score and Integrated Offender Management (IOM) model. This evaluation focused specifically on the beta-testing phase of the project, which was being conducted within two WMP Local Offender Management Units (LOMUs) at the time the study was undertaken.

This study did not seek to evaluate the technical performance of the application, nor to assess the overall impact of the project on the force's wider offender management processes. The evaluation sought to scrutinise the implementation of the beta-testing phase of the project specifically, and assess the potential benefits and limitations of the application in an operational policing environment.

The RFSDi score is a statistical harm score calculated for nominals in the force database who have previously been charged with an offence, corresponding to the overall level of harm associated with their current offending. Alongside this, the IOM model is a predictive model which produces machine learning forecasts at the individual level, calculating each offender's likelihood of escalating from low-level to more serious offending, or of offending at a scale that cumulatively leads to more harm generated. The RFSDi score and model outputs are presented to offender managers via an interactive dashboard. The overall objective of the dashboard is to enable WMP officers to more effectively manage risk and target preventative interventions towards the highest-risk offenders.

The RFSDi score and accompanying IOM model were rolled out for beta-testing in October 2021, for a period of approximately 8 months until May 2022.

The methods used for the process evaluation included semi-structured interviews, focus groups, and a practitioner survey of all WMP officers involved in the beta-testing phase of the project. The purpose was to examine the real-world use of the RFSDi score and accompanying IOM model in their operational context, through direct engagement with those required to use the application.

Key Findings

The evaluation did not establish sufficient evidence to support wider operational deployment of the RFSDi dashboard and accompanying IOM model. Further testing and evaluation are required to assess the potential benefits of the system if it is to be deployed for enduring use.

The research found that the practical implementation of the application is yet to result in the operational improvements for offender managers that were envisaged at the time it was rolled out for beta-testing. As a result of perceived deficiencies in the application discussed further below, PCs and Sergeants report that the RFSDi score and predictive IOM model are not routinely used as part of offender management processes.

The research highlighted a fundamental divergence in views between PCs and Sergeants on the one hand, and Inspectors on the other. Inspectors were significantly more positive regarding the new application, reporting that the RFSDi score and IOM model are useful, and represent an improvement over the previous application (Corvus). In stark contrast, no PC or Sergeant reported finding the RFSDi score or IOM model useful, and none agreed that the application represented an improvement over Corvus.

[1] Recency, Frequency, Severity, Drugs and Intelligence

This divergence in views is partly explained by the fact that Inspectors interviewed for the research had been more closely involved in the development of the application and are therefore likely to have a more detailed understanding of its strengths and limitations. This demonstrates the critical importance of ensuring wide consultation and engagement in the early planning stages of a new data-driven system, to ensure end-users have had sufficient opportunity to contribute to the early development process.

Three main reasons were identified for the negative feedback reported by PCs and Sergeants. The first relates to the user experience of the dashboard itself, with officers reporting that the application interface is not user friendly and is difficult to navigate. The second relates to perceived deficiencies in the underlying statistical application leading to erroneous outputs, damaging officers' confidence in the overall validity of the system. PCs and Sergeants report experiencing a high number of over-classification errors, with a disproportionately high number of nominals being scored as 'high' or 'super high', resulting in an unmanageably large list of individuals to review. As well as over-classification errors, officers also report a large number of 'missing nominals'; individuals who should be scored as high-risk but are not being identified by the application (either because they have not yet been charged with an offence, or because the system relies solely on data held on WMP's internal database). Third, officers report receiving insufficient training and written guidance regarding the application before being required to use it operationally.

If these issues were to be resolved, Inspectors emphasised three main strengths of the RFSDi harm score which they believe could provide operational benefit. The first is improved precision of targeting, with Inspectors reporting that the tool should enable officers to more precisely monitor changes in an individual's harm score and track an increase or decrease in their level of risk. The second perceived benefit is increased confidence in the risk assessment process, as the factors that contributed to each score can be identified and triangulated across other data sources, providing another source of information to support individual-level risk assessments. The third perceived benefit is the identification of 'hidden risk'; the ability to identify individuals who are not currently subject to offender management orders, but are flagged by the system as requiring further, in-depth risk assessment. It is important to note that these perceived benefits described by Inspectors were not reflected at the operational level in the experiences reported by PCs and Sergeants.

It is also notable that the majority of interview and survey responses relate specifically to the RFSDi harm score, rather than the predictive IOM model. The distinction between the two components of the application was not clearly recognised by most respondents. When prompted, interviewees had relatively little insight into the predictive modelling element of the application specifically. This is concerning, as the outputs from the predictive model represent a probabilistic forecast associated with an inherent degree of uncertainty, and should therefore be interpreted in a different way to the purely descriptive RFSDi harm score.

Looking beyond the current performance of the application, the research identified three priority areas for future improvement. The first is training and guidance, with most respondents reporting that they received insufficient training and guidance on the application before being asked to use it. The second is the inclusion of more selection criteria within the dashboard, to allow officers to filter according to particular crime types. The third relates to integration and cross-compatibility with other data systems, most notably the Connect information management system. Several interviewees suggested that RFSDi should be integrated within Connect, to allow officers to view custody images and other intelligence for nominals who are assessed using the RFSDi dashboard.

In conclusion, the evaluation has not established sufficient positive evidence in favour of deploying the RFSDi dashboard and accompanying IOM model for long-term operational use. The system

should be subject to further, detailed evaluation research to conclusively establish its benefits and limitations before it is deployed for wider operational use. This should include the development of a detailed evaluation plan, including measurable criteria against which to assess the ongoing business case for the project and demonstrate that it is delivering its intended outcomes.

Recommendations

1. The RFSDi dashboard and IOM model should be subject to further evaluation research to conclusively establish their benefits and limitations before being deployed for wider operational use.

2. The force should establish a clear impact evaluation plan to measure the outcomes of the RFSDi dashboard and IOM model on an ongoing basis. This should include the development of measurable evaluation criteria and a basic logic model or Theory of Change to describe the intended outcomes of the project, as outlined in the College of Policing's Evaluation Toolkit.

3. Offender Managers who are required to trial a new data-driven risk assessment tool should be consulted at an early stage in the project development, giving them an opportunity to directly contribute to the application development process. An initial survey of end-users should be distributed, requesting feedback on the user interface and design requirements for any new system.

4. Any future development of the RFSDi score should focus on identifying nominals not currently subject to offender management orders who should be subject to more in-depth risk assessment. To avoid the risk of false negatives (which could lead to high-priority nominals being erroneously de-selected), individuals already subject to offender management orders should be excluded from the RFSDi harm scoring system.

5. The research was inconclusive regarding the potential benefits offered by the predictive modelling component of the application specifically. The IOM model should be subject to dedicated, controlled evaluation research before it is deployed operationally.

6. If the IOM model is to be deployed for enduring use, predictive model outputs should be more clearly distinguished from the descriptive harm scores. A caveat should be included alongside the model outputs, with the following 'health warning': *Prediction generated by statistical model. Accuracy and confidence may vary depending on context. Validate alongside other data sources before taking further action.*

7. Any future development of the application should focus on improving the front-end user experience, incorporating best practice in data visualisation and software accessibility. The dashboard should incorporate a 'Feedback' section, where users can provide feedback on the application and submit suggestions for improvement. Monthly feedback meetings should be held for officers to provide verbal feedback to the development team. Where possible, behavioural scientists should be consulted to advise on the most effective visual presentation of outputs to support the decision-making process and minimise risk of cognitive bias.

8. The most pertinent data points from other information management systems (most notably custody images) should be included within the RFSDi dashboard. Integrating the dashboard within the existing Connect system is likely to achieve this and should be a priority for any future development of the application.

9. Additional training should be delivered to all officers with access to the RFSDi dashboard and IOM model. This should cover how the application is intended to be used, the input variables used to calculate the RFSDi score and to build the predictive model, and an overview of the inherent limitations of the statistical techniques underpinning the system.

10. Written guidance should be developed for all officers with access to the RFSDi dashboard and IOM model. This guidance should include a summary of how the application generates the harm scores and statistical predictions, as well as a workflow diagram of how the algorithmically-generated insights should be integrated into existing offender management processes.

1. Introduction

This report presents the findings of an independent process evaluation of the West Midlands Police (WMP) RFSDi harm score and Integrated Offender Management (IOM) model. This process evaluation focused specifically on the beta-testing phase of the project, which was being conducted within two WMP Local Offender Management Units (LOMUs) at the time the study was undertaken. The purpose of beta-testing (or 'user testing') is to trial a piece of software or application with a group of target users to evaluate its performance in an operational environment. This research did not seek to evaluate the technical performance of the application. The reader is directed to the briefing notes provided by WMP to its independent Data Ethics Committee in April 2019, July 2019 and January 2020 for further discussion of the technical evaluation of the application.

When implementing a new intervention, it is important to conduct a pilot trial in a small number of settings to 'develop and refine the approach and test [the intervention's] feasibility' (Education Endowment Foundation, 2015). However, pilots in the criminal justice system are often 'implemented prematurely with insufficient time and resource put into first developing a sound theory of change and then testing key elements prior to a larger pilot' (Fox *et al.*, 2018, p. 40). The purpose of a pilot trial is to test whether a new intervention has potential in its operational context, through engaging directly with end users through primary qualitative research methods such as interviews and surveys. As such, this process evaluation sought to scrutinise the implementation of the beta-testing phase of the project, and assess the potential benefits and limitations of the intervention in the context in which it is to be implemented.

This evaluation report is structured as follows. The remainder of this section provides a brief overview of the RFSDi harm score and IOM model, which together form the two components of the application under evaluation. Section 2 outlines the research methodology used for the evaluation, including ethical considerations and limitations of the research. Section 3 summarises the key findings of the evaluation, including identified strengths and limitations of the application, and priorities for further improvement. Finally, Section 4 provides an analysis of the evaluation findings and includes 10 recommendations for the force regarding the future development of the application.

1.1 About the RFSDi harm score and IOM model

The application under consideration was first developed in 2019 and has two distinct components. The first component is referred to as the 'RFSDi score'.[2] This is a statistical harm score calculated for nominals in the force database who have previously been charged with an offence, corresponding to the overall level of harm associated with their current offending (ranging from 'low' to 'super high'). The definition of the differing levels of harm was derived from the Cambridge Crime Harm Index (Sherman, Neyroud and Neyroud, 2016). The second component of the application (the 'IOM model' or 'harm model') is a predictive model which produces forecasts at the individual level calculating each offender's likelihood of escalating from low-level offending to more serious offending, or of offending at a scale that cumulatively leads to more harm generated (i.e. the probability of moving from the 'low' or 'medium' harm groups into the 'high' or 'super high' harm groups).

The RFSDi harm score is a descriptive measure corresponding to an individual's *current* level of offending. By contrast, the output from the IOM model (the harm model) is a predictive forecast corresponding to their expected risk of *future* offending. The model predictions are calculated using Xgboost (extreme gradient boosting), a type of supervised machine learning typically used in

[2] Recency, Frequency, Severity, Drugs and Intelligence

regression and classification tasks (Chen *et al.*, 2015). The harm score and the outputs of the predictive modelling are both available to offender managers on-demand via an interactive dashboard. Throughout this report, the RFSDi score is referred to interchangeably as the 'harm score' and the IOM model outputs are referred to as 'predictions'. Both components together constitute the 'application' or 'dashboard'.

The police data scientist responsible for developing the project summarised the intended purpose of the application as follows:

> 'The basic idea is that we will provide that information in a dashboard to offender managers. There are two ways in which they will potentially use it. First of all to look at the harm score and identify whether they are managing who they should be managing. And then in terms of the harm model, it provides them with a filtered list of people that they can then go away and look at using their own normal processes.' (Police data scientist)

The overall objective is to enable WMP officers to more effectively manage risk and target preventative interventions towards the highest-risk offenders. As explained by one Inspector responsible for overseeing the project:

> 'Risk is a massive amount of what we do… we're coming away from enforcement more into rehabilitation areas through use of partners and pathways… we're trying to understand by way of engagement with individuals those crime-causing catalysts, and divert individuals away from further offending by addressing those needs... [we're moving] away from enforcement towards more of a prevention approach… morphing into the use of pathways, partners and policing skills to address and understand why people commit crime.' (Inspector)

It is important to note that neither the RFSDi score nor the predictive model are intended alone to constitute a full risk assessment; the purpose is to enable officers to more efficiently prioritise higher-harm nominals who require more detailed individual risk assessment. Any decisions related to further management or supervision will be taken by the officer, who will be expected to conduct a subsequent (manual) risk assessment in addition to the initial machine-generated forecast. The numerical scores alone do not provide insight into the underlying factors related to an individual's offending, or supportive interventions that could address their individual criminogenic needs – hence the need for a more detailed, individualised risk assessment to develop a bespoke risk management plan for each nominal under supervision.

The RFSDi score and accompanying IOM model were rolled out for beta-testing in October 2021, for a period of approximately 8 months until May 2022. Offender managers within the two pilot LOMUs were required to use the dashboard in conjunction with existing data systems to support offender risk assessment processes. All participants took part in a training session to brief them on the purpose and function of the application before being required to use it operationally (the author also attended one of these training sessions). A total of 9 training sessions were delivered throughout 2021, 6 of which were in person and 3 of which were delivered online.

Both components of the application were first reviewed by the WMP Data Ethics Committee in April 2019. The minutes from this meeting contain a detailed explanation of how the RFSDi score and model predictions are calculated. Various questions were raised at this meeting, which were addressed in writing by the force in July 2019. This written response noted that:

> ‘The model seeks to address a capacity issue and allows the assessment of more data (than previously utilised) related to individuals in order to identify those people who require enhanced support and offender management. This cannot be done manually and as a result opportunities are currently missed to help offenders by providing preventative interventions and also protect future victims from harm.’ The response went on to note that the overall anticipated benefit of using the model is that ‘there will be fewer occurrences of harmful or vulnerable offenders being missed’ (West Midlands Police, 2019)

The reader is directed to the minutes from these ethics committee meetings for further detail regarding the technical development and pre-deployment testing of the application.

2. Methodology

This section summarises the research methodology used for the process evaluation. The methodology is based on the guidance described in the College of Policing's Policing Evaluation Toolkit (College of Policing, 2018). The toolkit provides a framework to assist the police in evaluating the impact of tactics, projects or policies in their local area. Stage 2.3 of the toolkit (pp. 26–28) outlines the key factors to consider when conducting a process evaluation of a policing intervention. It encourages the use of interviews and survey methods to understand participants' perceptions of an intervention, identify whether the intervention was delivered as intended, what worked well and what could be improved. The methodology used here is based directly on the principles described in Stage 2.3 of the toolkit.

2.1 Data collection

Three data collection methods were used to conduct the process evaluation: semi-structured interviews; focus groups; and a practitioner survey. The purpose was to examine the real-world use of the RFSDi harm score and accompanying IOM model in their operational context.

Interviews and focus groups were conducted throughout January 2022 with 15 respondents within the force who were taking part in the beta-testing phase of the project. This included officers of varying ranks between Constable and Inspector, and several police staff such as police data scientists. Interviews are the most widely used method of data collection in qualitative social research (Cassell, 2005; Nunkoosing, 2005). Interviews were appropriate in this context given their inherent flexibility (Bryman, 2016), and their value not only in obtaining insight into social issues by exploring individuals' experience, but also in capturing important contextual information (Denzin, 2001).

A semi-structured approach was adopted to ensure data collection remained focused on the target research questions, while allowing sufficient flexibility to explore other areas of interest not initially anticipated in the research design. Semi-structured interviews are conversational in tone, and allow for an open response rather than a binary "yes" or "no" answer (Longhurst, 2003). A semi-structured interview guide is an outline of key questions to guide the discussion, but does not constitute an exhaustive list of all questions the researcher may ask in interview (Newcomer, Hatry and Wholey, 2015). Semi-structured interviews are well suited to studies such as this, where the focus is on exploring under-researched territory and allowing interviewees maximum latitude to pursue new angles of enquiry (Newcomer, Hatry and Wholey, 2015).

Interviews and focus groups were conducted on an anonymised basis. Interview request letters were sent in advance (by email), alongside a project information sheet so respondents had a clear understanding of the purpose of the project and were able to give their informed consent to the interview. Once participants had reviewed the information sheet and had an opportunity to ask any questions regarding the project, they were requested to return a signed consent form, which was stored securely and separate to any interview data. All interviews and focus groups were conducted remotely using a secure videoconferencing platform (Microsoft Teams). The sessions were not video or audio recorded. Instead, notes were transcribed directly into a secure document at the time of interview. Bearing in mind the sensitive nature of the subject matter and the background of respondents, it is unlikely that participants would consent to a recording being made of their interview or focus group, and any who did consent would likely be inhibited from exploring sensitive or potentially contentious issues.

Finally, a closed-ended statistical survey was distributed to all officers involved in the beta-testing exercise, eliciting a total of 11 responses. The survey explored the extent to which users find the harm score and predictive modelling useful, whether they use it regularly as part of their offender management responsibilities, and whether it has delivered operational benefit in their force area. Due to the anonymised data collection method used, it is not possible to assess the degree of intersection between the interview / focus group sample and written survey responses. It is possible that some users who took part in a research interview or focus group did not complete a written questionnaire, and vice-versa. Nevertheless, consistent themes and findings emerged across all research activities, which are discussed further in the following sections.

2.2 Data analysis

Ensuring sound qualitative research requires developing a systematic and rigorous approach to the collection and analysis of data, and the interpretation and reporting of findings (Fossey *et al.*, 2002). Interview data was analysed following a general inductive approach, whereby the aim is to derive theory from the data, rather than test pre-defined hypotheses (Bryman, 2016). The inductive approach involves adopting a systematic procedure for analysing qualitative data, guided by specific research objectives (Thomas, 2003). The inductive approach is appropriate in this context as it enables transparent and defensible links to be established between the research questions and resultant findings (Gioia, Corley and Hamilton, 2013).

A preliminary open coding process allowed recurring themes and categories to be identified, and a more granular analysis allowed trends and patterns within these themes to be explored in detail (Corbin and Strauss, 2014). Following close reading of interview transcripts, an “in vivo” coding process allowed distinct categories to be derived from units of phrases (also referred to as “verbatim coding” or “literal coding”) (Saldaña, 2014). Pertinent text segments were then extracted into their relevant category, allowing interview data to be analysed thematically.

Once interview data had been organised according to these broad categories, a more granular analysis allowed subtopics to be explored within each category, providing detailed insights into the range of views and perspectives put forward by participants. This next stage of analysis could be described as a form of “axial coding” (Corbin and Strauss, 1990), where data is scrutinised to identify relationships between categories and sub-categories. As noted by Brown et al., there are four analytical processes involved in axial coding: continually relating categories to subcategories; comparing those categories with collected data; expanding the density of categories by exploring their properties and dimensions; and exploring variations in the phenomena (Brown *et al.*, 2002). This axial approach is intended to be iterative, whereby the analysis and coding of data informs the questions to be posed in later interviews.

This multi-stage process of open and axial coding bears close resemblance to the systematic set of procedures involved in Grounded Theory (Corbin and Strauss, 1990, 2014; Charmaz and Belgrave, 2007; Glaser and Strauss, 2017). However, this study did not adopt a Grounded Theory methodology in its true sense. The overall purpose of Grounded Theory is to *explain* a substantive topic at a broad conceptual level (Creswell, 2002). The primary focus of the current study is rather to critically assess the strengths, limitations, opportunities and risks associated with the project under investigation. Nevertheless, adopting elements of the Grounded Theory approach ensures a degree of systematic rigour to inductively derive key observations and findings. A main advantage of this inductive approach is that it enabled detailed analysis of the content of interviews, including identifying divergences of views and contradictions between participants.

2.3 Ethical Considerations

When conducting research involving human participants, tensions may arise between the investigative aims of the project and participants' right to privacy (Fujii, 2012). Potential harms can be prevented or reduced through the application of appropriate ethical principles (Orb, Eisenhauer and Wynaden, 2001). Numerous examples of research ethics frameworks have been developed over the years, which aim to provide a set of guiding principles for researchers to follow (Pimple, 2002; Smith, 2003). The most relevant frameworks for the current study are the ESRC framework for research ethics[3] and the UWL Research Ethics Code of Practice 2018.[4] Both are guided by the over-arching principles of beneficence and non-maleficence: research should maximise beneficial outcomes while minimising potential risk and harms.

All stages of the research were guided by the two ethics frameworks mentioned above. Informed consent was received from all participants as a pre-requisite to their participation. A project information sheet was provided to participants outlining the parameters of the project, the aims of the research, and contact details of a third-party they could contact for further information. Participants were able to follow up with any questions regarding the study before they agreed to participate. Once participants had the opportunity to review the information sheet and ask any questions, they were asked to provide a signed consent form, acknowledging that they understood the purpose of the project and what their data would be used for, and that they freely consented to taking part in an interview. These consent forms were stored securely in a separate location to any interview data.

Research for this study was conducted on an anonymised (non-attributable) basis, and all collected data was stripped of any personal information relating to participants. Prior to engaging in any research activities, participants were requested not to discuss any classified or otherwise sensitive information that should not be disclosed in the public domain. Another important ethical consideration in this context is the need for transparency of data collection, analysis and reporting (Fossey *et al.*, 2002). To ensure transparency, participants were offered the opportunity to request a copy of their interview transcript. One participant made such a request and their transcript was provided to them one day after the interview was conducted.

A final ethical issue taken into account when designing the study was the potential for the methods used to cause unintended distress, discomfort or otherwise adverse outcomes for participants. The risk of psychological harm or distress to participants arising from the interview process was assessed to be negligible. Interview questions were non-personal: at no stage were participants asked for information that did not relate directly to their professional role. In addition, research activities were conducted with careful consideration of issues of diversity and inclusion in the selection of stakeholders and case studies.

[3] https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/our-core-principles/

[4] https://www.uwl.ac.uk/sites/default/files/Departments/Research/Web/PDF/research_ethics_code_of_practice_january_2019f.pdf

2.4 Limitations

This study has several limitations. The primary limitation relates to participant recruitment and sampling. Although efforts were made to ensure that all offender managers involved in the beta-testing phase of the project were engaged in the process evaluation, staffing changes within the force meant that some research participants had been using the application for longer than others, potentially influencing their perspectives on its strengths and limitations. Another challenge to internal validity is the risk of self-censorship: although all research activities were conducted on an anonymised basis, it is possible that some participants felt reluctant to express views that could reflect negatively on the force.

External validity of the process evaluation is also limited. As detailed previously, the evaluation focused specifically on the beta-testing phase of the project. The findings may not be reflective of other data-driven offender management projects currently in development, either within the same force or nationwide. Moreover, as noted by the College of Policing, process evaluation is not a substitute for good impact evaluation (College of Policing, 2018, p. 28). As such, further evaluation research is needed to assess the overall *impact* of the project within the force in question, and whether it has resulted in improvements in the force's overall approach to offender risk management. It was beyond the scope of the current process evaluation to assess the overall impact of the project on the force's wider offender management processes.

Finally, internal (descriptive) validity may also be affected by the fact that all research activities were conducted remotely, and the method of live transcription used to record interview notes. It is possible that research participants conveyed certain non-verbal cues that were not apparent over a video call, and the interview notes did not record every word uttered by participants. The interview notes are therefore likely to be less descriptively rich than a verbatim interview transcript derived from audio recordings.

3. Key Findings

This section summarises the key findings of the process evaluation. The following section (Section 4) discusses the implications of these findings and recommendations for future practice.

Before discussing the specific strengths and limitations of the application identified in the process evaluation, it is important to reflect on three cross-cutting findings that emerged from the research.

3.1 Cross-cutting findings

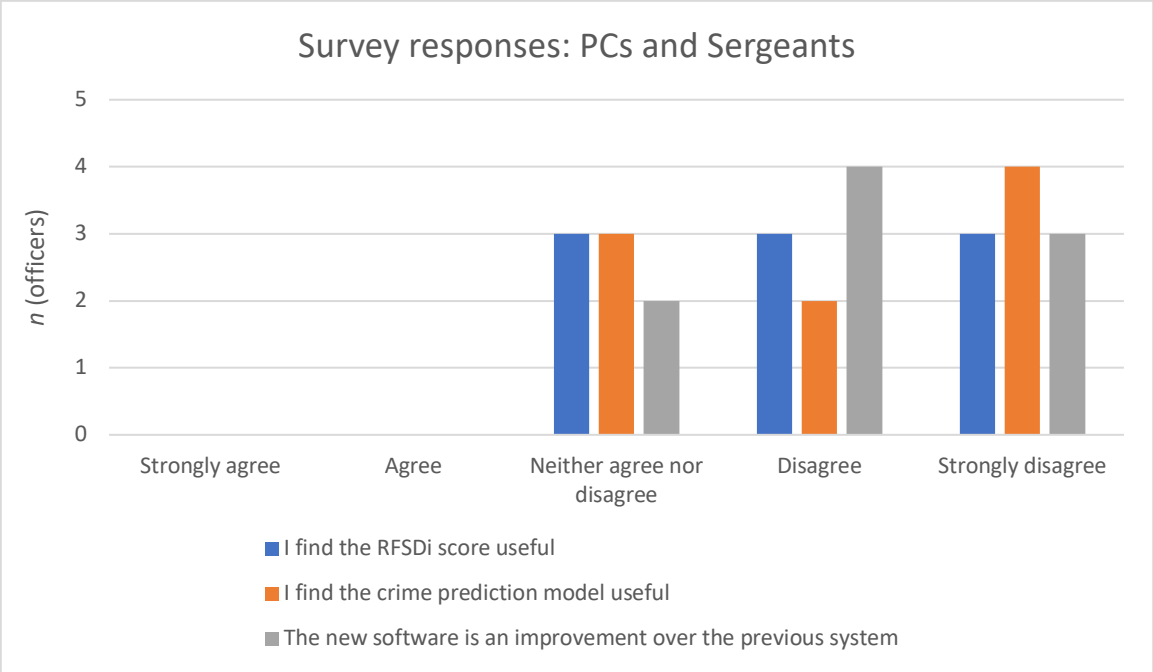
First, the research highlighted a fundamental divergence in views between junior officers (PCs and Sergeants) on the one hand, and more senior officers (Inspectors) on the other. The latter group were significantly more positive regarding the application, as indicated by comments such as:

> 'I'm infinitely more happy with RFSDi than Corvus. I was not convinced with the maths that sat behind the scoring. I'm much happier with the transparency around RFSDi, the filters that it's brought in. It's more interactive, that gives me a lot more confidence that it's a more precise tool.' (Inspector)

This is in stark contrast to the comments returned by PCs and Sergeants:

> 'I have no idea how the RFSDi and crime predictor tool come up with the data. The training is an hour over Skype but I left as confused as when I started. The system only works on charges but this does not show intelligence so will only pick up people charged, not arrested or those with lots of intelligence to suggest offending. I don't use it.' (Sergeant)

These findings were also reflected in the survey data. In response to the prompt 'I find the RFSDi score useful', no PC or Sergeant respondents 'agreed' or 'strongly agreed' with this statement. 33% (*n* = 9) neither agreed nor disagreed; 33% (*n* = 9) disagreed; while 33% (*n* = 9) strongly disagreed. In relation to the predictive IOM Model, again no PC or Sergeant respondents reported that they find the modelling useful: 33% (*n* = 9) neither agreed nor disagreed; 22% (*n* = 9) disagreed; while 44% (*n* = 9) strongly disagreed. When asked whether RFSDi is an overall improvement over Corvus, no PC or Sergeant respondents 'agreed' or 'strongly agreed' with this statement: 22% (*n* = 9) neither agreed nor disagreed; 44% (*n* = 9) disagreed; while 33% (*n* =9) strongly disagreed.



By contrast, both Inspectors who responded to the questionnaire 'agreed' that they find the RFSDi score and the crime prediction model useful, and both 'agreed' that the application represents an improvement over Corvus. This further emphasises the divergence in perspectives between PCs and Sergeants on the one hand, and Inspectors on the other. This is an important finding of the research, and its implications are discussed further in the following section.

The second key finding of the research is that the practical implementation of the application does not appear to have resulted in a better user experience for officers. In theory, officers believed that the new system should enable them to 'be a bit more strategic to recommend people… to management in neighbourhoods' (PC). However, in practice, PCs and Sergeants reported that the user experience of the new application was not an improvement over Corvus, as indicated by comments such as:

> 'Corvus was an easier system to navigate. RFSDi throws up a lot of names, but the interface isn't as user-friendly as Corvus. I think Corvus was better as people are able to manage better with visual aids.' (Sergeant)

> 'I've had a play with it over the last few days. I've put a nominal's name in and it just takes forever to load. 3, 4, 5 minutes just to look at one person.' (PC)

This poor user experience appears to be the main reason that respondents reported not regularly using the new system as part of their offender management responsibilities, nor finding the RFSDi score or predictive modelling useful, as reflected in survey data.

Based on these observations, it appears that the practical implementation of the application is yet to result in the improvements for offender managers that were envisaged at the time it was rolled out for beta-testing. It remains unclear whether this is due to limitations in the underlying statistical application itself; deficiencies in the user experience of the dashboard itself; or a lack of sufficient training and guidance for users (or a combination of these factors). These issues are discussed further in the following section.

A third and final cross-cutting finding relates to the distinction between the two components of the application: the descriptive RFSDi score on the one hand; and the predictive IOM model on the other. These represent two distinct components of the dashboard: the harm score is a descriptive measure relating to individuals' *current* offending level; while the IOM model produces a probabilistic forecast indicating an individual's likelihood to progress to higher-harm offending.[5] However, this distinction was not clearly recognised by the majority of respondents. Most interview quotes and survey data relate specifically to the descriptive RFSDi score. Although prompted several times to provide an opinion on the probability modelling component specifically, interviewees had little insight into this aspect of the application. As such, all findings discussed in this section should be interpreted as relating primarily to the descriptive harm score element, unless specific reference is made to the probability modelling element of the dashboard.

3.2 Strengths

As mentioned previously, there was a clear divergence of perspectives between Inspectors on the one hand, and PCs and Sergeants on the other, which must be taken into account when interpreting the findings presented in this section. At times the feedback provided by PCs and Sergeants directly contradicted the comments provided by Inspectors. Potential causes and implications of this are discussed further in the following section. For this reason, the perceived strengths of the application reported in this section derive largely from the comments returned by the Inspectors interviewed for the project, rather than PCs or Sergeants.

Inspectors responsible for overseeing the beta-testing of the application recognised that existing approaches to offender risk management were 'inconsistent', with one interviewee explaining that 'everyone's done it differently. Inconsistency with regard to risk is the big one for me.' (Inspector). It was suggested that the new application could 'give us a good push in the right direction to be more precise with our risk assessment.' This interviewee described the new application as 'more smart really', adding that 'I have a better understanding at where the scores are coming from.' (Inspector). In relation specifically to the predictive IOM Model, the Inspector reported that:

> 'I don't know much about this one. My understanding is that it's going to predict harm and predict who may re-offend… that could allow us to intervene sooner with someone on an upward trajectory. If that was really nice and clear, that would allow us to intervene really quickly by way of a more intensive offender management visit. At the moment, we're reliant on local knowledge but there's been no formal system to feed that into.' (Inspector)

Based on interviews and questionnaire responses from the two Inspectors specifically, the research highlighted three main strengths of the new application: precision; transparency; and identification of 'hidden risk'. These are discussed in turn below.

Precision

The first perceived benefit of the RFSDi harm score is improved precision of targeting. It is important to note that precision is distinct to accuracy: accuracy refers to the overall validity of the statistical predictions (i.e. the proportion of individuals that are assigned the 'correct' risk score), while precision refers to the granularity of these predictions. Precision is directly linked to efficiency; in a resource-constrained environment it is essential to efficiently prioritise limited resources to those areas of greatest need or risk.

[5] For example, an individual's *current* harm score may be 'LOW', while their risk of escalating from low harm to higher-harm offending may be a high probability, such as '90%'.

As explained by one interviewee, 'with reduced officer capacity, there is an absolute requirement for us to act with precision. RFSDi helps us to do that beautifully.' (Inspector). As a result, 'many individuals have been de-selected on the basis of RFSDi', meaning the tool is enabling the force to more effectively screen out lower-risk offenders who no longer require supervision. An important feature in this regard was said to be the ability to monitor changes in an individual's harm score across specific offending behaviour, to track 'an increase or decrease in that risk, with numerous sub-filters.' (Inspector).

This is consistent with the anticipated benefits described by the senior data scientist responsible for overseeing the project, who explained that:

> 'RFSDi would provide that filtering capability that they might not have had before… if it means that offender managers can start looking at the people who are committing the higher types of harm in a consistent manner, then that risk can be mitigated and managed within the force, allowing them to do that more efficiently.' (Police data scientist)

It remains unclear whether these intended benefits have been realised in practice at the operational level, as discussed further below.

Confidence

The second perceived benefit of the system is increased confidence in the risk assessment process. One interviewee suggested that the application has 'brought an element of transparency to the IOM process, so you can see where the scores are coming from.' (Inspector). This transparency has helped to build confidence in the validity of the outputs, as the factors that contributed to a particular harm score can be identified and triangulated across other data sources:

> 'I would say it is very accurate. When we've gone through the RFSDi scores we've recognised names of individuals, and we cross-reference to our intel system, and you can see where the score has been generated from… I'm very confident that it's identifying the right people.' (Inspector)

This degree of transparency was viewed as essential to maintain confidence in the output: 'if it was a black box, I think you'd naturally feel less confident about making that decision. I'd feel significantly less confident if I didn't know what sat behind that.' (Inspector). As the RFSDi score is purely descriptive, it is possible to identify the specific factors that led to each individual harm score (in contrast to the predictive IOM model).

Identification of 'hidden risk'

A main advantage identified in the research is the ability to identify high-risk nominals who may otherwise not have reached the threshold for offender management or further scrutiny. While not an original intended purpose of the system, Inspectors report using the harm score to verify and triangulate the risk assessments conducted by other agencies (most notably the probation service) using traditional risk assessment frameworks such as OASys and OGRS:

> 'We use it now to take to Day One selection meetings with Probation. To confirm that Probation are also selecting the right people with their OGRS scores.' (Inspector)

However, a more significant development is the use of the harm score to identify individuals who may not have previously been subject to offender management orders:

> 'We've been using RFSDi to identify [an Under-25 cohort] for either offender management or neighbourhood management, depending on the level of risk… the vast majority were not being managed or were being managed as part of a wider grouping… they were not given any proactive management plans… Previously, there would be no system that was flagging them to us. The neighbourhood officers would probably know them, but they would have no way of carrying that forward, managing that. The system to flag them.' (Inspector)

Assuming this is an accurate observation, this could be interpreted as both a strength and limitation. While the surfacing of 'hidden risk' could lead to pertinent individuals being identified who may otherwise have gone unnoticed, this also increases the scope for over-classification of risk, which could place increased demand on already-stretched offender managers to scrutinise a larger number of cases.

3.3 Limitations

User experience

As mentioned previously, a main limitation identified in the research relates to the user experience of the interface itself. As reported by interviewees:

> 'It's certainly not user-friendly if I'm honest. I'm clicking around, pressing buttons, I don't really know how to navigate around it if I'm honest… I don't think the tool itself is user friendly.' (PC)

A common issue mentioned was that the dashboard is not integrated with other policing systems, meaning officers must access multiple systems to triangulate the information provided by the harm score:

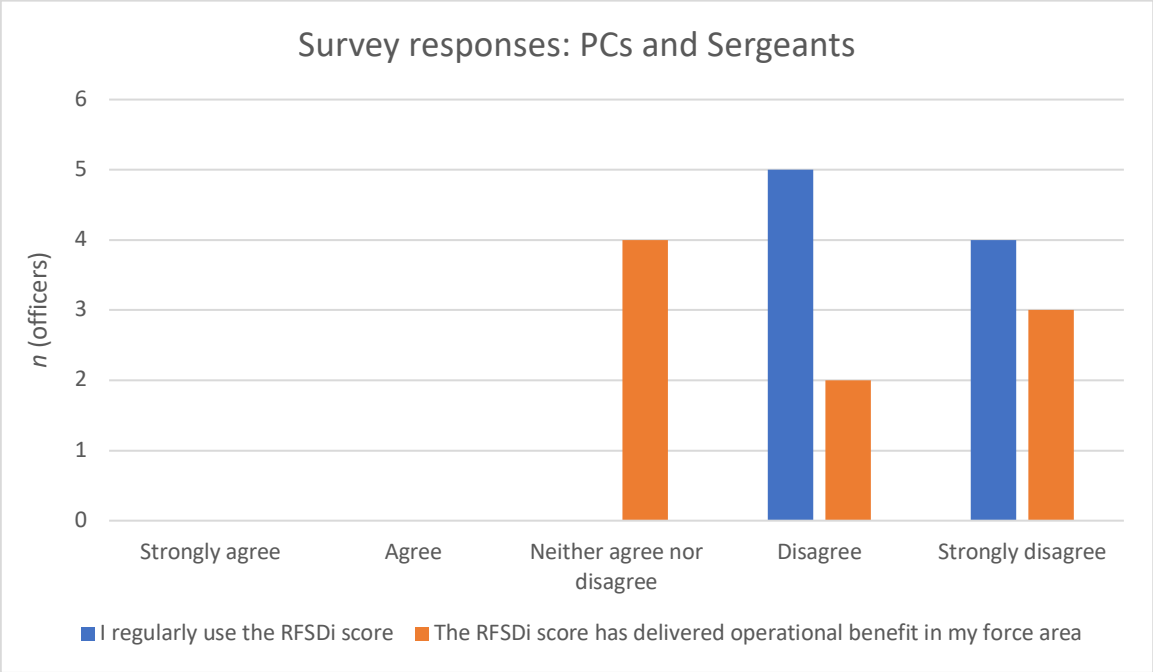
> 'There was no way in the app of knowing why people are scored in a certain way. You have to then go into a different system to look at one individual.' (PC)

> 'It needs to be a bit more user-friendly and available on the same system as everything else… it needs to be very simple, very clear. You really have to spend time on it, and time is of the essence, especially in policing… Customers wouldn't use it, but since it's a police system we have to use it and we have to make it work.' (Sergeant)

This was recognised by the senior data scientist responsible for developing the project, who explained that:

> 'They [offender managers] would still need to go off and do their own risk assessment, because at the end of the day the offender managers will have access to information that we don't… They would go away and apply their own processes and if they think that person should be managed then they will be managed in the normal fashion.' (Police data scientist)

Nevertheless, this poor user experience appears to be one of the main reasons why no PCs or Sergeants who participated in the questionnaire reported using the RFSDi score regularly as part of their offender management responsibilities, and none reported that the application has delivered operational benefit in their force area.



Over-classification of risk

A second limitation identified in the research relates to the over-classification of risk: individuals who receive high harm scores but on further examination are assessed not to pose an immediate risk. As described by one interviewee, 'Sometimes it's incorrect, there are people in prison who have high RFSDi scores.' (PC). As a result, officers report that the system is scoring a disproportionately large number of nominals as 'high' or 'super high', resulting in an unmanageably large list of individuals to review:

> 'I've seen some inconsistencies in it. It wasn't showing any of our current youths that we would look at. I've had another look recently and a lot of ours are starting to feature in our "high" cohort, and there's now one who's scoring as "super high", but I have some doubt as to how that scoring is calculated… having looked at him I don't think he needs to be on that radar.' (Sergeant)

> 'On the harm side of it, I've had a look at it and there were 900 offenders who scored high across my area. I'm not going to go through all of them, it's too time-consuming. And the breakdown of crimes is not detailed, it just says "acquisitive", it doesn't say burglary or robbery or whatever. So it's really difficult for me to work out who I'm meant to look at.' (PC)

There is a risk that the large number of high harm scores generated by the system create additional demand for already-stretched offender manager teams. This risk was well recognised by the Inspector responsible for overseeing the delivery of the project:

> 'If we're going out looking for people to manage, that does create more demand. But we've made an agreement with the neighbourhoods that we're not going to overload them and we're only going to choose the highest risk individuals. The system will help us identify those who are causing the most risk.' (Inspector)

Nevertheless, they did recognise that 'it does generate a lot of people, and you have to go in and double-check that you've got the right offending type.' (Inspector). This is an important limitation to consider when assessing whether the system has resulted in overall efficiency gains for offender management units.

Missing nominals

Conversely, as well as over-classification errors, officers also report that a large proportion of individuals who should be scored as high-risk are not currently being identified by the system. There appear to be two main reasons for this. First, the risk scoring only includes nominals who are classed as 'defendants', i.e. those who have been previously charged with an offence. It does not include 'suspects' (those who have been arrested but not yet charged). Second, the system only uses local data, meaning that data from other police forces or national databases will not be incorporated into the risk scoring:

> 'Individuals who are committing offences are not necessarily on the system if they are "suspects" and not "defendants"… If someone moves in from our area and they are a prolific burglar, they won't feature whatsoever because nothing is transferred.' (Sergeant)

As summarised by one Sergeant who responded to the written questionnaire:

> 'I have concerns over how up to date the info is that is then being used to give the scores. Having checked when the trial initially started, no offenders we manage were featuring in the high brackets… I'm managing youths of significant risk, which [the dashboard] isn't picking up. It is therefore not something I would use as part of my daily business nor on a regular basis. It would be something I would check probably on a monthly basis to see if there are any names on there which we are not aware of.' (Sergeant)

It is important to note that the discrepancy between the officer's assessment and the machine-generated harm score is based on the subjective judgement of the officer; it was not clear what factors led to the officer concluding that the youths in question are 'of significant risk', or why these factors had not been identified by the dashboard.

This perceived limitation appears to be one of the main reasons why no PCs or Sergeants report having confidence in the accuracy of the RFSDi score, or assess that it has delivered operational benefit in their force area (0%, n = 9). In relation specifically to the IOM Model (the machine learning component of the application), it was suggested that:

> 'The suspect thing would come in handy there. If they're a suspect for offending but not a defendant, it can show us the frequency even if they're not charged or convicted… Where the system could come in useful is if people are getting arrested, there's lots of intel coming in about them… The people in the "Pursue" cohort, but where there are no conditions on them but they are on our radar.' (Sergeant)

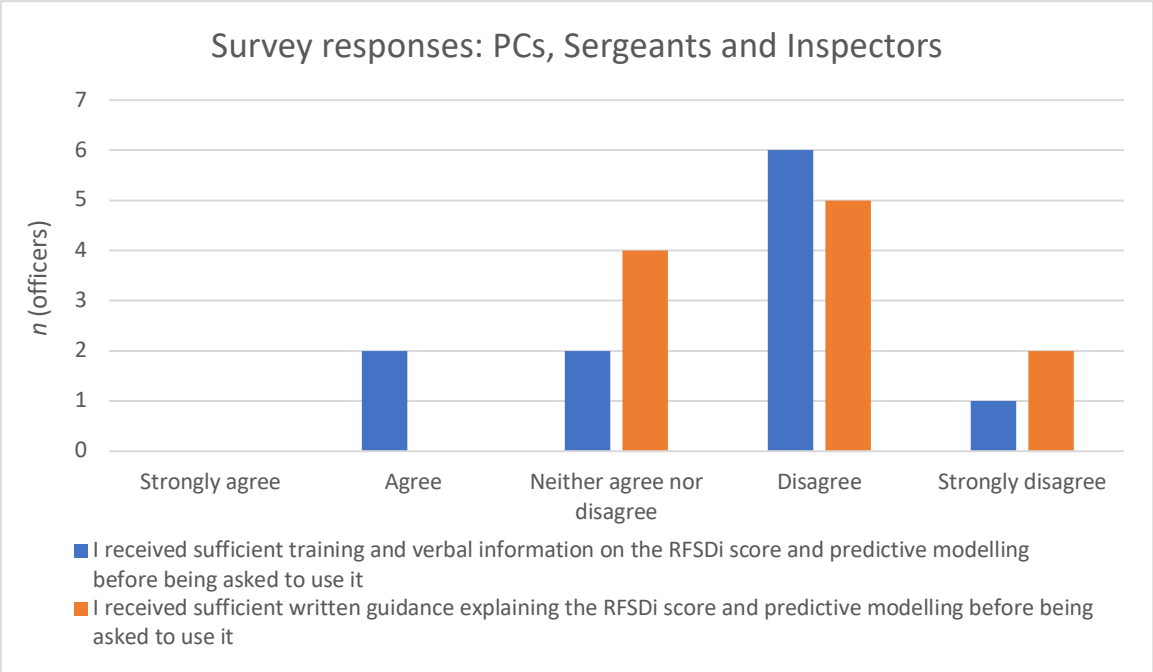
It was reported that the decision to exclude suspects from the dashboard was made on the basis of advice provided by the WMP Data Ethics Committee regarding the ethical considerations associated with 'risk scoring' individuals who have not yet been charged with an offence. However, it was pointed out that the force will continue to actively monitor, risk assess and manage others who have not been charged with an offence, but without the use of the app. It is therefore unclear whether the choice to exclude suspects from the dashboard does in fact represent a more 'ethical' approach, particularly if this is damaging users' trust and confidence in the overall system.

3.4 Priorities for further improvement

Beyond the specific limitations discussed in the previous section, respondents also identified several areas of focus for future improvement, if the application were to be deployed for enduring use.

Training

The first and most significant priority for the future use of the application relates to the training provided to officers. As indicated by survey responses (reported below), the majority of respondents reported that they have not received sufficient training or written guidance on the RFSDi scoring system or IOM Model before being required to use the application.



Inspectors and Sergeants alike recognised the importance of further training, as indicated by comments such as:

> 'The training is an hour over Skype but I left as confused as when I started.' (Sergeant)

> 'I think there should have been some written training guidance that we could refer to when using the system… The training given could have been more detailed.' (Inspector)

> 'Some refreshed and specific training inputs would be beneficial prior to (or at the point of) further roll out of RFSDi to ensure that all users are aware of the system and its benefits – maybe some written instructions that can be retained.' (Inspector)

The perceived lack of sufficient training and guidance materials may have partially contributed to the divergence of views between Inspectors on the one hand, and PCs and Sergeants on the other. This is discussed further in the following section.

Filters

One important missing feature that interviewees wanted to see added to a future version of the dashboard is the ability to filter by more specific crime types. PCs explained that 'because the selector criteria are not filtered, it's difficult to narrow down according to crime type.' (PC). To address this, they requested: 'can we have a top 20 or top 30 of offenders in a particular crime type, like burglary, knife crime etc.' (PC). This requirement was also recognised by Inspectors interviewed, who suggested it would be particularly helpful to filter for domestic-related incidents specifically:

> 'I'd like to see more specific filters for violence, domestic abuse… I'd like to see a "domestic" tab in there as well, that would be a really useful filter to have in so we can divide out anything with a domestic element… And just to retain that ability to have filters built into it depending on what we'd like to look at… Perhaps something just to filter through anything that's domestic.' (Inspector)

Lack of this functionality may be one of the factors leading most respondents to report not regularly using the harm score as part of their offender management responsibilities.

Data integration

The final missing feature that several respondents requested was the ability to view other intelligence related to an individual within the dashboard, most notably pictures of offenders:

> 'When we look at it there are no pictures of offenders, it's just names. On Connect you would see pictures and other Intel, RFSDi is just words… it's too data-driven.' (PC)

This primarily appears to be an issue of data integration and cross-compatibility with other information management systems. As explained by one PC:

> 'It wouldn't be the sole basis of a decision. You'd go through other things, you'd speak to your neighbourhoods… RFSDi is more of a selection process tool. It might be better if RFSDi was part of Connect in the future.' (PC)

Indeed, this was also recognised by the project's lead data scientist as one of the main limitations of the current application:

> 'The main limitations for the model are the data available to it. We don't have data from other forces, or information from other sources.' (Police data scientist)

However, it must be borne in mind that this is not unique to the tool being assessed but is a recurring issue across all police information management systems, as discussed further in the following section.

4. Analysis and Recommendations

4.1 Testing and evaluation

For reasons discussed below, this process evaluation was not able to establish sufficient evidence to support wider operational deployment of the RFSDi dashboard and accompanying IOM model. Further testing and evaluation are required to assess the potential benefits of the system if it is to be deployed for enduring use.

First, due to a combination of staffing changes, absence and non-participation, not all officers involved in the Beta-testing of the application took part in this process evaluation. As such, there may be other important factors relevant to the development and use of the system that were not captured by this research. As detailed below, the evaluation findings demonstrate only limited positive evidence in favour of wider deployment of the application, and the little evidence that was established was provided primarily by Inspectors who took part in the project, rather than PCs or Sergeants.

Second, most officers engaged for the research did not have a clear understanding of the distinction between the RFSDi harm score on the one hand, and the predictive model on the other. For this reason, it has not been possible to meaningfully assess the benefits and limitations of the predictive model specifically, and further evaluation research is needed to assess this component of the system specifically, distinct to the RFSDi harm score.

> **Recommendation 1:** This process evaluation has not established sufficient positive evidence in favour of deploying the RFSDi dashboard and accompanying IOM model for long-term operational use. The application should be subject to further, detailed evaluation research to conclusively establish its benefits and limitations before it is deployed for wider operational use.

Importantly, there are no pre-defined evaluation metrics in place by which the overall impact of the project will be assessed. Without such evaluation criteria, it remains unclear what the intended outcomes for the project are and how these will be measured. As part of the longitudinal evaluation plan specified above, it is essential to develop clear evaluation criteria against which the project will be assessed, alongside a theory of change or logic model articulating the overall intended outcomes of the project.

> **Recommendation 2:** The force should establish a clear impact evaluation plan to measure the outcomes of the RFSDi dashboard and IOM model on an ongoing basis. This should include developing a basic logic model or Theory of Change to describe the intended outputs and outcomes of the project, as outlined in the College of Policing's Evaluation Toolkit. It is important to define measurable evaluation criteria to assess the ongoing business case for the project and demonstrate that it is delivering its intended outcomes. The Maryland Scientific Methods Scale provides useful methodological guidance for ensuring the validity of policing evaluation research.

The impact evaluation plan described above should be time-bound, and the project should not proceed unless it can be demonstrated that it is delivering its intended outcomes at the end of a specified evaluation period. These results should be independently reviewed by an impact evaluation specialist and presented to the West Midlands Police Data Ethics Committee following completion of the evaluation period.

4.2 Does the RFSDi harm score help officers to more effectively manage risk?

Perhaps the most significant finding of the evaluation is a fundamental divergence in views between PCs and Sergeants on the one hand, and Inspectors on the other. Both Inspectors who responded to the questionnaire reported that they find both the RFSDi score and the predictive model useful, and that the application represents an improvement over Corvus. In stark contrast, no PC or Sergeant reported finding the RFSDi score or the crime prediction model useful, and none reported that it represents an improvement over Corvus. Based on these findings, it appears that the RFSDi score and predictive model are significantly more useful for Inspectors than they are for PCs and Sergeants. There are two likely reasons for this.

First, the Inspectors who participated in the evaluation have been more directly involved in the planning and development of the system and are therefore likely to have a more detailed understanding of its strengths and limitations. They are likely to be more familiar with the rationale for its deployment and have had longer to familiarise themselves with the application. This demonstrates the critical importance of ensuring wide consultation and engagement in the early planning stages of a new data-driven system, to ensure end-users have had sufficient opportunity to contribute to the early development process.

> **Recommendation 3:** Offender Managers who are required to pilot or trial any new data-driven risk assessment tool should be consulted at an early stage in the project development, giving them an opportunity to directly contribute to the application development process. An initial survey of end-users should be distributed, requesting feedback on the limitations of existing processes, and the user interface and design requirements for any new system.

A second potential reason for this divergence in views relates to the need for more senior officers to understand strategic-level insights across their force area. At the individual offender level, PCs and Sergeants reported numerous perceived over-classification errors, requiring users to validate risk scores manually via other systems. Conversely, they also reported numerous perceived under-classification errors and false negatives (missing nominals), suggesting that a large proportion of individuals who should be scored as high risk are not being identified by the system. The perceived occurrence of both over-classification errors and false negatives appears to have significantly damaged PCs and Sergeant's trust in the overall validity of the system, leading none of them to report that the harm score has delivered operational benefit in their force area. One Sergeant explained that the occurrence of false negatives 'discredits the info for me personally', demonstrating that the experience of even a small number of false negatives could cause users to lose trust in the validity of the system as a whole. (It is important to note, however, that these perceived errors are based on the subjective judgement of officers – there is no way to validate that the purported 'false negatives' and 'missing nominals' are in fact genuine errors of the system).

By contrast, at the more strategic level, Inspectors report being very confident that the system is identifying the right people, explaining how the RFSDi scoring has identified a new cohort of under-25 offenders, the majority of whom were not previously subject to proactive management plans. As such, despite individual-level over-classification and under-classification errors that appear to have weakened officers' confidence in the validity of individual outputs, at aggregate level, the system appears to be surfacing high-risk nominals who should be subject to further intervention, but may have otherwise gone unnoticed. This identification of nominals who may otherwise not be subject to further scrutiny appears to be the single greatest strength of the RFSDi score (as reported by Inspectors) and should be the focus of any future development of the application.

Recommendation 4: Any future development of the RFSDi harm score should focus on identifying nominals not currently subject to offender management orders who should be subject to more in-depth risk assessment. To avoid the risk of false negatives (which could lead to high-priority nominals being erroneously de-selected), individuals already subject to offender management plans should be excluded from the RFSDi harm scoring system.

4.3 Does the IOM model help officers to more effectively manage risk?

PCs and Sergeants interviewed for this research did not distinguish clearly between the descriptive RFSDi score on the one hand, and the predictive model on the other. The majority of interview and survey responses related specifically to the RFSDi harm score. When prompted to provide feedback on the predictive modelling component as distinct from the RFSDi score, most interviewees reported that the predictive model is not yet routinely used. The evaluation was therefore inconclusive regarding the potential benefits offered by the predictive modelling component of the application. Further evaluation research is required to determine whether the predictive model is useful in practice to help officers manage risk more effectively.

Recommendation 5: The research was inconclusive regarding the potential benefits offered by the predictive modelling component of the application. The IOM model should be subject to dedicated, controlled evaluation research before it is deployed operationally.

Nevertheless, one important conclusion can be drawn in this regard. From a statistical perspective, the RFSDi score and the offender escalation predictions are fundamentally different categories of output. The RFSDi score is a descriptive score corresponding to an individual's *current* level of offending. By contrast, the predictions produced by the IOM model are based on machine learning forecasting of *future* risk, and therefore entail a degree of inherent uncertainty. However, this important distinction has not been clearly articulated to users, reducing their ability to critically assess the validity of model outputs.

It is essential that end users are made fully aware of this inherent uncertainty associated with machine learning predictions, and that model outputs are treated with a significantly greater degree of caution than descriptive harm scores. It is concerning that officers interviewed for this evaluation were not aware of this crucial distinction between the descriptive RFSDi score and the predictive modelling. If the application were to be deployed for enduring use, it will be essential to clearly communicate to users that the outputs of the predictive model are inherently uncertain and probabilistic, and should be treated with a higher degree of caution and scrutiny than the descriptive harm score.

Recommendation 6: If the IOM model is to be deployed for enduring use, predictive model outputs should be more clearly distinguished from the descriptive harm scores. A caveat should be included alongside the model outputs, with the following 'health warning': *Prediction generated by statistical model. Accuracy and confidence may vary depending on context. Validate alongside other data sources before taking further action.*

4.4 User experience and data integration

The evaluation has found that the implementation of the dashboard has resulted in a poor user experience for officers. There are two main reasons for this. The first relates to the user interface, with officers reporting that the graphical interface of the dashboard is not user friendly. Regardless of the technical performance of the system, a poor user interface is likely to deter officers from regularly accessing the dashboard, resulting in low adoption levels. Any future development of the

dashboard should focus on continuously improving user interface and accessibility features, by requesting regular feedback from officers.

> **Recommendation 7:** Any future development of the application should focus on improving the front-end user experience, incorporating best practice in data visualisation and software accessibility. The dashboard should incorporate a 'Feedback' section, where users can provide written feedback on the application and submit suggestions for improvement. Monthly feedback meetings should be held for officers to provide verbal feedback to the development team. Where possible, behavioural scientists should be consulted to advise on the most effective visual presentation of outputs to support decision-making and minimise risk of cognitive bias.

The second reason for the poor user experience relates specifically to data integration challenges. As is common for police data systems, officers report needing to access multiple systems separately to triangulate the information provided by the RFSDi score. A common request was for all relevant information to be available on the same system, rather than needing to access the dashboard in parallel to other existing police data systems. While this is symptomatic of a broader data integration challenge across UK policing, future efforts should focus on extracting the most pertinent data points from other key databases to be presented alongside the RFSDi score and model outputs.

> **Recommendation 8:** The most pertinent data points from other information management systems (most notably custody images) should be included within the RFSDi dashboard. Integrating the dashboard within the existing Connect system is likely to achieve this and should be a priority for any future development of the application.

4.5 Training and guidance

Finally, there was broad consensus that further training and written guidance would be required if the application were to be deployed for enduring use. In addition to the specific limitations of the system identified by interviewees, the research also highlighted a general lack of sufficient understanding among PCs and Sergeants of how the application works in practice, and how the insights derived from the system are expected to be integrated within existing offender management processes. The fact that one Sergeant reported leaving the training session 'as confused as when I started' is concerning, and suggests that further training should be a high priority if the application is to be deployed operationally.

This lack of training could also be a main reason why PCs and Sergeants reported not finding the application useful, and not routinely using it as part of their offender management responsibilities. If users do not sufficiently understand how the system works – including the strengths and limitations inherent in the application – they are unlikely to feel comfortable using it to inform operational decision-making. This is particularly important when the decisions informed by the application will have a direct impact on individuals being assessed, requiring a high degree of confidence and accountability throughout all stages of the decision-making process.

> **Recommendation 9:** Additional training should be delivered to all officers with access to the RFSDi dashboard and predictive IOM model. This should cover how the application is intended to be used, the input variables used to calculate the RFSDi score and to build the predictive model, and an overview of the inherent limitations of the statistical techniques underpinning the system.

Recommendation 10: Written guidance should be developed for all officers with access to the RFSDi dashboard and IOM model. This guidance should include a summary of how the application generates the harm scores and statistical predictions, as well as a workflow diagram of how the algorithmically-generated insights should be integrated into existing offender management processes.

4.6 Conclusions

This process evaluation sought to scrutinise the implementation of the beta-testing phase of the WMP RFSDi harm score and Integrated Offender Management model, and assess the potential benefits and limitations of the application in an operational policing environment. The research comprised semi-structured interviews, focus groups, and a practitioner survey of all WMP officers involved in the beta-testing phase of the project.

The research did not establish sufficient evidence in favour of deploying the RFSDi dashboard and accompanying IOM model for long-term operational use. Further, detailed evaluation research is required to conclusively establish the potential benefits and limitations of the system before it is deployed for wider operational use. This should include developing clear longitudinal evaluation metrics, and a theory of change describing the overall intended outcomes of the project.

Perhaps the most notable finding of the research is a fundamental divergence in views between PCs and Sergeants on the one hand, and Inspectors on the other. Inspectors were significantly more positive regarding the new application, while the feedback returned by PCs and Sergeants was unanimously negative. One potential reason for this is that Inspectors involved in the research had been more directly involved in the development of the application, emphasising the critical importance of ensuring end-user engagement at an early stage in the design process. The beta-testing may have yielded a more positive outcome if PCs and Sergeants had been more closely involved in the early application development stage, for instance to advise on user interface requirements. Perceived deficiencies in the user experience were a main reason reported by PCs and Sergeants for why the RFSDi score and IOM model are not routinely used as part of offender management processes.

At an operational level, PCs and Sergeants report encountering both over-classification errors (i.e., an over-estimation of risk) and under-classification errors (i.e., individuals not being flagged by the system despite posing a high level of risk). In the offender management context, false negatives are likely to be a more 'costly' error, as they could result in high-risk nominals being erroneously de-selected. For this reason, the RFSDi score should not be used to assess offenders who are currently subject to offender management orders; such individuals should be subject to individual risk assessment incorporating structured professional judgement to assess whether they are eligible for de-selection. Instead, the RFSDi score could provide greater value as a 'risk identification' tool, enabling the identification of high-risk nominals not currently subject to offender management orders, who should be prioritised for more in-depth risk assessment.

Three priority issues should be addressed if the application is to be deployed for enduring use. This includes ensuring a comprehensive training plan for users, including written guidance summarising how the harm score and model outputs are calculated. This is important not just to ensure officers understand how to use the system in conjunction with existing processes, but also to maintain accountability throughout the full decision-making chain. At the technical level, efforts should be made to include more selection criteria within the dashboard to enable users to filter according to specific crime types, and to extract relevant data points from other systems to be integrated within

the dashboard itself. The ideal outcome described by interviewees would be for RFSDi harm scores to be integrated within the existing Connect information management system.

The findings reported in this evaluation report relate specifically to the beta-testing phase of this project. They should not be interpreted as generalisable to other data-driven risk assessment tools, whether used within West Midlands Police or elsewhere. The recommendations are intended to prioritise areas of focus for the future development of the RFSDi score and IOM model. The research remains inconclusive on whether the application will ultimately provide the operational benefits for the force that were envisaged at the time of its development.

References

Brown, S.C. *et al.* (2002) 'Exploring complex phenomena: Grounded theory in student affairs research', *Journal of college student development*, 43(2), pp. 173–183.

Bryman, A. (2016) *Social research methods*. Oxford university press.

Cassell, C. (2005) 'Creating the interviewer: identity work in the management research process', *Qualitative research*, 5(2), pp. 167–179.

Charmaz, K. and Belgrave, L.L. (2007) 'Grounded theory', *The Blackwell encyclopedia of sociology* [Preprint].

Chen, T. *et al.* (2015) 'Xgboost: extreme gradient boosting', *R package version 0.4-2*, 1(4), pp. 1–4.

College of Policing (2018) *The Policing Evaluation Toolkit*.

Corbin, J. and Strauss, A. (2014) *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Corbin, J.M. and Strauss, A. (1990) 'Grounded theory research: Procedures, canons, and evaluative criteria', *Qualitative sociology*, 13(1), pp. 3–21.

Creswell, J.W. (2002) *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall Upper Saddle River, NJ.

Denzin, N.K. (2001) 'The reflexive interview and a performative social science', *Qualitative research*, 1(1), pp. 23–46.

Education Endowment Foundation (2015) 'EEF evaluation: A cumulative approach', *London: EEF* [Preprint].

Fossey, E. *et al.* (2002) 'Understanding and evaluating qualitative research', *Australian & New Zealand Journal of Psychiatry*, 36(6), pp. 717–732.

Fox, C. *et al.* (2018) 'Piloting different approaches to personalised offender management in the English criminal justice system', *International Review of Sociology*, 28(1), pp. 35–61. Available at: https://doi.org/10.1080/03906701.2017.1422886.

Fujii, L.A. (2012) 'Research ethics 101: Dilemmas and responsibilities', *PS: Political Science & Politics*, 45(4), pp. 717–723.

Gioia, D.A., Corley, K.G. and Hamilton, A.L. (2013) 'Seeking qualitative rigor in inductive research: Notes on the Gioia methodology', *Organizational research methods*, 16(1), pp. 15–31.

Glaser, B.G. and Strauss, A.L. (2017) *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Longhurst, R. (2003) ‘Semi-structured interviews and focus groups’, *Key methods in geography*, 3(2), pp. 143–156.

Newcomer, K.E., Hatry, H.P. and Wholey, J.S. (2015) ‘Conducting semi-structured interviews’, *Handbook of practical program evaluation*, 492.

Nunkoosing, K. (2005) ‘The problems with interviews’, *Qualitative health research*, 15(5), pp. 698–706.

Orb, A., Eisenhauer, L. and Wynaden, D. (2001) ‘Ethics in qualitative research’, *Journal of nursing scholarship*, 33(1), pp. 93–96.

Pimple, K.D. (2002) ‘Six domains of research ethics’, *Science and engineering ethics*, 8(2), pp. 191–205.

Saldaña, J. (2014) ‘Coding and analysis strategies’, in *The Oxford handbook of qualitative research*.

Sherman, L., Neyroud, P.W. and 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(3), pp. 171–183.

Smith, D. (2003) ‘Five principles for research ethics’, *Monitor on psychology*, 34(1), p. 56.

Thomas, D.R. (2003) ‘A general inductive approach for qualitative data analysis’.

West Midlands Police (2019) *Minutes, Ethics Committee Meeting July 2019.* Available at: https://www.westmidlands-pcc.gov.uk/archive/april-2019/.