Ministry of Justice

Revalidation: Risk of recidivism tools

An evaluation of the actuarial instruments developed to assess recidivism risk in England and Wales

Andrew Craik, Lu Han, Liam Sullivan,
Dr Julia Landsiedel, Dr Thomas Travers,
Christopher Spaull, and Dr Philip Howard

Ministry of Justice

Ministry of Justice Analytical Series
2024

Data and Analysis exists to improve policy making, decision taking and practice by the Ministry of Justice. It does this by providing robust, timely and relevant data and advice drawn from research and analysis undertaken by the department's analysts and by the wider research community.

Disclaimer

The views expressed are those of the authors and are not necessarily shared by the Ministry of Justice (nor do they represent Government policy).

First published 2024

Any enquiries regarding this publication should be sent to us at

Evidence_partnerships@justice.gov.uk

This publication is available for download at
http://www.justice.gov.uk/publications/research-and-analysis/moj

ISBN 978-1-911691-20-4

Acknowledgements

We acknowledge the example set to us by Megan Whewell and Ben Fortescue, of Ministry of Justice, whose earlier work on these topics provided us with a useful guide on the direction and structure for this study.

We acknowledge our external reviewers of both methodology and content.

Further, we acknowledge the subject matter advice and guidance provided by our colleagues in HMPPS, Dr Mark Farmer, Cindy Keehner and Helen Walton.

The authors

Andrew Craik, Lu Han, Liam Sullivan, Dr Julia Landsiedel, Dr Thomas Travers, Christopher Spaull and Dr Philip Howard

Contents

List of tables

1. Terms and abbreviations 1

2. Executive Summary 7

3. Action-based summary 10

4. Introduction 13
4.1 Serious reoffending (predicted by RSR) 14
4.2 Other types of reoffending 15
4.3 Research Questions 16
4.4 Structure of the report 16
4.5 Population subgroups 17

5. Method 19
5.1 Risk Predictors 19
5.2 Analytical Datasets 23
5.3 Model Evaluation 25
5.4 Limitations 28

6. Results 29
6.1 Descriptive statistics of population and missing data 29
6.2 Proven rates of reoffending 30
6.3 Accuracy in the prediction of all (or general) reoffending 33
6.4 Accuracy in the prediction of broad (OVP-type) violent reoffending 36
6.5 Accuracy in the prediction of serious nonsexual violent (SNSV) reoffending 39
6.6 Accuracy in the prediction of contact sexual reoffending 41
6.7 Accuracy in the prediction of reoffending involving indecent images of children 43
6.8 Accuracy in the prediction of all serious (RSR-type) reoffending 44

7. Conclusions 47
7.1 Overview 47
7.2 Predictors by offence type 47

8. References 50

List of tables

Table 1: Table of abbreviations and terms used in the study 1

Table 2: Summary of results by risk predictor 8

Table 3: Summary of risk predictors and offending outcomes 20

Table 4: Reporting convention for calibration 26

Table 5: Reporting convention for discrimination 28

1. Terms and abbreviations

Table 1: Table of abbreviations and terms used in the study

Term or abbreviation	Description
Actuarial	Relating to actuarial science, the discipline that applies mathematical and statistical methods to assess risk
AUC	Area under the curve. The AUC is an aggregated accuracy metric used to say how well a model has performed in its predictions analogous to Harrell's C-index used in this report. A value of 0.5 would indicate that a model has performed no better than 'random guessing'. The higher (and closer to 1.0) the value is, the better model's predictions. Conversely, the lower (and closer to 0.0) the value is, the poorer the model's predictions are.
Baseline rate	Used in the context of sexual reoffending risk for women with sexual offending history. A calculated static rate of reoffending added to a woman's total RSR score, based on an historic group of women with sexual offending history
Broad violence	Used in the context of all violent offences that fall under the definition of violence that the OVP predictor was designed to predict (hence, OVP-type violence). See Appendix B - Types of reoffending
Calibration	Also referred to as accuracy - a comparison of a predictor's (see Predictor) mean predicted risk score against the actual rates of reoffending observed. Also see residual.
Caseload	One of the two cohorts studied in the revalidation study. A snapshot of the community (probation) population as at a single point in time, currently 30 June 2018. Have varied offence-free time. Additionally, see Starts and offence-free-time.
Censored	Used in the context of survival analysis, and specifically relating to 'right censored' data. A right-censored data point is an individual who has been removed from the study (partway) or reasons not related to the event (offence) being studied. This can be a prison recall or a custodial sentence. A 'left censored' data point would be one where the study start date is unknown (not an issue within this study)
Censoring	See censored
C-index	See Harrell's C-index

Term or abbreviation	Description
Cohen’s h	Cohen's h, popularized by Jacob Cohen, is a measure of distance between two proportions, allowing us to describe the difference between two proportions as "small", "medium", or "large". Referred to as the 'effect size'
Cohort	A group of offenders - used primarily in the context of either the 'caseload' or the 'starts' cohorts. See caseload, starts
Concordance Index	See discriminative validity
Cox proportional hazards survival	The Cox proportional hazards is a statistical model used to study the relationship (or association) between a varying value (such as a person's predicted risk score) and the time for an 'event' of interest to happen. In the context of this study, we are looking to understand the relationship between the predicted risk score and the time till a reoffence.
CRC	Community Rehabilitation Company. Run by a mix of providers from private, statutory and voluntary sectors, contracted to deliver community sentences for medium and low-risk offenders, and paid, in part, for results achieved in reducing reoffending
Discrimination	See discriminative validity
Discriminative validity	Used in the context of the accuracy of a risk predictor. That is, how well a risk score (the prediction) was at discriminating between lower- and higher-risk offenders. In this study we use (for survival analysis) Harell's C-index. The higher the value of the C-Index, the better discrimination. It can be interpreted as the 'probability that a randomly selected individual who reoffended had a higher risk score than another randomly selected individual who did not reoffend (or reoffended later)'. See 'Model discrimination - Harell's C-index' in report. As with AUC a value of 0.5 would indicate that a model has performed no better than 'random guessing'. The higher (and closer to 1.0) the value is, the better model's discrimination.
DV	Domestic violence
Dynamic factors	Any input to a risk predictor which could change (i.e. is dynamic). These include things like (for example) severity of need for accommodation, employment, substance misuse and levels of impulsivity
Dynamic predictor	See 'static/dynamic predictor'
Effect size	See Cohen’s h
Harrell's C-index	See discriminative validity
HMPPS	His Majesty's Prison and Probation Service

Term or abbreviation	Description
IIOC	Indecent images of children
LDC	Learning disabilities and challenges
MoJ	Ministry of Justice
NPS	National Probation Service
OASys	Offender Assessment System, a structured assessment instrument used to record the risks and needs of eligible offenders in prisons and probation trusts across England and Wales.
Offence-free	See offence-free time
Offence-free time	The number of whole months that an individual has been in the community without a proven (conviction) reoffence
OFM	Offence-free months, see also offence-free time
OGP/OGP1	OASys General reoffending Predictor - version 1.
OGP2	OASys General reoffending Predictor - version 2. See also footnote on page 5 TBC.
OGRS	Offender Group Reconviction Scale
OGRS3	Offender Group Reconviction Scale - version 3. No reference to version 1 or 2 are made in the report.
OGRS4	Offender Group Reconviction Scale - version 4. Umbrella term used for both OGRS4/G and OGRS4/V predictors. OGRS4 was developed but never implemented. It would have enabled offenders with no OASys layer 3 assessment to have a predictor of violent reoffending (OGRS4/V)
OGRS4/G	OGRS4 general predictor - for 'all reoffending risk'
OGRS4/V	OGRS4 general predictor - for 'violent reoffending risk'
OSP	OASys sexual predictor. Umbrella term used for sexual prediction tools (which form part of RSR)
OSP/C	OASys sexual predictor for contact sexual reoffending
OSP/I	OASys sexual predictor for reoffending involving indecent images
OVP	OASys violence predictor
OVP1	OASys violence predictor - version 1
OVP2	OASys violence predictor - version 2
OVP-type violence	See broad violence
PNC	Police National Computer. The PNC is a national database of information available to all police forces, law enforcement agencies and other specified bodies throughout UK.

Term or abbreviation	Description
Predictive validity	See discriminative validity
Predictor	An actuarial risk instrument used to estimate the likelihood of reoffending
Proven reoffending	A conviction for an offence. The risk tools are designed to predict proven reoffending. To make the distinction between offending which is not detected by police or proven at court
p-value	Used in this report in the context of the two-tailed Z-test (see Z-test). The probability (p) value provides a measure of how likely it is that the difference between the actual and predicted rates of reoffending is due to chance. A Z-value is calculated from the two proportions (actual and predicted) and is compared to the 'normal' distribution to determine the probability or p-value. Smaller values of p indicate differences that are less likely to be due to chance (given the underlying assumptions made in the statistical test), that is, there is evidence to suggest that the actual and predicted rates of reoffending are different. See statistical significance and Z-value.
Residual	The difference (in percentage points) between the actual and the predicted rate of reoffending
Risk band	Risk scores can be grouped into bands such as 'low', 'medium', 'high'
RSR	Risk of serious recidivism
RSR SNSV	The component of the RSR predictor designed to predict serious nonsexual violence (SNSV). Umbrella term for 'SNSV static' and 'SNSV static/dynamic'
Serious reoffending	All offending under the definition of the RSR predictor - see Appendix B - Types of reoffending
Sexual history	A current or historic proven conviction for a sexual offence
SNSV	Serious nonsexual violence. A subset of all violence considered the most seriously harmful
SNSV brief	See SNSV static
SNSV static	The brief version of the RSR SNSV predictor. Uses a reduced set of questions (static factors only). Available for everyone. See RSR SNSV, static factors

Term or abbreviation	Description
SNSV static/dynamic	The extended version of the RSR SNSV predictor. Uses a combination of static and dynamic factors to calculate risk. Available for the subset of people that have a full OASys layer 3 assessment See RSR SNSV, static factors, dynamic factors
Starts	One of the two cohorts studied in the revalidation study. Consists of individuals who have started a community sentence between 1 July 2018 and 31 December 2018. Everyone has zero offence-free time. Additionally, see caseload and offence-free time
Static factors	Any input to a risk predictor which cannot change (i.e. is static) over the course of a sentence, such as 'age at the commencement of risk' and criminal history.
Static predictor	Refers to any predictor which does not require dynamic factors (or a OASys layer 3) to calculate the risk score. See static factors, dynamic factors
Static/dynamic predictor	Refers to any predictor which requires dynamic factors (or a OASys layer 3) in addition to static factors to calculate the risk scores. See static factors, dynamic factors
Statistical significance	Used in the context of the two-tailed Z-test. A result is considered 'statistically significant' when the p-value is below an arbitrary threshold (often cited as 0.05) to indicate that there is evidence to reject the 'null hypothesis'. In the context of this report, the null hypothesis is that "the actual rate of reoffending is the same as the predicted rate". A different threshold may be set where multiple comparisons are made to reduce the likelihood of 'false positive' results.
Statistically significant	See statistical significance
Survival analysis	Survival analysis is a branch of statistics for analysing the expected duration of time until one event occurs, such as a proven reoffence
Tiering	The tiering model takes into account risk (risk of serious harm, serious recidivism, MAPPA level and additional risk-related factors) and needs (for example, other reoffending risk and criminogenic needs). Cases are allocated based on their tier combined with clinical and professional judgement to determine grade of probation practitioner most appropriate for the case.

Term or abbreviation	Description
Two-tailed test	Refers to a statistical test where we want to check whether the result is greater than or less than a hypothesised value. As opposed to a one-tailed test which specifies a particular direction to check for differences. Also see Z-test/Z-value.
Z-test	A one-proportion z-test. A statistical test used to compare the proportion of a sample (the actual rate of reoffending) to a known (or hypothesised) proportion. It is used to test a hypothesis about the predicted rate of reoffending (the hypothesised proportion) and assumes that the sample is drawn from a population with a statistically 'normal' (or gaussian) distribution.
Z-value	The z-value represents the number of standard errors that the actual rate of reoffending is from the predicted rate and can be converted into a probability (p) value. It is used to determine whether the difference between the two rates is 'statistically significant'. Also see statistical significance.

2. Executive Summary

In its policies for risk management, targeting of rehabilitative interventions, and the intensity of supervision ("tiering") for offenders, His Majesty's Prison and Probation Service (HMPPS) recognises a need to estimate the risk of reoffending. Understanding the likelihood of an individual reoffending allows the Ministry of Justice (MoJ) to plan the management of a person in prison or on probation, and to incorporate appropriate interventions and support that will ensure an offender's greatest chances of successful rehabilitation with reduced reoffending. To calculate the likelihood of reoffending HMPPS has access to several risk predictors, each designed to calculate the risk of one of five types of reoffending:

- All (or general) reoffending
  - Offender Group Reconviction Scale – version 3 (OGRS3)
  - Offender Group Reconviction Scale – version 4: General predictor (OGRS4/G)
  - The Offender Assessment System (OASys) General reoffending Predictor – version 2 (OGP2)
- Broad nonsexual violence
  - Offender Group Reconviction Scale – version 4: Violent reoffending predictor (OGRS4/V);
  - OASys violence predictor – version 1 (OVP1)
  - OASys violence predictor – version 2 (OVP2)
- Serious nonsexual violence (SNSV)
  - Risk of serious recidivism Serious nonsexual violence (RSR SNSV)
- Contact sexual reoffending
  - OASys sexual predictor for contact sexual offending (OSP/C)
- Sexual reoffending involving indecent images
  - OASys sexual predictor for indecent images of children (OSP/I)

RSR SNSV, OSP/C and OSP/I are components of the Risk of Serious Recidivism (RSR) predictor for all serious reoffending.

Over time, as reoffending patterns change, it may be that some of these risk predictors no longer perform optimally at calculating individuals' risk of reoffending. It is therefore important for the department periodically to check that the risk predictors in use continue to work correctly.

This report describes recent analysis to evaluate the effectiveness of the predictors, by checking that they still accurately calculate the reoffending risk of individuals who were in the prison and probation system in 2018.

To understand risk predictor performance, two separate aspects of their predictions were checked:

- **Model Calibration**, which tests how accurately the models can calculate how likely reoffending is to occur. This answers the question, 'If a model predicts a reoffending rate of X% for a group of offenders, how far away from this is the actual rate of reoffending?'
- **Model discriminative validity**, which assesses how well the models can differentiate between high and low risk individuals. This answers the question, 'If a model says individual X is high risk and individual Y is low risk, how likely is that statement to be true?'

The following table sets out the results of this work. The definitions of each of the headers are as follows:

- Predictor – Name of the predictor
- Implemented – Is the predictor currently in use by HMPPS?
- Calibration – Are these predictors well calibrated?
- Discriminative validity – How well do these predictors discriminate between high and low risk individuals?

Table 2: Summary of results by risk predictor

Predictor	Implemented	Calibration	Discriminative Validity
OGP2	No	Well calibrated	Good
OGRS3	Yes	Well calibrated	Good
OGRS4/G	No	Well calibrated	Good

Predictor	Implemented	Calibration	Discriminative Validity
OGRS4/V	No	Small miscalibration	Good
OVP1	Yes	Well calibrated	Acceptable
OVP2	No	Well calibrated	Good
RSR SNSV Static	Yes	Well calibrated	Good
RSR SNSV Static/Dynamic	Yes	Well calibrated	Good
OSP/C	Yes	Very large miscalibration	Acceptable
OSP/I	Yes	Large miscalibration	Excellent
RSR	Yes	Small miscalibration	Good

The work done shows that most of the predictors are well calibrated and have good discriminative validity. However, there are a few exceptions. The two sexual predictors, OASys Sexual Predictor of Contact offending (OSP/C) and OASys Sexual Predictor of Indecent Image based offending (OSP/I) show very large and large miscalibrations, respectively, meaning the predicted rate of proven reoffending does not accurately match the actual rate of proven reoffending.

The recommendation from this report is therefore for further work to be done to analyse the performance of two sexual reoffending predictors. This work has now been completed and published in a companion report (Emeagi et al., 2024)

3. Action-based summary

What's working well

Overall, most predictors in use are well suited to predict their specific types of reoffending.

With one exception, all predictors of general, broad violent and serious nonsexual violent reoffending were well calibrated overall: that is, the rates of reoffending they predicted across all offenders were similar to the actual rates of reoffending across all offenders, rather than being over- or underestimates of the actual rates. Nearly all predictors perform well at telling the difference between higher and lower risk offenders and assigning appropriate risk scores meaning they have good discrimination. The indecent images predictor had excellent discrimination, meaning that it was even better at assigning risk scores to higher and lower risk offenders.

What requires improvement

The two sexual predictors in operational use (OSP/C and OPS/I) underpredicted rates of reoffending leading to the conclusion that they were not well calibrated. The only predictor for contact sexual reoffending, OSP/C, was found to perform worse at differentiating between higher and lower risk offenders than in previously published research. This means that the predictor struggled in some instances to assign higher risk scores to those that should have them, although it was still above the threshold for acceptable performance.

The RSR algorithm (comprising RSR SNSV, OSP/C and OSP/I) marginally underpredicted the rate of reoffending, which equated to a small miscalibration.

A full OASys assessment is a more in-depth assessment of an individual that includes recording dynamic risk factors - aspects of a person's life that can indicate risk of reoffending and change over time such as accommodation, alcohol misuse and emotional well-being. Including risk factors like these improves model performance. However, full OASys assessments are not conducted for many people and as such their dynamic risk factor data would be missing. Currently, none of the predictors of general reoffending in operational use account for dynamic risk factors and are not benefiting from the increase

in performance these factors can provide. Conversely, the sole predictor in use for broad (OVP-type) violent reoffending, OVP1, requires a full OASys assessment. Therefore, there is no applicable predictor of this type of reoffending for a large proportion of individuals.

As the time since someone's last offence increases (offence-free time) their risk of reoffending goes down. Including a measure of offence-free time improves model performance when used to assess people who are in the middle of their sentence. However, there are no predictors of all (or general) and broad (OVP-type) violent offence types in operational use that consider offence-free time.

Proposed next steps

Predictors for all violent (broad, OVP-type) reoffending

Bringing existing predictors into operational use, like OGP2 and OGRS4/V, or developing two new predictors that account for dynamic risk factors and do not require a full OASys assessment respectively would fill the gaps in our predictive abilities. This would allow prediction of general offending to benefit from the improvements consideration of dynamic risk factors bring and facilitate the prediction of OVP-type violent offending for the large proportion of the population that do not have a full OASys assessment.

Predictors for contact sexual reoffending

A specific study will be necessary to understand the root cause of the underperformance of OSP/C and improve its performance moving forward. This work has been undertaken and is published alongside this report (Emeagi et al., 2024).

Offence-free time

Currently, with the exception of RSR, operational practice for assessment of risk of reoffending is to assess an individual at the point at which they are at risk in the community (i.e. when they leave prison and enter the probation system). At that point individuals will have zero offence-free time and, therefore, predictors have not needed to account for it. However, in the future the MoJ and HMPPS may wish to expand the practice of reconsidering actuarial risk partway through a sentence to all reoffending types; allowing them to reflect the general fall in rates of reoffending as offence-free time increases. Therefore, it would be worthwhile to ensure newly developed predictors account for offence-free time, or to bring existing predictors that do into operational use where they

exist. Currently, the only operational predictors of general reoffending (OGRS3) and OVP-type violent reoffending (OVP1) do not account for offence-free time. Yet, both types of reoffending have existing, but not operational, predictors (OGRS4/G, and OGRS4/V and OVP2 respectively) that do. Utilising these existing resources or developing wholly new predictors that do account for offence-free time would be beneficial.

4. Introduction

In its policies for risk management, targeting to rehabilitative interventions, and the intensity of supervision ("tiering"), His Majesty's Prison and Probation Service (HMPPS) currently recognises a need to estimate the risk of five types of reoffending:

- All (or general) reoffending;
- Broad nonsexual violence;
- Serious nonsexual violence (SNSV);
- Contact sexual reoffending; and
- Sexual reoffending involving indecent images.

Understanding the likelihood of an individual reoffending allows the MoJ to plan the management of a person in prison or on probation, to incorporate appropriate interventions and support that will ensure their greatest chances of successful rehabilitation with reduced reoffending.

Twelve actuarial risk assessment instruments have been developed by the Ministry of Justice (MoJ) of which eight are in use by HMPPS. These instruments (the predictors, going forward) provide an estimate of the one- and two-year probability that an individual will have a proven reoffence. Risk predictors are an important objective component in practitioner decision making. Risk predictors aim to remove individual bias by including an element of consistency and fairness between practitioners and are used as part of the overall risk assessment process alongside professional judgement.

Other assessments, based on structured professional judgment, are used in the management of risk of intimate partner abuse (through the Spousal Assault Risk Assessment) and offences judged to cause "serious physical and/or psychological harm" (Risk of Serious Harm). These are not actuarial tools and were out of scope of this study.

For each of the types of offending described above, there have been one or more predictors developed. HMPPS uses a set of predictors generated in the 2000s and 2010s (Table 3); replacements to predictors created in the 2000s were generated in 2010s but it was not possible, due to operational resource pressures, to implement them.

This report describes work done to evaluate the effectiveness of the predictors as set out in Section 4.3 - Research Questions.

4.1 Serious reoffending (predicted by RSR)

One focus of this study was on the prediction of **serious reoffending** – estimated through the 'risk of serious recidivism' (or RSR) predictor. There have been no prior publications on the development or validation of the RSR predictor.

RSR was initially designed and built through the follow up of a sample of offenders between 2010 and 2012. As well as informing Risk of Serious Harm (RoSH) judgements, between 2014 and the unification of the probation service which concluded in 2021, RSR was used for allocation between National Probation Service (NPS) and Community Rehabilitation Company (CRC) caseloads.

RSR comprises three sub-predictors, used to estimate risk of three of the five types of reoffending referred to earlier. These three types of reoffending are considered the most seriously harmful:

- Serious nonsexual violence (or SNSV): predicted by RSR SNSV;
- Contact sexual offending: predicted by the Offender Assessment System (OASys) predictor for contact sexual offending or OSP/C (Howard & Wakeling, 2021); and
- Sexual offending involving indecent images: predicted by OSP/I (the OASys Sexual reoffending predictor for indecent images of children; Howard & Wakeling, 2021).

The RSR score is calculated as the (arithmetic) sum of the three component risk scores mentioned above: RSR SNSV + OSP/C + OSP/I. In practice, practitioners are only presented with the two OSP risk levels (where calculated) and the total RSR score. Where OSP scores are not calculated (that is, for all females and males with no sexual offending history), the total RSR is equal to the SNSV score.

SNSV offences are a subset of all violent offences (see section 4.2, Other types of reoffending, below). That is, the violent offences considered to be the most serious under the RSR definition (e.g. murder, manslaughter, grievous bodily harm). See Appendix B for a complete list.

4.2 Other types of reoffending

This study also investigated a suite of other predictors developed by the MoJ for the prediction of the two remaining types of reoffending: all reoffending and broad violent reoffending. Further detail on the definition of types of reoffending are provided in Appendix B.

All reoffending

For the prediction of all (or general) reoffending three predictors have been developed and one, OGRS3 (Howard, 2009a), is in use operationally.

The general reoffending predictors are:

- OGRS3: Offender Group Reconviction Scale (OGRS) – version 3
- OGRS4/G: Offender Group Reconviction Scale – version 4 (Howard, 2015)
- OGP2:[1] OASys General Reoffending Predictor – version 2 (Howard, 2014)

Broad (OVP-type) violence

Broad violent reoffending can be described as any (serious and nonserious) nonsexual violent offence. Due to the establishment of a functional classification of nonsexual violent offences that took place in the development of version 1 of the OASys Violence Predictor (OVP; Howard, 2009b), such offences are referred to as "OVP-type" offences in this study. OVP-type violence does include the more serious forms of violent reoffending (under the RSR definition), including murder and manslaughter. For broad violent reoffending three predictors have been developed, only OVP1 is in use operationally:

- OGRS4/V: Offender Group Reconviction Scale – version 4 (Howard, 2015)
- OVP1: OASys Violence Predictor (OVP) – version 1 (Howard, 2009b)
- OVP2: OASys Violence Predictor (OVP) – version 2 (Howard, 2014)

[1] A fourth 'general' predictor, OGP1, is in use but is calibrated for nonviolent offending only.

4.3 Research Questions

The key research questions this report addressed were:

- How have risk predictors performed in identifying those more likely to commit reoffences?; and
- Do the risk predictors correctly estimate rates of reoffending?

To answer these questions, the predictors were evaluated through two measures:

- **Calibration**: a measure of how close the **predicted rate** of reoffending is to the **actual rate** of reoffending; and
- **Discriminative validity**: a measure of how well an individual predictor discriminates between higher- and lower-risk offenders, measured via Harrell's C-Index (see Appendix C).

4.4 Structure of the report

The main body of the report focusses on analysis of the **starts population**.[2] To understand the relationship of model accuracy and offence-free time (see section Offence-free time in 5.1), results based on the **caseload population** are presented. Tables of analyses are referenced in the main report and provided in Appendix A of the accompanying technical appendix.

First, the proven rates of reoffending were analysed for different subgroups of the offender population, provided as two-year proven reoffending rates.

Results are structured by **type of reoffending** (e.g. all proven reoffending, broad violence, and so on) and, within each type of reoffending calibration and model accuracy is assessed (see section 5.3 - Model Evaluation).

[2] Starts – a cohort of individuals who are scored at the point at which they entered the HMPPS community caseload. The starts cohort will have higher mean scores than the caseload population (i.e. all those on the HMPPS community caseload) for two reasons. First, higher risk individuals will tend to reoffend early and not be represented in the caseload population. Second, on those predictors that incorporate 'offence-free time' (time passed on the community caseload without reoffending), scores on the caseload population will be lower to recognise where an individual has been offence-free for some time.

4.5 Population subgroups

Throughout the report results are presented by various characteristics for which comprehensive data were available for. These include some protected characteristics (age, gender, ethnicity and disability) and other case characteristics.

In some cases, analyses were only provided for specific subsets of the population. These include all offenders where we had a full OASys assessment, men with sexual offending history and all women and men without sexual offending history.[3]

Case characteristics

In addition to protected characteristics, it was possible to identify those with probable Learning Disabilities and Challenges (LDC) using a screen included in OASys. Domestic violence (DV) perpetration status was best understood using OASys data. Both are only available where we have an OASys assessment, however in some cases this data may still be missing even when an OASys was completed.

'Former' DV perpetrators were those recognised as lifetime perpetrators in the Relationships section of OASys, but without 'physical violence towards partner' noted in the Analysis of Offences section, which deals with current offences. The term 'former' is therefore shorthand, also encompassing some cases of current domestic violence where physical violence to partner does not occur.

OASys assessment (with OASys)

Some predictors are only available for the subset of the study population where we have a complete OASys Layer 3[4] assessment (i.e. 'with OASys'). These predictors are said to be static/dynamic.[5] Where comparisons are made between predictors, it was considered more accurate to ensure they were compared on the same population. That is, if a predictor required an OASys to calculate it (e.g. OVP1) its performance would be

[3] Also referred to as 'women and men with no history'

[4] See section OASys (Layer 3) assessments in Appendix B

[5] Static risk factors are those that those which cannot be deliberately changed over the course of a sentence (e.g. criminal history). Dynamic factors/criminogenic needs include 'dynamic' risk factors which change as the service user passes through their rehabilitative journey of the criminal justice system.

compared to other predictors for the 'with OASys' population only (even if the other predictors did not require the OASys assessment to calculate it, e.g. OGRS4/V).

Men with sexual offending history

Performance of the two sexual predictors is only provided for men that have been sanctioned for sexual offending (i.e. have a sexual offending history).

The OASys sexual predictors, OSP/C and OSP/I (for sexual contact and indecent image offending, respectively), were designed to predict the rates of sexual reoffending for **men with a sexual offending history**. That is, no OSP score is calculated for men with no sexual history[6] or any women. In practice, a baseline rate[7] of contact sexual risk *is* added to a woman's total RSR score (where she has sexual offending history), but no separate OSP/C risk score is calculated.

[6] In practice an individual may be scored for OSP if they are lacking a conviction for sexual offence but have a current sexually-motivated conviction.

[7] The baseline rate for women with sexual offending history is 1/193, as this rate of contact sexual reoffending was observed for such women in the sample used to construct RSR.

5. Method

The process for evaluating the performance of the predictors was as follows:

- Produce an analytical dataset and, for each offender and type of reoffending being studied, derive the two-year probability of reoffending using the predictors described above;
- Identify the actual reoffending (the 'offending outcomes') over the two-year study period; and
- Evaluate the performance of the predictors for each of the outcomes.

This section starts by outlining the tools used to predict different types of reoffending and then provides some high-level information which should help the reader in understanding the analysis set out in the report. More detail on the study data is provided in Appendix C – Method for evaluation of models.

5.1 Risk Predictors

Within the scope of this study were actuarial risk instruments developed or in use by the MoJ. As noted in the Introduction, HMPPS currently recognises a need to estimate the risk of five types of reoffending:

- All reoffending;
- Broad nonsexual violence (under the OVP definition, see Appendix B); and
- All serious reoffending which is composed of three subtypes of reoffending:
  - Serious nonsexual violence;
  - Contact sexual reoffending; and
  - Sexual reoffending involving indecent images.

There are a total of twelve distinct predictors which have been developed, or are in use, by the MoJ. The predictors studied in this report not only vary in the type of reoffending they were designed to predict, but also their data requirements and use of offence-free time (Table 3)

Table 3: Summary of risk predictors and offending outcomes

Risk Predictor	Reoffending outcome	Accounts for offence-free months?	Implemented	Requires full OASys?	Cohort used for development
OGRS3	All reoffending	No	Yes	No	January – March 2002
OGRS4/G	All reoffending	Yes	No	No	2005 to 2008
OGP1	Nonsexual, Nonviolent reoffending	No	Yes	Yes	2002 to 2004
OGP2	All reoffending	Yes	No	Yes	2005-2008 (with an OASys)
OGRS4/V	Broad violence	Yes	No	No	2005 to 2008
OVP1	Broad violence	No	Yes	Yes	2002 to 2004
OVP2	Broad violence	Yes	No	Yes	2005-2008 (with an OASys)
RSR SNSV Static	Serious (nonsexual violence)	Yes	Yes	No	2005-2008 (with an OASys)
RSR SNSV static/dynamic	Serious (nonsexual violence)	Yes	Yes	Yes	2005-2008 (with an OASys)
OSP/C	Serious (contact sexual offences)	No[8]	Yes	No	Men with sexual history and OASys data (completed up to March 2008)
OSP/I	Serious (indecent images)	No	Yes	No	Men with sexual history and OASys data (completed up to March 2008)
Total RSR	All serious (RSR definition)	Yes	Yes	No	As with RSR SNSV, OSP/C and OSP/I

[8] An evidence-based 5-year risk reduction rule is in use by HMPPS, but not an official part of OSP/C scoring

Predictors by offending outcome

All (or general) reoffending including nonviolent offending

OGRS3 was an update to OGRS2 which was launched in 2000. OGRS2 was the first to be introduced to the then-new computerised Offender Assessment System (OASys). The first generation of OGRS developed in the 1990s and was scored by hand.

OGRS4/G was developed as another general reoffending predictor in 2009 and previous research has shown it to statistically outperform OGRS3. It is not in use operationally due to resource pressures at the time of inception. OGRS4/G includes an 'offence-free time' element, recognising that an offender's probability of future proven reoffending falls with time after community sentence or discharge from custody without yet reoffending. The predictor thus allows a more accurate comparison of offenders at different stages of community supervision, assisting with the targeting of supervision and treatment resources.

OGP2 was an update to OGP1 (Howard, 2009b), both of which rely on static and dynamic risk factors (see Static and dynamic risk factors, below). OGP1 was designed to predict nonsexual, nonviolent reoffending only, as opposed to OGP2 which was designed for all (general) reoffending.

OGP1 is the only predictor designed specifically for nonviolent reoffending. No HMPPS business processes (e.g., targeting rehabilitative interventions; Risk of Serious Harm guidance) utilise nonviolent reoffending risk, and therefore OGP1 has limited practice value. As such, it is out of scope for this study. Both OGRS3 and OGRS4/G are calculated based on static factors only and can be calculated for everyone in the study, whereas the static/dynamic predictors OGP1 and OGP2 require a full OASys assessment.

Broad violent reoffending

OVP1 is the only predictor designed for broad violent reoffending in use operationally. OVP2 was developed in 2009 on a more recent cohort and accounts for offence-free time. Both versions of OVP include static and dynamic risk factors, meaning they rely on a full OASys assessment to calculate.

OGRS4/V was developed at the same time as OVP2 and provides a predictor for violent reoffending which can be calculated using static risk factors only. However, as with OGRS4/G, OGRS4/V is not in use operationally.

Serious nonsexual violent (SNSV) reoffending

The RSR SNSV predictor was developed in 2009 to focus specifically on the most serious forms of violence. Along with OSP/C and OSP/I it forms part of the total risk of serious recidivism (RSR). It can be calculated based on static risk factors or static and dynamic risk factors, where available, and accounts for offence-free time.

Serious contact sexual reoffending

OSP/C is the only predictor in operational use for contact sexual reoffending and replaced the paper-based Risk Matrix 2000 as a predictor of sexual reoffending. OSP/C is based on static risk factors only. Due to the small number of women with a sexual offending history, and their low reoffending rates, OSP/C is calculated for men with a sexual offending history only.

Serious indecent images reoffending

Developed alongside OSP/C, OSP/I is the only predictor in operational use for the prediction of reoffending involving indecent images of children (IIOC). As with OSP/C, it is only calculated for men with a history of sexual offending.

Static and dynamic risk factors

All risk predictors use **static risk factors** (i.e. those which cannot be deliberately changed over the course of a sentence), such as 'age at risk',[9] 'Offence category'[10] and gender (noting that sexual predictors are for men with sexual offending history only). Static risk factors typically have the strongest association with proven reoffending.

The five risk predictors that are based solely on static risk factors can be calculated for all offenders, irrespective of whether they have a full OASys assessment (see "OASys (Layer 3) assessments" in Appendix C).

[9] For assessment purposes, the age of the individual is set at the point which they entered or will enter the community

[10] Categorisation of offences varies between risk predictors

Several predictors include **dynamic risk factors** (also referred to as criminogenic needs). That is, factors which change as the offender passes through their rehabilitative journey of the criminal justice system. Examples include: Accommodation, Employment, Alcohol Misuse, Emotional Well-being, Thinking and Behaviour, and Attitudes. Information on criminogenic needs is only collected as part of a detailed OASys layer 3 assessment. Predictors that require dynamic risk factors can therefore only be calculated for the subset of offenders with a complete OASys. These predictors are said to be based on static/dynamic risk factors.

Offence-free time

In addition to the division between static and static/dynamic assessment tools, the newer generation of tools – OGRS4, OGP2, OVP2 and RSR – incorporate offence-free time (or offence-free months, OFM). That is, the length of time (in months) that the offender has been continuously on the community caseload (without recall or reoffending). Offence-free time becomes a factor in those predictor scores, **causing them to fall** as each offence-free month passes.

For example, someone who has been offence-free through to 15 months of the community portion of their sentence has a lower predicted RSR than an identical person just released from custody.

5.2 Analytical Datasets

Two different study populations (or cohorts) were used in this study, and are described in detail in Appendix C:

- Starts: A population of individuals who are newly sentenced, derived over a six-month period from 1 July 2018 to 31 December 2018; and
- Caseload: A snapshot of the probation caseload, as at 30 June 2018.

The two cohorts were then matched with data from the Police National Computer (PNC) and the Offender Assessment System (OASys) to derive a full profile of previous and subsequent offending (from the PNC data) and risk and criminogenic needs (OASys).

Starts and caseload

The population of starts broadly reflects individuals who have 'just entered the community' – and have therefore all been offence-free for zero months. The starts population generally reflects how the predictors are used in practice: assessing an individual on or before they are sentenced/enter the community. Therefore, the population of 'starts' is the primary focus of this report in assessing predictor performance.

The analysis is replicated for the caseload (a snapshot of the probation caseload), all with varied offence-free time. These analyses provided insight into the performance of the predictors, depending on how long an individual had been in the community offence-free and whether predictors account for offence-free time (calculated in full months) in their calculation of risk. This is especially important for cohorts such as sexual offenders who may spend a long time on the caseload.

Offending Outcomes

An offending outcome is shorthand for 'proven reoffending involving an offence of interest being committed within the specified time period'. The PNC extract was used to identify **proven reoffences** that were committed in the subsequent two-year follow-up period which led to a recorded conviction, simple caution or Conditional Caution by the date the PNC extract was taken, on 2 November 2022 (i.e. allowing a further waiting period of about two years for conviction etc. to occur).[11]

For each offence of interest (see Appendix B) we identify the number of days from the start of the study up to a period of 731 days (a two year follow up which includes a leap day). It was recognised that a complete two-year follow-up is often not possible, such as if the offender was recalled to custody.

For example, when studying serious reoffending (predicted by RSR), an offender might be imprisoned for a nonserious offence during the two-year period. Using the language of 'survival analysis', these cases were labelled censored. Censoring is a form of missing

[11] It is not always clear what waiting period rules were applied when the predictors were originally generated, but the production of RSR involved tracing the caseload of 31 March 2010 using a PNC extract taken in spring 2013, and a one-year waiting period would replicate this closely.

data in our study and the approach used for measuring a predictor's discriminative validity accounts for censoring to avoid total loss of these cases' data.

Therefore, for each reoffending outcome, every offender was categorised by the earliest of one of the following outcomes:

- Reoffended;
- Sentenced to immediate custody (either for an offence not of interest or for any offence committed before their study 'start' date)
- Standard recall to custody[12]
- None of the above, for the full two years of the follow-up period.

An offender may have multiple reoffences of different types during the follow up period (e.g. OVP-type violence on day 100 and an RSR offence on day 150). Each reoffence was considered separately.

5.3 Model Evaluation

The primary research questions were focussed on two areas:

1. How well the models are calibrated – measured by comparing the **actual rate** of reoffending with the mean **predicted rate**; and
2. How well an individual predictor discriminates between higher- and lower-risk offenders for specific outcomes, measured via Harrell's C-Index (Harrell, Lee & Mark, 1996).

Model calibration

Model calibration is one aspect of model accuracy that was evaluated in this study. In this context, calibration relates to the accuracy of the risk scores (the probabilities or percentages) that each risk instrument produces. If the risk score is used to estimate the rate of reoffending (or forecast the volume of reoffenders) then these estimates will be 'out' by the same amount as any miscalibration observed, assuming that 2018 to 2020 patterns of reoffending persist. A poorly calibrated model can still discriminate well.

[12] Those subject to standard recall (including emergency standard recall) can be imprisoned until the sentence end date, which may be a substantial duration. Fixed-term recalls, which last two or four weeks only, were considered to cause insufficient disruption to the follow-up to be counted as censoring events.

Model calibration is measured through the difference between the observed (actual) rate of reoffending and the mean predicted rate. The difference between these (i.e. the actual rate minus the predicted rate) is referred to as the residual, and residuals are reported as 'percentage point differences',[13] or in 'points'. A well calibrated model would effectively estimate the rate of reoffending it was designed to predict. Reference is made to 'statistical significance' when comparing the rates of reoffending, allowing us to understand whether the predicted rate of reoffending differs significantly from the observed (actual) rate (see Appendix C). In addition to statistical significance, a measure of 'effect size' is reported for the difference in those proportions. Effect sizes are a quantitative measure of the difference between two measurements and help identify differences of practical importance. Here, odds ratios were adopted. Odds ratios are a continuous measure. Therefore, to gauge the calibration of the predictors thresholds of 25, 50, 75 and 100 per cent change in odds were selected to describe small, medium, large and very large effects respectively (see Appendix C). Any difference below the small threshold was described as negligible. Odds ratios between predicted and actual rates of reoffending were compared to those representing the above thresholds in either direction and together with statistical significance are used to determine the calibration of a predictor (Table 4).

In the results section, when reporting odds ratios (effect sizes) for multiple predictors or subgroups the result closest to the next threshold was reported to give an indication of the largest/most extreme differences. For example, if the overall calibration results for three predictors being evaluated are 0.80, 0.78 and 1.12 (all negligible) then the result of 0.78 would be the one highlighted. This result would be the odds ratio which is furthest from 1.00 and closest to the threshold for a small effect size.

Table 4: Reporting convention for calibration

Statistical significance (at 5 per cent level)[14]	Odds ratios (lower threshold; upper threshold)	Reporting convention
Not statistically significant	Any	Well calibrated
Statistically significant	0.801 – 1.249 (negligible)	Well calibrated

[13] That is, the difference between 'three per cent actual reoffending' and 'two per cent predicted reoffending' is one percentage point (or one point).

[14] Significant at 5 per cent means that there's a 5 per cent probability of rejecting the null hypothesis (i.e. that there is a difference) when the null hypothesis is true (i.e. that there is no difference)

Statistical significance (at 5 per cent level)[14]	Odds ratios (lower threshold; upper threshold)	Reporting convention
Statistically significant	0.667 – 0.800; 1.250 – 1.499 (small)	Small miscalibration
Statistically significant	0.572 – 0.666; 1.500 – 1.749 (medium)	Moderate miscalibration
Statistically significant	0.499 – 0.571; 1.750 – 1.999 (large)	Large miscalibration
Statistically significant	≤ 0.500; ≥ 2.000 (very large)	Very large miscalibration

Model calibration, all women and men with no sexual history

For the serious reoffending outcomes, a complication comes from the fact that sexual reoffending risk is only calculated for men with known sexual offending history. Results for OSP/C and OSP/I in this study will underestimate sexual reoffending risk for all women and any man without a sexual history, where it is known that a small volume of women and men without history do reoffend. However, as mentioned in section 4.5 for women with a sexual history, a baseline rate of 1/193 is added to the total risk of serious recidivism (that is, total RSR) score.

Discriminative Validity: Harrell's C-Index

Discriminative validity is the risk predictor's ability to successfully distinguish higher- from lower-risk offenders (Table 5). Harrell's C-Index, also known as the Concordance Index,[15] is the discriminative validity metric used in combination with survival analytic methods such as the selection of reoffending outcome described above. Unlike other model performance measures the C-Index can account for individuals being removed partway through the study. For example, if someone is imprisoned for burglary after six months, and had no violent reoffences prior to that, their violent reoffending follow-up would read "no violence, censored at six months", and they can be compared with people who did reoffend violently within six months though not those who did so at a later point.

[15] Many studies of predictive validity use the area under the receiver-operator characteristic curve (AUC) discriminative predictive validity metric. The C-Index can be interpreted in the same way as the AUC, and in fact the AUC is a special case of the C-Index, where all subjects have the same length of follow-up (that is, survival methods are not used).

Further information on C-Indices, why they may vary and our chosen reporting conventions is available in Appendix C.

Table 5: Reporting convention for discrimination

Harrell's C-Index score	Reporting convention
Less than 0.556	Poor discrimination
0.556 to 0.638	Moderate discrimination
0.639 to 0.713	Acceptable discrimination
0.714 to 0.784	Good discrimination
Greater than 0.785	Excellent discrimination

5.4 Limitations

There are some limitations to the analytical approach.

Firstly, this study only accounts for proven reoffending during the follow-up period. Any offending that is either not reported or convicted are excluded from this analysis.

Next, the use of a follow-up period of two years will only give a partial picture of the reoffending habitats of offenders, any reoffences which occur after the follow up period has ended will not be included in the study. However, previous studies show that individual risk of reoffending is highest within those two years.

Finally, the choice of snapshot date was guided in part by the COVID-19 pandemic. To ensure that the assessment of actuarial tools pertained to normal reoffending patterns data from 2018 to 2020 was used. This meant the most up to date data was not included but ensured that changes to offending patterns and conviction rates caused by lockdowns and court backlogs did not produce spurious conclusions regarding the effectiveness of the suite of risk predictors.

6. Results

6.1 Descriptive statistics of population and missing data

Descriptive statistics

There were 81,258 offenders that entered the community[16] in the second half of 2018 (the starts; Table A1). A large majority studied were male (86.42 per cent), more than half (54.99 per cent) had an OASys assessment and one in three (32.50 per cent) were aged 30 to 39. Three in every four starts (73.63 per cent) were White, but a large proportion (9.77 per cent) missing data on ethnicity; a further 6.88 per cent were Black and 5.23 per cent were Asian. The distribution of ethnicities was broadly similar for those with an OASys, and when gender and sexual history were considered.

Overall, one in six of all starts (17.15 per cent) were screened with likely learning disability and challenges (LDC), this was higher when only considering those with an OASys (31.18 per cent) and men with a sexual offending history (26.89 per cent). A large proportion (60.27 per cent) of all starts had a disability and, of those with an OASys, 45.82 per cent were a current or former domestic violence perpetrator.

Missing data (censoring)

The concept of censoring is introduced in section 5.1, above. Censoring is a form of missing data where an individual is removed from the study (and cannot be followed up for the full two-year period). The amount of censoring (and the average time until the censor date; the date the individual 'leaves' the study) may vary by offending outcome (e.g. any offending and OVP-type offending can have different amounts of censoring). As described above, calculations of the C index (discriminative validity) can account for censoring so the individual can remain in the study.

For any reoffending one in ten (9.62 per cent) starts were censored and the median number of days[17] for any reoffending was 31 days (Table A2). That is, of the 9.62 per cent of individuals that had any censoring (for any reoffending) half were censored (or removed

[16] Includes community orders, suspended sentences and people released on licence from custody

[17] The median is the number of days that half of the censored cases were removed from the study

from the study) at 31 days. There was little variation across subgroups that were analysed but people screened with likely LDC and former domestic violence (DV) perpetrators had higher rates of censoring with 17.25 and 15.61 per cent, respectively.

Rates of censoring were higher when the offending outcome was more specific (e.g. contact sexual or OVP-type) from 14.48 per cent for broad violent (OVP-type) reoffending up to 19.76 per cent for serious reoffending (either contact sexual, indecent images or serious nonsexual violence, SNSV) (Table A2 and Table A3). The median days until censor were higher for serious reoffending (around 72 days). This pattern of higher rates of censoring and median days until censor for specific offending is a consequence of a narrower offence definition. Specific offences are, by nature of not including all reoffending, rarer. Therefore, there is a larger window within which a censoring event can occur leading to higher median days to a censoring event and higher rates of censoring.

6.2 Proven rates of reoffending

Any (all) offending and OVP-type violence

Overall, for the population of starts (who all have had zero offence-free time), the two-year proven reoffending rate was at 45.07 per cent (Table A4). Rates of reoffending fell in the older age bands (those aged fifty and over), was lower for females (38.55 per cent), and higher for people screened with likely LDC (learning and disability challenges; 68.14 per cent) and former DV perpetrators[18] (65.29 per cent).

OVP-type (broad violent) reoffending was the most common subtype with over one in four (27.63 per cent) starts having a proven violent reoffence within two years.

Rates of reoffending were higher for people with an OASys assessment, overall 55.95 per cent of the starts population with an OASys had reoffended over the two-year study period (Table A5). Higher rates of reoffending for people with an OASys reflects the fact that individuals with more prolific or serious reoffending history are more likely to have had a full OASys assessment and subsequently have a higher risk of reoffending overall.

[18] The LDC screening and DV perpetration were only available with a full OASys assessment.

Broad patterns of reoffending rates for all and OVP-type reoffending were similar across all subgroups, including when accounting for those with and without an OASys.

Serious (RSR) reoffending

There are three subtypes of serious reoffending (RSR-type reoffending) that fall under the definition of the 'risk of serious recidivism' (RSR) tool:

- Serious nonsexual violence or SNSV (accounts for most of RSR-type offending) (Table B2);
- Contact sexual offending (Table B3); and
- Sexual offending involving indecent images (Table B4).

Overall, the rate of all proven RSR reoffences was 2.13 per cent. Of all RSR reoffending, serious nonsexual violence (SNSV) is the most common with 1.62 per cent of the starts having a proven reoffence for SNSV within two years. Rates of sexual reoffending were lower; contact sexual reoffending (0.40 per cent) and indecent images (0.13 per cent). The total RSR reoffending rate (of 2.13 per cent) is lower than the sum of SNSV, contact sexual and indecent images, this indicates that a small number of people have proven reoffences of multiple types (for example, contact sexual and SNSV) – though rare (Table A5).

A steep fall in total serious reoffending was observed by age. Of those aged 18 to 20, 3.97 per cent had any serious (RSR-type) reoffences compared to 0.95 per cent of those aged 60 and over.

Sexual offending

Table A6 provides the rates of sexual reoffending (contact and indecent images) for men with a sexual offending history (note the small case numbers for subgroups). As described in section 4.5, Population subgroups, above, the sexual predictors were designed for the prediction of sexual reoffending for men with a sexual history. As such, it was of particular interest to study rates of reoffending for this group. Additionally, Table A7 provides rates of sexual reoffending by gender, sexual history and a combination of age or risk band (where calculated).

Overall rates of contact sexual and offences involving indecent images were higher for men with a sexual history (2.16 per cent and 1.50 per cent respectively) when compared to the total population of starts (0.40 per cent and 0.13 per cent respectively).

Rates of **contact sexual reoffending** increased in line with the associated **OSP/C** risk band (0.80 per cent with low OSP/C scores increasing to 8.33 per cent for those with very high OSP/C). The same pattern was observed for rates of proven reoffending for **indecent images** and the associated **OSP/I** risk band.

There is a relatively high rate (albeit low number of cases) of two-year proven reoffending for **contact sexual offences** in the high **OSP/I risk band** (1.72 per cent; Table A7**)**. This contrasts past research (Howard, Barnett & Mann, 2015), indecent image specialists do not go on to contact sexual offending. As a topic for further research it would be interesting to understand how many of those in the high OSP/I band had any history of non-indecent image sexual offences.

Women and men without sexual history

There were a small number of women with no sexual history (7 out of 10,963) who went on to commit a contact sexual reoffence over the two-year study period and no women had a proven indecent images reoffence. For men without a sexual history, rates of contact sexual reoffending broadly fell with age[19] with 0.38 per cent of men with no sexual history aged 18 to 20 (20 out of 5,206 men) down to 0.08 per cent of men with no history aged 60 and over (1 man out of 1,280) having a contact sexual reoffence. For men without sexual history, indecent images reoffences were most common in the 18 to 20 age group (0.13 per cent or 3 in 8,715 men) and rates were flat in the older age groups (maximum, 0.03 per cent).

[19] With a peak of 0.45 per cent in those aged 25 to 29

6.3 Accuracy in the prediction of all (or general) reoffending

Due to its more onerous data requirements, needing dynamic factors that necessitate a full OASys assessment, OGP2 cannot be calculated for all individuals. Therefore, this section will discuss the performance of predictors in two parts:

- All individuals' – comparing OGRS3 and OGRS4/G; and
- 'With an OASys assessment' – comparing OGRS3, OGRS4/G, and OGP2.

Summary of findings

- All predictors designed for all (or general) reoffending were overall well calibrated. Residuals had negligible effect sizes indicating minimal difference between predicted and actual rates of offending;
- OGRS3 and OGRS4/G both had good discrimination (for all cases) and discrimination was acceptable when assessing those with an OASys only;
- OGP2 had good discrimination and performed marginally better than the two static predictors (for those with an OASys); and
- OGRS4/G and OGP2 performed better than OGRS3 when assessing the full probation caseload due to their ability to account for offence-free time.

Model calibration

In absolute terms, two of the three predictors designed for 'any proven reoffending' had over-predicted the rate of reoffending (Table A8), these results are statistically significant,[20] however, the largest effect size was found to be negligible[21] indicating that differences were minimal (i.e. the tools are well calibrated).

All individuals

OGRS3 is the only predictor in operational use capable of calculating reoffending rates for all individuals and was found to, in absolute terms, over-predict the overall rate of reoffending by 3.82 points (Table A8). OGRS4/G, the closest comparable predictor,

[20] Two-tailed tests, $p<0.0001$

[21] OGRS4/G odds ratio: 0.822

also over-predicted but to a greater degree (4.90 points). The largest effect size of which was found to be below the threshold for a small effect size.[22]

When assessing across risk bands OGRS3 continued to over-predict the actual rate of reoffending, with the largest residual in the high-risk band (6.25 points), although OGRS4/G had higher residuals in the lower risk bands (Table A9). Across risk bands, with one exception,[23] all differences were statistically significant. Small miscalibration was observed in the OGRS3 high and prolific risk bands and the OGRS4/G medium and high bands[24] and a moderate miscalibration in the OGRS4/G low risk band.[25]

With an OASys assessment

Considering only those individuals with an OASys assessment, OGP2 was found to under-predict reoffending by 1.52 points (Table A8), OGRS3 was marginally better calibrated, over-predicting by 1.05 points, despite not using dynamic factors. Results for OGP2, OGRS3 and OGRS4/G were statistically significant but below the threshold for a small effect size,[26] leading to the conclusion that the predictors are well calibrated.

Across lower risk bands OGP2 was slightly better calibrated than OGRS3 (Table A9). However, OGP2 tended to under-predict while OGRS3 over-predicted, a pattern that reversed for low-risk individuals. All results were below the threshold for a small effect size.[27]

OGP2 and OGRS4/G varied in their performance across subgroups with no clear pattern (Table A17 and Table A19). Generally, OGRS4/G calibration improves (residuals become smaller) as age increases, while OGP2 worsens. Across ethnicity groups OGRS4/G performs best, except for White individuals, where OGP2 had smaller residuals. Across all subgroups, except one,[28] where results were statistically significant, effect sizes were

[22] OGRS4/G, all cases, odds ratio 0.822
[23] OGRS4/G, prolific risk band, Z = -1.330, p = 0.184
[24] Largest effect size, OGRS4/G medium risk band with odds ratio of 0.751
[25] Odds ratio of 0.658
[26] Largest effect size, OGRS4/G with odds ratio of 0.934
[27] Largest effect size, OGP2, with OASys, very high and prolific risk bands with odds ratios of 1.243
[28] OGRS3, with OASys, 50 – 59 age band, odds ratio 1.327

negligible, leading to the conclusion that the predictors were well calibrated across subgroups.[29]

Model discrimination

All individuals

Considering static predictors (those which are available for the entire population), OGRS3 has virtually equal discriminative validity as measured by Harrell's C-index to OGRS4/G (0.736 and 0.737, respectively; Table A20). This would be expected as the main factor differentiating the two predictors is the inclusion of offence-free time for OGRS4/G (which is not a factor in the population of starts).

With an OASys assessment

When an OASys assessment was present OGRS3 was marginally outperformed by the static/dynamic predictor OGP2 but had equal discriminative validity to OGRS4/G with C-indices of 0.711, 0.717 and 0.711 respectively. A similar pattern was noted through all subgroups analysed (Table A20).

General observations across subgroups that were analysed were:

- Increasing discrimination as age increased;
- Better discrimination in females than for males;
- Poorer discrimination in Black and Mixed ethnicities compared to White and Asian ethnicities; and
- Poorer discrimination in those screened with likely LDC.

Offence-free time

When investigating how the predictors perform on the caseload data[30] the value of accounting for offence-free time becomes apparent. For instance, across risk bands all predictors were found to over-predict the actual rate of all reoffending (Table A33). However, OGRS3 (which does not account for offence-free time) reported substantial differences, with over-predictions of 13.55, 17.11 and 13.40 points in medium, high and very high risk bands for all cases respectively. A similar pattern of performance was noted

[29] Largest effect size, OGRS3, with OASys, 60 and over age band, odds ratio 1.161

[30] Who have varied offence-free time as opposed to the starts who have all been offence-free for zero months

for those with OASys. Meanwhile, OGP2 and OGRS4/G, which both account for offence-free time, over-predicted by a maximum of 4.48 and 7.06 points respectively, where OGP2 had better calibration overall.

Residuals were generally seen to increase as age and offence-free time increased. The rates of proven reoffending fell as both age and offence-free time increased (Table A37) therefore, OGRS3 and OGRS4/G predictors were not well calibrated for the falling rates. In other words – the predicted rates fell at a lower rate than the actual rates fell, leading to larger residuals.

The influence of offence-free time can also be seen in the model discriminations. OGRS3 had lower overall discrimination in the caseload when compared to the two predictors which account for offence-free time; OGP2 and OGRS4/G. As might be expected, given it does not account for it, OGRS3 discrimination falls as offence-free time increases. OGP2 remained relatively flat as the length of time without an offence increased until that time reached 12 months or more, at which point discrimination increased. The discrimination of OGRS4/G displayed a similar relationship to OGP2 as offence-free time increased when considering individuals with an OASys. OGRS4/G's discrimination declined as offence-free time increased when evaluating all individuals, though not as steep a decline as with OGRS3 (Table A44).

6.4 Accuracy in the prediction of broad (OVP-type) violent reoffending

Of the predictors of OVP-type violence OVP1 is the sole predictor in operational use. Both OVP1 and OVP2 require dynamic factors and can only be produced for those individuals with a full OASys assessment. OGRS4/V is the only static predictor and does not require an OASys assessment but is not in operational use. However, the comparison of static versus dynamic predictors could yield interesting insights. Therefore, this section will discuss the performance of predictors in two parts:

- 'All individuals' – describing the performance of OGRS4/V; and
- 'With an OASys assessment' - comparing OVP1, OVP2 and OGRS4/V.

Summary of findings

- Overall, OVP1, OVP2 were well calibrated while OGRS4/V had small miscalibration;
- By risk band, OGRS4/V had large miscalibration in the prolific risk band (those with predicted rates between 90 and 100 per cent) whereas OVP1 and OVP2 were well calibrated in all risk bands;
- OVP2 had the best discrimination overall for those with an OASys assessment; and
- OVP1 had poorer calibration than both OGRS4/V and OVP2 when offence-free time was considered.

Model Calibration

In absolute terms, all of the predictors of OVP-type violence over predicted the rates of reoffending (Table A8) though effect sizes were negligible.

All individuals

Overall OGRS4/V had a small miscalibration. OGRS4/V was found to over-predict reoffending of OVP-type violence by 5.06 points – a statistically significant difference with a small effect size.[31] When assessing across risk bands (Table A10), OGRS4/V had small miscalibration in all bands except the prolific risk band with a residual of 4.86 points (a large miscalibration[32]). In other risk bands residuals ranged from 2.96 points to 6.61 points in the low and high bands respectively.[33]

With an OASys assessment

OVP1 and OVP2, which both consider static/dynamic factors, performed better than OGRS4/V when considering individuals with an OASys assessment and both were well calibrated overall: OVP1 and OVP2's residuals were not statistically significant and odds ratios were negligible.[34] OVP1 and OVP2 were well calibrated across all population

[31] Two-tailed test, z = -29.906, p < 0.0001; odds ratio 0.789
[32] Odds ratio of 0.551
[33] Small miscalibration, largest effect size: OGRS4/V, all cases, low risk band, odds ratio 0.741
[34] Largest effect size, OVP2, with OASys, odds ratio 0.986

subgroups[35] and had smaller absolute residuals across most subgroups when compared to OGRS4/V (Table A17 and Table A19).

Turning to predictor performance across risk bands OVP1 and OVP2 performed better than OGRS4/V at each risk level (maximum residuals of 3.73, 2.20 and 4.68 points respectively; Table A10). OVP1 and OVP2 displayed similar trends in performance across risk bands (larger residuals as risk increase) but OVP2 had smaller absolute residuals overall and was very well calibrated for lower risk bands.

Model discrimination

Considering static predictors (available for the entire population), OGRS4/V displayed good discrimination with an overall index of 0.742 (Table A21).

For those, that do have an OASys assessment, OVP2 had the best discrimination overall and across all subgroups when compared to the static/dynamic predictor OVP1 and the static predictor OGRS4/V (Table A21). Similar patterns of discrimination, such as increasing with age and better discrimination for females, were observed across subgroups for all three predictors.

Offence-free time

The caseload data highlights the value of accounting for offence-free time when assessing individuals partway through their sentence. In individuals with an OASys assessment all predictors had (in absolute terms) over-predicted reoffending, but OVP1 (which does not account for offence-free time) did so to a greater degree – 3.03, 5.31 and 8.11 points for OVP2, OGRS4/V and OVP1, respectively (Table A32). OVP2 was well calibrated (negligible effect size), OGRS4/V had a small miscalibration and OVP1 had a moderate miscalibration.[36] A similar effect can be seen across risk bands, where OVP1 generally performed worse except in very high-risk individuals where OGRS4/V was least calibrated (OVP1 performed similarly).

The influence of offence-free time was also noted in the C-indices (discrimination) of the models. Whilst OVP1 had lower discrimination when offence-free time was not a factor

[35] Largest effect size, OVP1, with OASys, female, odds ratio 1.217

[36] Largest effect size, OVP1, with OASys, odds ratio 0.648

(Table A21) the difference in discrimination between OVP1 and the other two predictors was more pronounced when offence-free time was a factor (Table A45).

6.5 Accuracy in the prediction of serious nonsexual violent (SNSV) reoffending

There is only one predictor developed specifically for serious nonsexual violence (SNSV), the RSR SNSV predictor. RSR SNSV comes in two forms; static, which is available for everyone in the study; and static/dynamic, which is only available to those individuals with an OASys assessment.

In this section the static and static/dynamic versions of RSR SNSV will be compared, and OGRS4/V where appropriate.

Summary of findings

- RSR SNSV static predictor was well calibrated for all individuals and those with an OASys assessment with residuals of less than 1 point in most cases;
- Both static and static/dynamic version of RSR SNSV were well calibrated for individuals with an OASys, across most risk bands and for most subgroups analysed; and
- Both static and static/dynamic versions of RSR SNSV demonstrated good overall discrimination.

Model calibration

Overall, RSR SNSV static and static/dynamic predictors were well calibrated.

All individuals

Overall, the RSR SNSV static predictor (which can be calculated for the entire population) was well calibrated. RSR SNSV had (in absolute terms) under-predicted the actual rate of reoffending by less than a tenth of a point (Table A8). This was not statistically significant and had a particularly small effect size – indicating that the difference between actual and predicted rates of reoffending were minimal.[37] RSR SNSV static was found to closely predict the rate of reoffending across most subgroups analysed, with small miscalibration

[37] Two-tailed test, z = 1.063, p = 0.288; odds ratio 1.030

for those in the 18 – 20 age band (under by 0.84 points) and females (over by 0.28 points)[38] and a moderate miscalibration for Black and Mixed ethnicities under by 0.96 and 1.22 points respectively.[39]

RSR SNSV static was also well calibrated when considering risk bands, producing residuals of around half a point (i.e. less than one percentage point). Statistical analysis found that the predicted rate of reoffending did not differ significantly from the actual rate of reoffending in any of these instances and had negligible effect sizes.

With an OASys assessment

The RSR SNSV static/dynamic predictor was well calibrated for individuals with an OASys assessment with a residual of 0.23 points under (Table A8). This was found to be statistically significant but had a negligible effect size.[40] The static RSR SNSV predictor was similarly calibrated for those with an OASys, under-predicting by 0.19 points.

Studying across subgroups, both static/dynamic and static versions of RSR SNSV were found to be generally well calibrated with miscalibration in the same subgroups RSR SNSV static had on all cases, i.e. Black and Mixed ethnicities, 18 – 20 age band and females (Table A14).

Studying across risk bands, static/dynamic and static versions of RSR SNSV were found to be generally well calibrated. The only miscalibration was observed for medium risk band in the static/dynamic predictor (under by 0.46 points; Table A12).[41]

Model discrimination

All individuals

Overall, for the population of starts, the static version of the RSR SNSV predictor had best discrimination (0.761) when compared to the other static predictor (designed for broad violent reoffending) OGRS4/V (0.743; Table A22). RSR SNSV static performed similarly to, although usually marginally better than, OGRS4/V across all subgroups, with the

[38] Largest effect size, 18 – 20 age band, odds ratio 1.333

[39] Largest effect size, Mixed ethnicity, odds ratio 1.606

[40] Two-tailed test, $z = 3.459$, $p < 0.001$; odds ratio 1.121

[41] Odds ratio 1.264

exception of people aged 60 and over and Black ethnicity groups where RSR SNSV static had substantially better discrimination (Table A22).

With an OASys assessment

When we consider the subset of the population with an OASys, we can compare discrimination with static/dynamic predictors. Here the RSR SNSV static predictor was outperformed by its extended static/dynamic counterpart – C-indices of 0.737 and 0.751 respectively. Overall, both RSR SNSV predictors demonstrated good (or better) levels of discrimination across most subgroups (Table A22). However, there were a few exceptions. For instance, for the static/dynamic version of RSR SNSV, those aged 60 and over had moderate discrimination (0.637) and for those in some of the younger age bands (18 – 20, 21 – 24 and 30 – 39), females and those with Black ethnicity, discrimination was acceptable (ranging from 0.697 to 0.713).

6.6 Accuracy in the prediction of contact sexual reoffending

Only one predictor is available for the prediction of contact sexual reoffending, OSP/C, and a standalone risk score is only calculated for men with a sexual offending history.[42] Furthermore, OSP/C is a static predictor and does not rely on an OASys assessment. Therefore, unless specified, results in this section relate to men with a sexual offending history only and will not be discussed in terms of all individuals and those with an OASys assessment as in other sections.

Summary of findings

- OSP/C was found have a very large miscalibration, underpredicting the rate of contact sexual offending both overall and across most subgroups;
- OSP/C exhibited acceptable discrimination, however, the overall C-index was below that in the published research; and
- Discrimination was poorer in some subgroups, with results based on small number of cases with few proven reoffences.

[42] A baseline rate is added to any woman with a sexual offending history's total RSR score, see Men with sexual offending history

Model calibration

Overall, OSP/C under-predicted the rate of reoffending by 1.11 points (Table A8), this was found to be statistically significant and was a very large effect size,[43] leading to the conclusion that OSP/C is not well calibrated. A similar result was observed across most subgroups that were analysed with the largest differences in those in the 18 – 20 age band[44] and Black ethnicity[45] (with under predictions of 3.81 and 2.74 points respectively; Table A16). For some subgroups, residuals were not statistically significant, including Asian and Mixed ethnicities, 21 – 24 age band and current DV perpetrators.

OSP/C also under-predicted reoffending across risk bands. The largest absolute residual was observed in the highest risk band with a residual of 4.29 points with the largest effect size observed in the low risk band (0.49 points under[46]). A moderate effect size was observed in the high risk band;[47] all others were very large.

A standalone OSP/C score is not calculated for **all women** or **any man without sexual offending history** so results as shown in this study will, by definition, show an under-prediction (the actual rate of reoffending for all women and any man with no sexual history was 0.27 per cent; Table A8). As mentioned in section 4.5, a baseline rate of reoffending is added to a woman's total RSR score (if she has a sexual history).

Model discrimination

For contact sexual offending, overall, OSP/C had acceptable discrimination with a C-index of 0.676 (Table A23). However, this was notably lower than the 0.763[48] published in 2021 research.[49] In some subgroups the C-index was lower still (Table A23), such as 18 to 20 (0.365) and 30 to 39 (0.578) age bands, and Black (0.598) and Mixed (0.356) ethnicities. In contrast OSP/C exhibited excellent discrimination where ethnicity was not recorded (0.930) and in the 60 and over (0.814) subgroups.

---

43 Two-tailed test, z = 8.409, p-value < 0.0001; odds ratio 2.085

44 Odds ratio 3.760

45 Odds ratio 3.249

46 Odds ratio 2.618

47 Odds ratio 1.662

48 Based on the 64-point risk band rather than the continuous percentage score

49 See Howard, P., & Wakeling, H. Comparing two predictors of sexual recidivism: the Risk Matrix 2000 and the OASys Sexual Reoffending Predictor. *Ministry of Justice Analytical Report*.

Caution should be exercised when interpreting results for subgroups of a small cohort (men with a sexual offending history) for a rare offence type (contact sexual offending), see Appendix C for more information. Results may be subject to large variation between studies. For example, there were (of men with sexual offending history):

- 190 offenders aged 18 to 20, of which ten had a proven reoffence;
- 153 offenders with Mixed ethnicity, of which three had a proven reoffence; and
- 175 men where ethnicity was not recorded, of which two had a proven reoffence.

6.7 Accuracy in the prediction of reoffending involving indecent images of children

As with OSP/C, OSP/I is a standalone risk score calculated for men with a sexual offending history only and does not rely on an OASys assessment. Therefore, as above, unless specified otherwise results in this section relate to men with a sexual offending history only and will not be discussed in terms of all individuals and those with an OASys assessment.

Summary of findings

- OSP/I was not well calibrated overall and across most subgroups;
- OSP/I was well calibrated in the medium risk band, but had a large or very large miscalibration in the low and high bands; and
- OSP/I displayed good to excellent discrimination overall and across subgroups.

Model Calibration

Overall, OSP/I had a residual of 0.66 points under (Table A8) this was statistically significant and had a large effect size[50] leading to the conclusion that OSP/I was not well calibrated overall.

Across subgroups (Table A16) OSP/I tended to under-predict reoffending with some subgroups found to be well calibrated where residuals were not statistically significant including:

- Asian, Black and Mixed ethnicities (all three groups had no proven reoffences with predicted rates of 0.38, 0.29 and 0.38 per cent respectively);

[50] Two-tailed test: z = 5.603, p-value < 0.0001; odds ratio 1.804

- Those in the 21 – 24, 40 – 49 and 60 and over age bands (0.53, 0.05 and 0.08 points respectively); and
- Current or former DV perpetrators (0.01 and 0.13 points respectively).

For all other subgroups OSP/I was found to be miscalibrated with the largest differences in those aged 18 – 20, people where ethnicity was not recorded and likely LDC with 2.07, 1.97 and 0.82 points under respectively.[51]

Studying across risk bands, OSP/I had a very large miscalibration in the high risk band with a residual of 9.66 points under[52] but OSP/I was found to be well calibrated in the medium risk band where the residual was not statistically significant[53] (Table A11).

Model discrimination

For indecent images, overall, OSP/I had excellent discrimination at 0.857. Some of the highest C-indices observed in the study were for former DV perpetrator, and 21 to 24 and 30 to 39 age bands (all had C-indices of > 0.900; Table A23). The lowest discrimination score OSP/I achieved was for those where ethnicity was not recorded, but still had good discrimination (0.779) indicating that OSP/I had good or better discrimination across all subgroups. As with contact sexual reoffending, please exercise caution where results are based on smaller groups for rarer types of offending and will be subject to greater variation between studies.

6.8 Accuracy in the prediction of all serious (RSR-type) reoffending

All serious reoffending (total RSR-type) is any serious nonsexual violent (SNSV) offence or any serious sexual offence (contact sexual offending or sexual offending involving indecent images). The ‘total’ RSR score comprises the sum of the risk of SNSV (RSR SNSV), contact sexual (OSP/C) and sexual reoffending involving indecent images (OSP/I), and is therefore dependent on the performance of those three constituents. Both sexual predictors are static (for men with sexual offending history only), whereas RSR SNSV can be calculated using static or static/dynamic risk factors. As such, the RSR predictor can be

[51] Largest effect size, 18 – 20 age band, odds ratio 4.787
[52] Odds ratio 2.971
[53] Two-tailed test: Z = 1.592, p-value = 0.111, odds ratio 1.325

calculated for everyone in the study (using dynamic factors, where available, for the RSR SNSV tool).

In this section the RSR predictor will be assessed and compared with OGRS4/V and OVP2 where appropriate.

Summary of findings

- The RSR predictor had a small miscalibration and was well calibrated in some risk bands; and
- The RSR predictor showed good or better discriminative validity and performed better than the closest available comparative predictors.

Model calibration

All individuals

Overall, the RSR predictor was found to be right on the threshold for a small miscalibration,[54] under-predicting reoffending with a residual of 0.43 points (Table A8). Across subgroups large miscalibration was observed in Black and Mixed ethnicities[55] (1.37 and 1.68 points respectively) and moderate miscalibration observed in 18 – 20 and 50 – 59 age bands[56] (1.45 and 0.46 points respectively; Table A14). For all other subgroups the miscalibration was small or negligible.

Assessing risk bands, in absolute terms, RSR under-predicted reoffending and showed a decrease in calibration as risk band[57] increased (ranging from 0.14 points for low to 1.86 points for very high; Table A12). However, only the low and medium risk bands had small miscalibration.[58]

With an OASys Assessment

Overall, in absolute terms, the RSR predictor under-predicted reoffending (0.72 points; Table A8), a small miscalibration.[59] Across subgroups very large miscalibration was

[54] Two-tailed test: z = 9.300, p < 0.0001; odds ratio 1.255

[55] Largest effect size, Mixed ethnicity, odds ratio 1.797

[56] Largest effect size, 18-20 age band, odds ratio 1.556

[57] Operationally, RSR is presented in three risk bands, with the lowest covering the range of 0 to 2.99 per cent, but is presented here in four bands, of which the lower cover 0 to 0.99 and 1 to 2.99 per cent.

[58] Largest effect size, medium risk band, odds ratio 1.342

[59] Two-tailed test: z = 10.285, p < 0.0001; odds ratio 1.341

observed in Black ethnicity (2.14 points),[60] and large in Mixed ethnicity as well as those in the 50 – 59 and 60 and over age bands[61] (Table A18). Moderate miscalibration was observed in those aged 18 – 20 (2.19 points) and females (0.37 points over).

Considering risk bands for people with an OASys, RSR followed the pattern described for RSR when used for all cases – residuals increased as risk band increased (ranging from 0.21 points for low to 2.02 points for very high; Table A12) and only small miscalibration for those in the low and medium risk bands.[62]

Model discrimination

All individuals

Overall, discrimination was better for RSR predictor (0.752) than the closest available comparable predictor,[63] OGRS4/V (0.677;Table A24).

Across the subgroups analysed, RSR's discrimination was generally good or excellent (C-index > 0.714) with acceptable in discrimination in only three subgroups (those in the 40 – 49 age band, likely LDC and former DV perpetrators; Table A24). RSR generally showed better discriminative validity than OGRS4/V, with the exception of females (RSR = 0.740; OGRS4/V = 0.748). The RSR predictor's discrimination increased slightly with age, while OGRS4/V discrimination fell dramatically (people aged 60 and over with 0.475).

With an OASys Assessment

For the subset of the population for which an OASys assessment was available, the RSR predictor was seen to have better discrimination than OVP2 and OGRS4/V, both overall and in all subgroups that were analysed, except for females and Asian ethnicity, where RSR was outperformed by OGRS4/V and OVP2 respectively (Table A24).

60 Odds ratio 2.001

61 Largest effect size, 60 and over, odds ratio 1.999

62 Largest effect size, Medium risk band, odds ratio 1.484

63 Note that OGRS4/V was designed to predict broad violent reoffending and all RSR reoffending includes sexual reoffending

7. Conclusions

7.1 Overview

When assessing an offender's risk of reoffending it is useful to distinguish 'static' and 'dynamic' risk factors. Static risk factors are fixed factors such as age, sex, current offence and criminal history. These are available for all individuals in the study. Dynamic risk factors are changeable factors such as accommodation, employment, substance misuse, temper control and antisocial attitude. These are only available for individuals who have received a complete OASys Layer 3 assessment. Where dynamic risk information was available, it was possible to compare predictors that use only static factors with those using both static and dynamic factors. The predictors using both types of factor generally performed better overall: they were better calibrated and had better discrimination. That is, they were better at predicting rates of reoffending on average and were better at discriminating between lower and higher risk offenders.

The majority of the results reported were based on analyses run on the '*starts*' dataset, capturing those starting their orders and licences, as this best reflects operational practice. However, analyses of the 'caseload' dataset, a snapshot of the probation population on 30 June 2018, were also reported to investigate the importance of offence-free time; that is, the number of months an individual has been in the community without a proven reoffence. Accordingly, several results demonstrated that the newer generation of predictors that involved offence-free time (OGP2, OVP2, OGRS4/G, OGRS4/V and RSR SNSV) tended to be most successful on the caseload dataset.

7.2 Predictors by offence type

All proven reoffending

For all proven reoffending, all predictors were well calibrated, with OGRS3 possessing the smallest residual (difference between the actual and predicted rates) for those with an OASys assessment. All three predictors demonstrated good overall discrimination.

When offence-free time is a factor, both OGRS4/G and OGP2, which account for offence-free time, performed better than OGRS3, which does not.

Broad violent reoffending

OGRS4/V, OVP1 and OVP2 all predict broad ('OVP-type') violent reoffending. OGRS4/V is the only static predictor of broad violent reoffending tested and can be calculated for everyone in the population while OVP1 and OVP2 are dynamic predictors requiring an OASys Layer 3 assessment. Of the three, only OVP1 is currently in use. As such, there is an operational gap in the prediction of OVP-type violent reoffending for those without an OASys.

OVP1 and OVP2 were both well calibrated while OGRS4/V had a small miscalibration.

Both OGRS4/V and OVP2 demonstrated good discrimination, overall, and OVP1 had acceptable discrimination. OGRS4/V and OVP2 performed better than OVP1 when offence-free time was considered.

Serious nonsexual violence reoffending

Considering serious nonsexual violence (SNSV) reoffending, the RSR SNSV static predictor was well calibrated. Where an OASys was available, both the static and static/dynamic versions of RSR SNSV were well calibrated and the static/dynamic SNSV predictor showed better discrimination than predictors of broad violence. Both the static and static/dynamic version of RSR SNSV had good discrimination.

Both forms of RSR SNSV performed better when offence-free time was considered.

Sexual reoffending

With regards to sexual offending, two static predictors are available for men with sexual offending history to predict contact sexual offending (OSP/C) and sexual offending involving indecent images (OSP/I).

OSP/C was not found to be well calibrated overall (very large miscalibration). OSP/C was found to have acceptable discrimination but with an overall measure of discrimination ability (the C-index) below that reported in earlier work (Howard & Wakeling, 2021). This

loss of performance warranted further analysis, which is published alongside this study (Emeagi et al., 2024).

OSP/I was not well calibrated for the prediction of sexual offending involving indecent images (large miscalibration). OSP/I displayed good to excellent discrimination overall. The Emeagi et al. (2024) study investigates further the prediction of both types of sexual offending.

All Serious (RSR-type) reoffending

The RSR predictor had a small miscalibration for the prediction of reoffending rates, underestimating the actual rate of reoffending slightly. The degree of underestimation was smaller when considering all individuals than when considering just those with an OASys assessment.

The RSR predictor was found to have good discrimination.

8. References

Bell, K. (2018). *A quantitative study of the inter-rater reliability and completion timings of the OASys Sexual reoffending Predictor (OSP) compared to the Risk Matrix 2000 (RM2000/s)*. Report prepared for HMPPS. Unpublished.

Emeagi, C., Sullivan, L., Landsiedel, J., Craik, A. & Howard, P. (2023). *The Actuarial Prediction of Sexual Reoffending: Responding to Changing Offending Patterns. (Ministry of Justice Analytical Series, unnumbered)* [Manuscript in Preparation]. London: Ministry of Justice.

Hanson, R. K., Harris, A. J. R., Helmus, L. & Thornton, D. (2014). High-risk sex offenders may not be high risk forever. *Journal of Interpersonal Violence, 29*, 2792-2813. doi: 10.1177/0886260514526062

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*, 361–387. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Her Majesty's Prison and Probation Service [HMPPS] (2020). HMPPS risk of serious harm guidance. Available at: https://www.gov.uk/government/publications/hmpps-risk-of-serious-harm-guidance-2020 (accessed 06/10/2023)

Home Office (2006). Offender Assessment System Manual version 2. London: Home Office.

Howard, P (2009a). OGRS3: the revised Offender Group Reconviction Scale. London: Ministry of Justice. Available at https://core.ac.uk/download/pdf/1556521.pdf

Howard, P (2009b). Predictive validity of OASys – Improving prediction of violent and general reoffending. In Debidin, M. (Ed.) (2009) Compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009. Ministry of Justice Research Series 16/09.

Howard, P. (2014). OGP2 and OVP2: the revised OASys predictors. In Moore, R. (Ed.) (2015) *A compendium of research and analysis on the Offender Assessment System (OASys) 2009 – 2013*. (Ministry of Justice Analytical Series, unnumbered). London: Ministry of Justice. Available at https://assets.publishing.service.gov.uk/media/5a7f676fed915d74e33f6380/research-analysis-offender-assessment-system.pdf (accessed 06/10/2023)

Howard, P. (2015). OGRS4. The revised Offender Group Reconviction Scale. In Moore, R. (Ed.) (2015) *A compendium of research and analysis on the Offender Assessment System (OASys) 2009 – 2013*. (Ministry of Justice Analytical Series, unnumbered). London: Ministry of Justice. Available at https://assets.publishing.service.gov.uk/media/5a7f676fed915d74e33f6380/research-analysis-offender-assessment-system.pdf (accessed 06/10/2023)

Howard, P. D. (2017). The effect of sample heterogeneity and risk categorization on Area Under the Curve predictive validity metrics. *Criminal Justice and Behavior, 44 (1)*, 103-120.

Howard, P. D., Barnett, G. D., & Mann, R. E. (2015). Specialization in and within sexual offending in England and Wales. *Sexual Abuse: A Journal of Research and Treatment, 26*, 225–251. doi: 10.1177/1079063213486934

Howard, P.D. & Dixon, L. (2013). Identifying change in the likelihood of violent recidivism. Causal dynamic risk factors in the OASys Violence Predictor. *Law and Human Behaviour, 37*, 163-174.

Howard, P., Francis, B., Soothill, K., & Humphreys, L. (2009). *OGRS 3: the revised Offender Group Reconviction Scale* (Research Summary 7/09). London: Ministry of Justice.

Howard, P., & Moore, R. (2009). Measuring changes in risk and need over time using OASys. In Debidin, M., (ed), *A compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009*, p. 108-135. London: Ministry of Justice Research Series 16/09.

Howard, P., & Wakeling, H. (2021). Comparing two predictors of sexual recidivism: the Risk Matrix 2000 and the OASys Sexual Reoffending Predictor. *Ministry of Justice Analytical Report*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/955345/comparing-2-predictors-sexual-recidivism.pdf

Ministry of Justice (2021). End-to-End Rape Review Report on Findings and Actions. Available at https://www.gov.uk/government/publications/end-to-end-rape-review-report-on-findings-and-actions (accessed 06/10/2023)

Wakeling, H. (2018). The development of a screen to identify individuals who may need support with their learning. *Analytic Summary*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/740303/development-screen-identify-individuals-oasys-report.pdf (accessed 06/10/2023)