Ministry of Justice

Post-release reoffending outcomes for individuals with offence-related sexual paraphilias

An exploratory risk-band analysis

Ian A. Elliott

Eleanor Martin

Ministry of Justice Analytical Series

2023

The Data and Analysis Directorate 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 2023

Any enquiries regarding this publication should be sent to us at researchsupport@justice.gov.uk

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

ISBN 978-1-84099-9747

Acknowledgements

We would like to thank colleagues within the Ministry of Justice Data and Analysis Directorate Alana Diamond, Aidan Mews, Philip Howard, Olivia Hambly, and Wura Gerasimov. In addition, we thank two academic peer reviewers for their insightful comments on this report.

Final thanks are to HMPPS Intervention Services for their communication and support throughout the evaluation process.

Contents

List of tables

List of figures

1. Summary

Introduction and aims of the study

Within the wider population of individuals with a conviction for sexual offences, His Majesty's Prison and Probation Service (HMPPS) supervise a small sub-population that are assessed as having a sexual paraphilia and are predominantly considered to be a high risk of reoffending. Sexual paraphilia refers to interest in atypical sexual targets (e.g., animals or children) or unhealthy sexual behaviours (e.g., voyeurism/sadism).

To rehabilitate adult males with offence-related sexual interests, HMPPS provides the Healthy Sex Programme (HSP), an accredited cognitive-behavioural intervention that aims to increase sexual self-regulation skills toward leading safer lives. HSP is underpinned by a bio-psycho-social model of change primarily focusing on meaningful risk factors for sexual recidivism. Guided by the principles of "risk, need, and responsivity", HSP follows strengths-based organising principles to promote an offence-free life, and is delivered flexibly on a one-to-one basis. The latest published figures show that 109 people completed HSP during 2019/20 and between April 2015 and March 2018 (the dates that broadly align with the study period), 267 people completed the programme.

The aim of this report was to present the first analysis undertaken for HSP, specifically:

1. basic demographic details for a sample of individuals who have completed HSP and had returned to and resided in the community for longer than six months;
2. the frequency/nature of reconvictions for new offences and recalls to prison;
3. comparisons between the actual frequency of reconvictions among HSP participants with predicted rates (i.e. the frequency that would have been expected to be reconvicted based on a statistical assessment of their "future risk" of reoffending); and
4. relationships, if any, between observed reconvictions and individual characteristics.

Methodology and interpreting results

The sample consisted of 112 adult males who had completed HSP between 2015 and 2018 across 13 secure establishments and had subsequently been released from prison and spent more than six months residing in the community (hence the smaller number than the latest published figures).

Statistical analysis was conducted using a risk-band analysis (RBA) to assess the frequency of future proven reoffending among HSP participants. The actual number of HSP reconvictions were compared to the predicted number generated from a sexual offending risk assessment tool (i.e. RM200/s). This approach was adopted as it was not possible to create a matched comparison group (to act as a relevant real-life control group) from which to robustly assess comparative sexual reoffending rates, which is considered the minimum standard required for methodological robustness in evaluation.

The RBA included assumptions and estimates of important variables that may be prone to variation, bias, or measurement error. Due to short follow-up times for some participants statistical projections for 1-year reconviction rates had to be used, instead of published predicted rates. Differences between observed and predicted frequencies may indicate that the risk measure is not suitable or calibrated for the paraphilic population rather than an indication of higher or lower than expected outcomes for this sample. Therefore, this is an exploratory study and results must be considered indicative.

Key results

Acknowledging the methodological limitations for this study, the main results were:

- Amongst the HSP sample, 30% received some form of post-release reprimand. The proven sexual reconviction rate was 7% and a further 20% were recalled to prison for breaches of conditions of release (e.g., not disclosing electronic devices, unsupervised access to children, possession of pornography, or travel/curfew violations). The remaining 3% received a non-sexual reconviction or a conviction for a historical crime.

- The RBA analyses was used to assess whether actual HSP reoffending rates were higher or lower than statistically predicted rates. The main results indicated there was, no change, neither higher nor lower than predicted.

- Additional analyses found that 30% of the HSP sample obtained a score on a screening test for learning disabilities and challenges (LDC) that indicated potential for LDC (but this test does not confirm it). Almost half of the HSP sample scored highly on a test for paedophilic interests.

- Problems with access to permanent accommodation was predictive of reconviction and recall combined, whereas paedophilia, potential for LDC, relationship problems, cognitive deficits in thinking and behaviours, and pro-criminal attitudes were not.

Conclusions

This exploratory analysis of a sample of HSP participants aimed to examine simple frequencies of outcomes since, for reasons described above, a rigorous impact study was not possible. Given several methodological issues, results must be considered indicative.

This was a sample for whom reoffending risk assessment is more complex due to the additional presence of paraphilic interests. Analysis to assess whether actual HSP reoffending rates were higher or lower than predicted rates, suggested there was no statistically significant difference between actual and predicted rates. That said, it is noteworthy that new sexual offences were not observed after release at a frequency above that predicted, based on a wider general population of individuals with a sexual conviction. Further analysis should be considered when larger samples become available.

2. Background

Within the larger population of individuals with a conviction for sexual offences, His Majesty's Prison and Probation Service (HMPPS) and associated agencies supervise a relatively small sub-population that are considered to have a sexual paraphilia. Typically, a sexual paraphilia refers to an interest in anomalous sexual targets (such as animals or children) or atypical sexual behaviours (such as voyeuristic or sadistic activities) (Winder et al., 2018). If these sexual interests persist for at least six months and are accompanied by clinically-significant impairment or distress, are interfering with otherwise healthy sexual activity, and/or pose a risk of harm to the individual themselves or others around them, it can be classified and diagnosed as a paraphilic disorder. Paraphilic disorder is listed in the latest editions of both the International Classification of Diseases (ICD-11: World Health Organisation) and the Diagnostic and Statistical Manual of Mental Disorders (DSM-5: American Psychiatric Association, 2013).

In a review of meaningful risk factors for sexual offending, Mann et al. (2010) considered unhealthy sexual interests, to have strong support as a risk factor in terms of the amount of evidence indicating that it statistically predicts relevant future reconvictions. This includes the presence of multiple paraphilias. Nonetheless, paraphilic disorders can be treated and managed. As Winder et al. (2019) note in their review of the treatment of paraphilic disorder, cognitive-behavioural therapy (CBT) has been the main form of psychotherapeutic treatment for identified paraphilic disorders in the United Kingdom.

CBT has also been accompanied by the systematic use of pharmacological treatment in the early 2000s, such as selective serotonin reuptake inhibitors, synthetic steroidal analogues, and gonadotropin-releasing hormone analogues (see Holloyda & Kellahar, 2018). Notably, neither cognitive-behavioural or pharmacological approaches seek to "cure" a paraphilic disorder, which – at least for some – is increasingly considered to be outside of the volitional control of the individual. Instead, they aim to support individuals in managing sexual arousal and behaviours that result from it (Berlin, 2019).

To rehabilitate males with offence-related sexual interests, HMPPS provides the Healthy Sex Programme (HSP), a cognitive-behavioural intervention that aims to increase sexual self-regulation skills toward leading safer lives. It is targeted at adult males who have been incarcerated following a conviction for a sexual offence (or one with a sexual element or motivation) and who have specific treatment needs in relation to offence related sexual interests. HSP is underpinned by a bio-psycho-social model of change primarily focusing on meaningful risk factors for sexual recidivism (see Carter & Mann, 2016; Mann et al., 2010). Guided by the principles of “risk, need, and responsivity” (RNR: Andrews & Bonta, 2010), HSP follows strengths-based organising principles to promote an offence-free life. HSP is derived and adapted from the previously accredited Healthy Sexual Functioning programme (HSF). It is delivered with flexibility on a one-to-one basis and is accessible to individuals with learning difficulties and challenges.

The programme is designed to address needs as they relate to possessing “a sexual preference for children”, “preferring sex to include violence or humiliation” or “other offence-related sexual interest” on the Sexual Interest scale of the Structured Assessment of Risk, Need and Responsivity (SARNR). HSP was updated in 2019 and subsequently received accreditation from the Correctional Service Advise and Accreditation Panel (CSAAP). Broadly, the main revisions to the programme included a greater parity of clinical content to people with LDC, increased options to include structured work on mindfulness and personal values, introduction of exercises that aim target the effects of shame, and a more structured approach to relapse prevention.

HSP sessions are divided into five modules: (1) engagement; (2) understanding my sexual interests; (3) “new me” and sex; (4) sex and a better life; and (5) and bringing it all together. Dosage is dependent on the needs of the participant, lasting between 12 and 30 hours in total. Although HSP is clearly time-bounded not open-ended, the variation in dosage allows the therapist and client to develop a treatment plan that flexibly attends to strength, needs, and responsivity issues as they arise during skills practice. Although no specific data is collected on hourly dosage, programme developers indicated that it was rare for participants to complete only 12 hours, most completing at least 20 hours and many completing 25 to 30 hours. The latest published figures show that 109 people completed HSP during 2019/20 and between April 2015 and March 2018 (the dates that broadly align with the study period), 267 people completed the programme. (Ministry of Justice, 2020).

The aim of this study was to present the first analysis for HSP and explore criminal justice outcomes for individuals with a paraphilic sexual interest, via a sample of individuals who had completed HSP. It sought to examine:

(1) demographic details of a sample of HSP participants who have completed the programme and returned to the community for longer than six months;

(2) the frequency and nature of reconvictions for new crimes and recalls to prison for the sample;

(3) actual recidivism figures compared to estimated expected frequencies based on risk assessment results using risk-band analysis techniques; and

(4) any relationships between outcomes and scores on measures of dynamic risk, sexual interest, and learning disabilities and challenges (LDC).

3. Method

3.1 Sample

This study utilised participants on HSP as representative of a sample of adult males who, despite not necessarily being formally diagnosed, exhibit the symptoms and characteristics of paraphilic disorder. The current sample was drawn from an HMPPS clinical database of 120 HSP participants who had completed the first iteration of the programme (had accessed the pre-revised version of HSP) between 2015 and 2018 in 13 establishments.

After removing those who remained in custody, a final sample of 112 HSP participants remained for the planned analyses. In terms of completion, 34 had completed HSP in 2016, 48 in 2017, and 30 in 2018. Most the participants completed HSP at HMP Whatton (37.5%), HMP Wymott (14.3%), and HMP Usk (8.9%), with the remaining 39.3% completing the programme at various other secure establishments. Ages ranged from 21 to 78 years, with a mean age of 47.9 (standard deviation (SD) = 15.7). Of the participants, 83 had been serving a determinate sentence with 29 serving indeterminate sentences.

These participants had each completed the programme, been released from prison, and subsequently had spent from three months to three years residing in the community. For this sample, participants who had resided in the community for less than six months (n = 7) or for whom no accurate PNC number could be established for follow-up (n = 1) were excluded from the analysis. Of those 120, 99.2% were classified as “White – North European” for the purposes of law enforcement records (via PNC), with the remainder being classified as “Other”. Self-reported ethnicity data were not available.

Risk assessment

Risk assessment scores were available for the sample, based on the Risk Matrix 2000 (RM2000: Thornton, 2007; Thornton et al., 2003). The RM2000 is a static actuarial risk assessment tool designed for adult male sex offenders that consists of three scales that are designed to assess the likelihood of sexual (RM2000/s), non-sexual violent (RM2000/c), and any violent recidivism (RM2000/v) (Helmus et al., 2013).

Actuarial risk measures for recidivism differ from assessments based on clinical judgment by mathematically estimating the likelihood of recidivism over a set time using objective variables that have been found to be statistically predictive of reconviction for individuals with existing sexual convictions. RM2000/s is scored using "static", unchangeable demographic and criminal history information and assigns individuals to a final risk category of either low, moderate, high, or very high likelihood of a future sexual reconviction (Helmus et al., 2013). RM2000/s classifications are presented in Table 1.

Table 1: Risk classifications for the HSP sample based on the Risk Matrix 2000/s

	Risk Category				
	Low	Medium	High	Very high	Total
HSP sample	5 (4.5%)	32 (28.6%)	50 (44.6%)	25 (22.3%)	112

As can be seen in Table 1, two-thirds (66.9%) of the HSP sample were categorised as high or very high risk according to the RM2000 risk classifications.

3.2 Procedure

This study utilised a risk-band analysis (RBA) to statistically compare an observed outcome with an expected outcome based on a predictive measure (RM2000/s). Rather than seek to generate a control group of non-paraphilic participants with which to compare outcomes, it generates a counterfactual for comparison based on the estimated frequency of outcomes predicted by the sample’s profile on a relevant diagnostic tool.

At its most basic, the RBA estimates the predicted number of recidivists by multiplying the number of participants assigned to each risk band of the measure with the respective normative rates of that band. The sum of those estimates with the observed number of recidivists are then statistically compared. Because the normative rates are calibrated from untreated samples from an appropriate population, the underlying assumption is that the predicted number of recidivists represents what would be expected for an untreated sample with the same number of individuals in each risk band and represents a counterfactual to observed recidivism in a treated sample (treated vs. untreated).

There are three key variables necessary for an RBA with categorical risk assessment data: assessment that classify individuals to natural ordered categories, like the RM2000/s. The first are risk classifications resulting from a validated measure known to be statistically predictive of the outcome under investigation (see Sample section). The second is an observed (or known) frequency of the outcome or outcomes under consideration for the programme sample. The third is a normative rate or rates for the expected frequency of outcomes one can expect for scores or classes resulting from the predictive measure.

Chi-squared tests of association (Pearson's Chi-squared test and Fisher's Exact Test) were used to test the hypothesis that the observed frequency differs statistically significantly from the expected frequency, both of which derive the expected frequency via the method described above and are used in this study. These techniques involve entering the observed and expected number of recidivists into a 2x2 table and the calculation of a test statistic (a ratio or difference in ratios) that indicates how our predicted frequencies of recidivism or no-recidivism compare with our observed frequencies. This technique has been used in other risk band analyses in criminal justice (e.g., Woodrow & Bright, 2011).

Because it lacks a control group the RBA method is not considered a rigorous technique and cannot be used to draw "causal inferences" about the relationship between group membership and outcomes. Consequently, the findings are presented as a descriptive indication of outcomes and cannot be causally attributed to participation in HSP.

3.3 Outcomes

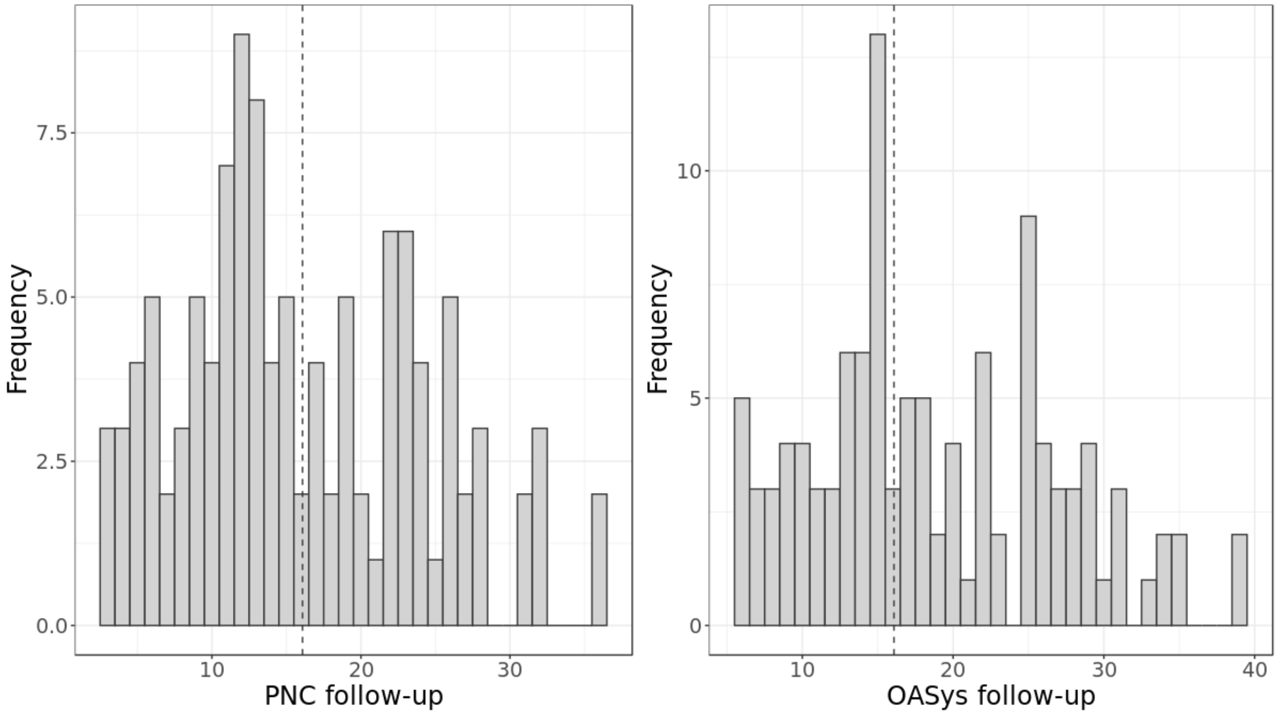
A criminal history search was obtained from the Police National Computer (PNC) Team in the MOJ Data and Analytical Services Directorate, to establish the number of sample participants who had proven criminal appearances post-release and up to the date of that PNC search. Follow-up times on the PNC ranged from 3.3 to 36.0 months, with a mean of 16.1 months (Median (Mdn) = 14.5) (see Figure 1). Follow-up lengths constitute the duration of time to the first new criminal justice appearance, after which follow-up stops.

Almost two-thirds (64.3%) had a follow-up duration of one year or greater. It is important to note that there is an acknowledged time lag between new crimes being processed and subsequently available on PNC searches that is estimated to be around six-months but could be considerably longer. It may be possible that some or all in the sample who have been returned to custody for breaches of conditions of release are consequently being investigated for new crimes that have not yet appeared on the PNC as new convictions.

Consequently, the sample was also subjected to a follow-up on the Offender Assessment System (OASys). OASys combines actuarial methods of prediction with structured professional judgement to provide standardised assessments of prisoners' and probationers' risks and needs, helping to link these risks and needs to individualised sentence plans and risk management plans (Moore, 2015). OASys follow-up durations ranged from 6.0 to 38.7 months with an average of 18.8 months (Mdn = 17.2 months) and 79.5% of the sample had a follow-up duration of greater than 1 year (see Figure 1).

OASys full "Return to custody" reports were examined and coded for (a) any return to custody for breaches of conditions of release and (b) broad reasons for those returns and details of breaches. It was hoped that these reports might note the presence of ongoing investigations for possible new crimes that could indicate recidivism over-and-above the initial breach. However, the information was not available with enough consistency and is not included for analysis.

Figure 1: Histograms of the PNC and OASys follow-up durations in months (the dashed line indicates the mean)



Estimating expected recidivism rates

This study used normative values derived from the 2007 Scoring Guide (Thornton, 2007), which are based on data from "a national sample of adult males sentenced to prison in England and Wales for sex offences and released in 1979… followed for 16 years using central police records and for 19 years using a central statistical database, the Offender Index, that holds data from the courts" (p. 13).

At the current time, neither these scoring guidelines or subsequent validation studies for the measure provide 1-year normative reconviction rates for the four risk bands that would be appropriate for short-term post-treatment samples (e.g., Barnett, Wakeling, & Howard, 2010; Grubin, 2008). Consequently, we generated estimates from the known longer normative follow-up rates for RM2000/s recidivism rates using linear, exponential, and logarithmic trend functions on Microsoft Excel (see Appendix A). Linear trends are those in which data points increase or decrease at a constant rate, exponential trends are those in which data points increase or decrease at increasingly greater rates, and logarithmic trends are those in which data points increase or decrease quickly and then level-out.

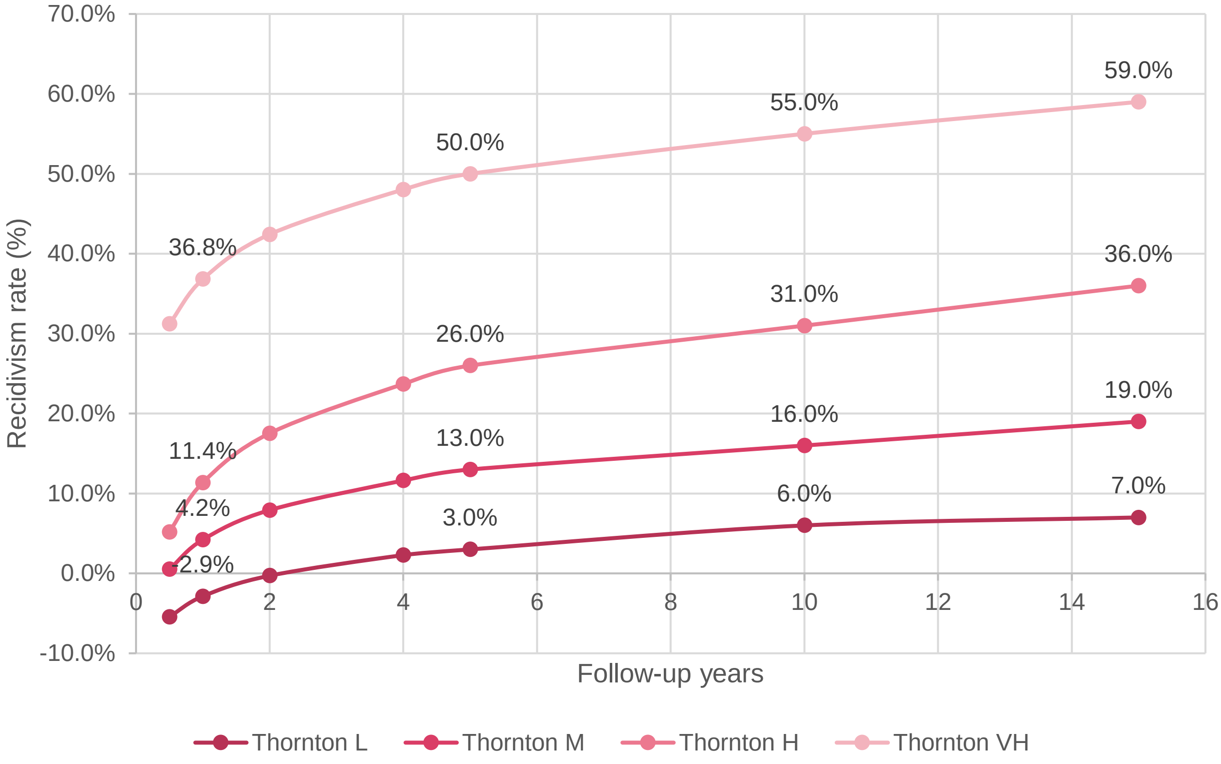
Table 2: Estimated linear, logarithmic, and exponential 1-year recidivism rates for RM2000/s

	Low	Medium	High	Very High
Linear	1.7%	10.6%	22.0%	46.6%
Exponential	2.3%	11.2%	22.9%	47.0%
Logarithmic	-2.9%	4.2%	11.4%	36.8%

Note: Where negative values were returned (e.g., -2.9%) a rate of 0.0% in the estimation function was applied.

These estimates are presented in Table 2 and the logarithmic trend lines from which the final estimates were derived can be seen in Figure 2. The 1-year rates were chosen based on the finding that the median follow-up times were 14.5 months for the PNC and 17.2 months for OASys. The logarithmic estimates were chosen for use in the analysis after consultation with internal MOJ analytical experts on recidivism rates for individuals with sexual convictions. They were considered to deliver the most conservative of the three sets of estimates and best represent the current historical trend towards lower absolute rates even at the very high-risk category (see for example, Barnett et al., 2010).

Figure 2: Logarithmic decay from the published RM2000/s normative reconviction rates to 1-year estimates



Adjusting for risk assessment measurement error

As described above, an RBA typically calculates the predicted number of recidivists as the number of sample participants in each risk band multiplied by the recidivism rate associated with that risk band. For example, if 100 individuals are assigned to the "Medium" risk band and we believe the reoffending rate for individuals with a medium risk score is 10%, we would expect 10 of those 100 individuals to reoffend. Implicit in this understanding of RBA is an assumption that the risk assessment is perfectly predictive of reoffending: that the "true" risk of reconviction for every individual is predicted by their scores and their classification. We know, however, that this is unlikely to be the case.

Although the predictive ability of RM2000/s has been found to be relatively stable across prior samples, its practical use includes some measurement error. Barnett et al. (2010) validated the scale using 2,755 individuals with sexual convictions and reported an AUC value of 0.68. The AUC represents the probability that for two randomly selected cases, one with a positive outcome and one with a negative outcome, the tool will have correctly classified the positive one as higher risk. Two meta-analyses combining data from international samples (the latter including the Barnett et al. study) reported AUCs of 0.68 (Hanson & Morton-Bourgon, 2009: *n* = 2,755) and 0.74 (Helmus et al., 2013: *n* = 10,644).

It is therefore reasonable to argue that not all cases in the HSP sample will have been correctly classified by RM2000/s. This has implications for the number of individuals with reoffences that a risk band analysis will predict. For the frequency of cases in each of the bands, the risk tool may have naturally underestimated risk (assigned individuals to “low” or “medium” bands when their true likelihood of reoffending is higher) or overestimated risk (assigned individuals to “high” or “very high” bands when their true risk is lower).

For RM2000/s a bootstrapping process was created that simulates that measurement error and its effect on the distribution of predicted recidivists (see Appendix A). Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This process was designed to simulate what happens if we re-assign individuals in the HSP sample to a risk band with a reasonable amount of measurement error accounted for. This produced a range of values for the total number of reoffenders that RM2000/s would predict, rather than one single value, representing an estimate of the extent to which that predicted total might increase or decrease when measurement error is present.

To conclude that the observed number of recidivists is statistically significantly different to the number of predicted recidivists, the observed value must be significantly lower than a reasonable lower boundary of the underlying distribution or significantly higher than a reasonable upper boundary of that distribution. Otherwise, it is fair to argue that differences between the observed value and the expected value could be a result of the risk tool under- or over-estimating risk due to measurement error. This gives us three scenarios to test:

(1) Is the observed value statistically significantly higher/lower than the predicted value if we presume RM2000/s has worked perfectly?

(2) Is the observed value statistically significantly higher/lower than the predicted value if RM2000/s has underestimated risk in our sample?

(3) Is the observed value statistically significantly higher/lower than the predicted value if RM2000/s has overestimated risk in our sample?

It is important to note that RM2000/s was only designed and validated to estimate the risk of future proven contact sexual reconvictions, not for the risk of other future criminal justice sanctions, such as recalls to prison and/or breaches of licence conditions. As such, the use of RM2000/s here to judge the expected likelihood of other criminal justice sanctions is a non-standard use of that tool and is not a form of analysis it was designed to support. These analyses are presented here on the basis that our qualitative analysis of the sanctions received in this sample arguably represent evidence of empirically established risk-related behaviours (e.g., unauthorised contact with children, unauthorised access to technology with potential access to prohibited imagery). It is also worth noting, however, that the RM2000/s was also not designed and validated for use with individuals with convictions relating to indecent images of children (IIOC) nor individuals with LDC.

3.4 Other predictors

The potential predictive validity of other relevant assessments was explored with the aim to provide some context for the RBA findings, via a binary logistic regression model (see Appendix A for details). Three sources of variables were chosen. The first set of variables were taken from the Offender Assessment System (OASys). These are considered relevant since they are based on variables that have been identified in the criminal justice literature as being those that are associated with likelihood of future offending. The second and third were taken from assessments administered as per the assessment criteria for HSP. These were scores on a screening assessment for the presence of paedophilic (pre-adolescence) or hebephilic (early adolescence) interests and scores on a screening assessment for the potential presence of learning disabilities and challenges (LDC).

Paedophilic interests were considered a relevant variable for predictive ability as it indicates the extent of paedophilia per individual (the most commonly observed paraphilic interest of those that are used to assess eligibility for HSP). LDC was of interest as some consensus exists that the related deficits in cognitive functioning and emotional regulation place members of the LDC population at heightened risk of involvement in the criminal justice system (e.g., Lindsay, 2011; Lindsay, Hastings, & Beech, 2011). As Lindsay (2011) notes, several studies have found higher rates of prior sex offending among prisoners with LDC. However, there are some that suggest rates of sexual convictions among persons with LDC are difficult to quantify and may parallel rates in general prison populations.

In total, 6 variables were entered into a binary logistic regression model.

- **Accommodation** – a lack of permanent decent accommodation (also a proxy for social exclusion).
- **Relationships** – problems in the development and maintenance of relationship stability and satisfaction.
- **Thinking and behaviour** – cognitive deficits in areas including impulse control, problem solving, perspective taking, and flexible thinking.
- **Attitudes** – pro-criminal attitudes, hostility towards others, and negative perceptions of supervision and restrictions.
- **Screening Scale for Pedophilic Interests – Version 2** (SSPI-2: Seto et al., 2015) – a brief actuarial measure of paedophilic or hebephilic sexual interest (sexual arousal to children).
- **Learning Screen Tool (LST)** – a measure to identify individuals who require further assessment of IQ for allocation onto an adapted suite of programmes.

3.5 Limitations

There are, however, important limitations to the RBA methodology that mean these findings should be considered indicative and descriptive. There is no authentic control group and consequently the methodology does not reach even the minimum level required for methodological robustness for impact evaluations (Level 1 on the Maryland Scientific Scale (MSS): Farrington et al., 2002). Systematic reviews of evidence often consider methods judged to be MSS Level 3 and above as being of the highest quality. Few risk band analyses exist in the published scientific literature and the lack of robustness is likely to be a key reason why that is the case.

It is important to stress the reliance of the RBA technique on the accuracy of the risk measure and the validity of that measure with the population being examined. Our findings are presented to provide rare outcome data for a small and atypical population. Any differences between observed and predicted frequencies, however, may indicate that the risk measure is not suitable or calibrated for this population rather than an indication of higher or lower than expected outcomes for this sample or their wider population.

The RBA also included assumptions and estimates of key variables that themselves may be prone to measurement error. For example, we were reliant on the variable quality and consistency of OASys and PNC data to calculate the RM2000/s classifications. As noted in the procedure, there were difficulties in establishing the true actual rates of recidivism. Of the sample, 36% had a PNC follow-up duration of less than 12 months, which indicates that for a sizeable minority of the sample we may have missed new convictions as they might not have appeared on the PNC reports available to us at the time.

The short-term follow-ups also meant that we needed to generate normative 1-year reconviction rates that were projections from existing published rates, and therefore may not be an accurate reflection of true rates. This is also based on an assumption that the Thornton (2007) normative data for RM2000/s are an accurate reflection of U.K. reconviction rates for each category.

As the guidelines state, “...those sex offenders allocated to high security prisons will typically differ from those allocated to lower security levels. Accordingly, anyone using RM2000 may wish to establish norms for the particular context in which they work." (p. 12). There are further questions about the predictive ability of RM2000/s. For example, a reanalysis of the Barnett et al. (2010) dataset found that the uncertainty associated with RM2000/s predictions that any one individual will re-offend was extremely large, and that the AUC statistic (small as it might be) is not a good reflection of the tool’s ability to accurately predict future behaviour (Cooke & Mitchie, 2014).

It is also important to note that although there is no set cut-off for AUC values and values are context dependent, a binary classifier with an AUC of less than .70 is unlikely to represent an accurate predictive tool. Howard (2001) also noted that AUCs can vary, even within a single study dataset, between subpopulations based on the heterogeneity of scores and the risk-score distributions within them. Finally, the pitfalls of using the AUC as a performance metric at all for tools of this kind are many (see, for example, Berrar and Flach, 2012) – albeit, it was the only performance metric available – and various alternatives have been proposed, such as the partial AUC (McClish, 1989), cost-curves (Drummond & Holte, 2006), or H-measure (Hand & Anagnostopoulos, 2014).

The Thornton rates are also based on a relatively older dataset. Barnett et al. (2010) reported 2-year and 4-year RM200/s rates that fall well below the rates in the 2007 guidelines. Their 4-year rate for medium risk individuals is 2.8% compared with Thornton's 5-year rate of 13.0%. Moreover, a published reanalysis of Barnett et al.’s data suggested only four of the RM2000/s variables are truly predictive: “the person’s age together with…[the] number of sexual appearances, stranger victim of a sexual offense and a noncontact sex offense.” (Cooke & Michie, 2014: p. 48). Finkelhor and Jones (2006; Finkelhor et al., 2015) have also published data suggesting that rates of sexual violence in many countries have shown a steady decline since the 1990s. This calls into question the current relevance of the England and Wales “1979 discharge sample” as normative data on which to judge RM2000/s bands for more modern cohorts and in the context of IIOC convictions.

Since the samples were not specifically assessed for paraphilic interests it is also possible that individuals with paraphilic interests may have been present in the Thornton (2007) normative sample. This would mean that our sample is not so different from that sample as a proxy for the general population of individuals with sexual convictions. The decision to choose the logarithmic rates over the linear and exponential rates was based in part on these trends for lower baseline offending but remains a speculative one.

Future use of the RBA method should utilise new advances in risk assessment for this population, such as the OASys Sexual Reoffending Predictor (OSP: Howard & Barnett, 2015). For routine use, RM2000/s was replaced by OSP in 2021. Furthermore, using the logarithmic method with what are essentially probabilities, albeit linear ones, could have affected the integrity of the estimated rates. In future, it may be beneficial to utilise proportional hazard approaches to account for rates as probabilities not linear outcomes.

Lastly in terms of the RBA, the correction procedure that intended to account for measurement error on the RM2000/s required some estimates and assumptions to be made. For example, we decided to use the AUC statistic as a proxy for the accuracy of the RM2000/s when used in the field. Although we defend this decision on the basis that the AUC is a measure of the ability of a tool to judge relative risk, it is not a common use of the statistic. The bootstrap procedure itself was developed specifically for these analyses and diverges from other examples of RBA. It is important to note that the bootstrapping procedure provided a more refined estimate of the underlying distribution of classifications.

There are also clear technical and statistical limitations to the techniques used. Although the intention of the of the bootstrap procedure was to produce a proxy for absolute upper and lower limits to the possibilities of RM2000/s (i.e., what might be the possible “best” and “worst” case scenarios for varying performance in the field), it produced very small 95% confidence intervals suggesting that the “true” value of the 50,000 samples was likely somewhere within a fraction of the mean of approximately 16 (a standard error of .007).

Using the 1.5*IQR boundaries, therefore, provides only a broad interval in which the “true” estimate lies. Furthermore, the small sample size also means that even the bootstrapping method will provide inadequate results as the data it draws on as “correct” is small. Consequently, these inefficiencies might result in false positives/negatives.

4. Results

4.1 Recidivistic outcomes

Of 112 participants with paraphilic sexual interests released into the community after completing their sentence, 34 (30.4%) were found to have received some form of post-release criminal justice sanction. These broad outcomes can be seen in Table 3, which shows that most of those who re-entered the criminal justice system did so via recalls to custody for breaches of conditions of release. Table 3 also includes the number of “pseudo-reconvictions”: instances where an individual received a new conviction post-release for a historical crime committed prior the index offence.

Table 3: Frequencies of known recidivistic outcomes

	Frequency
Reconviction	9 (8.0%)
Sexual	*7 (6.3%)*
Non-sexual	*2 (1.8%)*
Pseudo-reconviction (sexual)	2 (1.8%)
Breach/recall	23 (20.5%)
Total	**34 (30.4%)**
Total excl. pseudo	**32 (28.6%)**

Table 3 also indicates a proven reconviction rate for new offences (not including pseudo-reconvictions) of 8.0%, within an average 16-month follow-up. Seven of those were sexual reconvictions (listed in Table 4) resulting in a proven sexual reconviction rate of 6.3%. One of the 2 violent offences had been classified on the PNC as a “public order offence” but OASys details indicate that the offence involved masturbating in public. If this is included as sexual recidivism, the proven sexual reconviction rate would be 7.1%. Of the three contact sexual offences, two were against female victims under 13 years of age and one was against a female victim over 18 years of age. The grooming offence was against a female victim under 13 years of age.

Because of the two different types of outcome, we have used “proven sexual reconvictions” (n = 8) to refer to proven reconvictions alone (although it is acknowledged that the term “proven” is not accurate for the public order offence previously described) and “all relevant sanctions” (n = 31: 27.7%) to refer to both proven sexual reconvictions and breaches/recalls to prison combined. Since the RM2000/s was designed to predict proven sexual reconvictions, not for other forms of sanctions, the comparison sample to which the HSP sample is being compared did not account for breaches/recalls.

Therefore, the number of “all relevant sanctions” predicted by RM2000/s does not include breaches/recalls and will be an underestimation. However, due to the short follow-ups and the qualitative evidence that those sanctions are relevant (i.e., were indicative of empirically established risk related behaviours with the potential to lead to investigations into new sexual crimes) analyses using other sanctions in this study have been included.

Table 4: Classifications of proven reconvictions

	Frequency
Contact sexual	3 (33.3%)
Indecent images of children (IIOC)	3 (33.3%)
Grooming	1 (11.1%)
Violent non-sexual	2 (22.2%)
Total	**9**

It was also possible to group broad types of licence breach together for the 23 cases of recall to prison via offence information given in OASys reports. These themes and the number of cases represented are presented in Table 5 below. Most cases (56.5%) included two themes, 30.4% with only one theme and 13.0% with three themes. As the table shows the most common reasons for recall to prison were for concealed technology (typically possession of unregistered digital devices), unsupervised access to children, possessing or accessing pornography – either legal (but still prohibited by licence conditions), or potentially illegal – and being detected in prohibited locations or violating curfew.

Table 5: Types of breach and percentage of breach cases with that theme

	Frequency
Technology	15 (65.2%)
Access to children	10 (43.5%)
Pornography	7 (30.4%)
Location/curfew	5 (21.7%)
Substance abuse	2 (8.7%)
Elevated risk	2 (8.7%)
Prohibited peers	1 (4.3%)
Total	**42**

4.2 Risk-band analysis

In this section, we present comparisons to indicate whether any statistically significant associations exist between the numbers of observed and estimated recidivists, both for proven sexual reconvictions and any relevant sanctions (i.e., reconviction, recall to prison, and breach of licence conditions). RM2000/s is designed to predict proven reoffending but is not designed to predict other sanctions. Our data, however, suggest that many of these sanctions appear to have been serious enough to warrant law enforcement investigation and our follow-up duration will not have been long enough for those to have been proven.

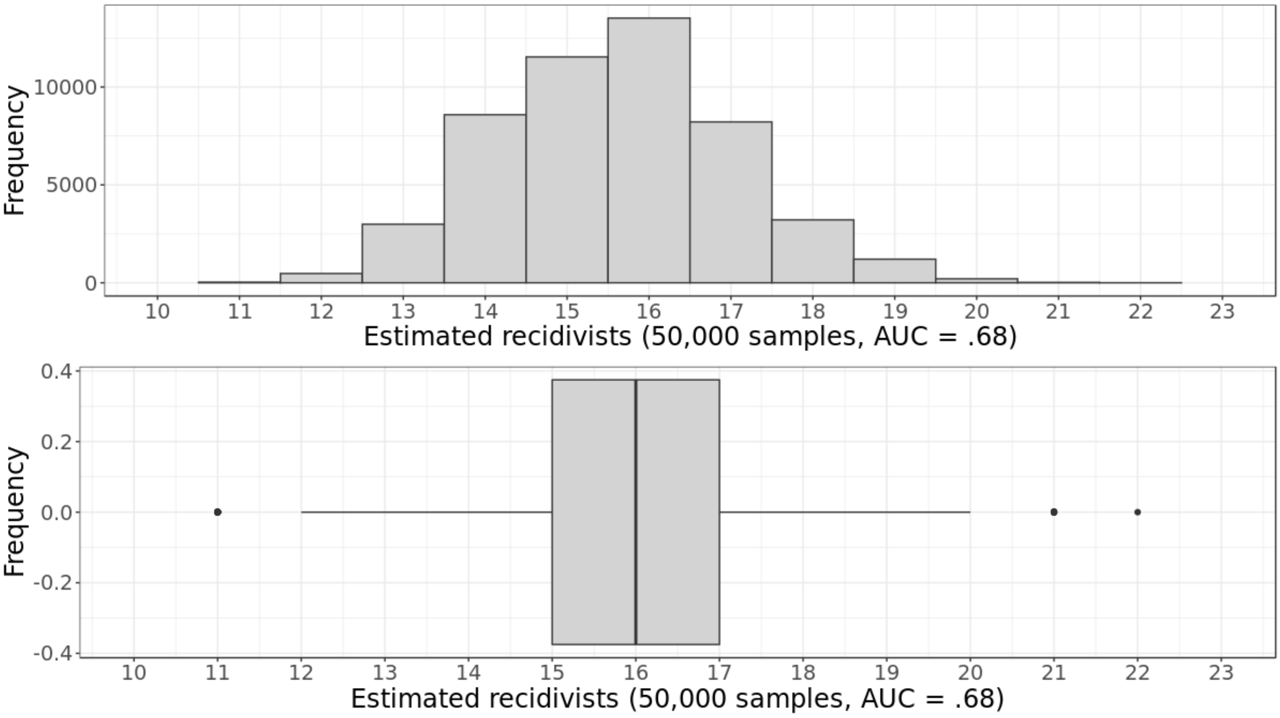
Table 6: Contingency tables for a standard 1-year RBA using proven sexual reconvictions and any relevant sanctions (combined) versus risk band estimates

RM2000/s risk category	**HSP sample frequency**	**RM2000/s predicted category rate of reconviction**	**Predicted number of reoffenders[1]**
Low	5	0.0%	0
Medium	32	4.2%	1
High	50	11.4%	6
Very high	25	36.8%	9
Total			**16**

[1] Rounded to the nearest whole number.

Our methodological analyses found that RM2000/s, when working perfectly, predicts 16 recidivists for our HSP sample (see Table 6), representing a comparative overall predicted proven reconviction rate of 14.3%.[2] Further analysis (using bootstrapping) also estimated that, excluding a small number of outliers, the plausible margin of error for measurement error ranges from 12 and 20 recidivists, for which 16 recidivists was the median. These predicted values represent predicted proven reconviction rates of 10.7% when RM2000/s is underestimating risk to 17.9% when RM2000/s is overestimating risk (see Figure 3).

Figure 3: Histograms and boxplots for the bootstrapped estimate of RM2000/s measurement error given the parameters of the sample and an AUC of .68



Two sets of three chi-squared tests of association were conducted. These tested whether our observed frequencies of reoffending differ statistically significantly from the predicted frequencies in our three scenarios. The first set of chi-squared analyses compared the observed rate of proven sexual reconvictions (8) to the each of the lower (12), median (16), and upper (20) values in the predicted range (Table 7). The second set compared the observed rate of any relevant sanctions (31) to the each of the lower (12), median (16), and upper (20) inter-quartile range values in the predicted distribution (Table 8).

[2] 16 ÷ 112 = 0.143 (14.3%)

Table 7: Contingency tables for a 1-year RBA with observed proven sexual reconvictions as the outcome versus lower, median, and upper predicted reconvictions

	Recidivism	No recidivism	Total
Lower range threshold			
Observed	8	104	112
Predicted	12	100	112
χ^2 (1) = 0.49, *p* = .482, FET = .483, OR = 0.64, *r* = .07.			
Median			
Observed	8	104	112
Predicted	16	96	112
χ2 (1) = 2.29, p = .131, FET = .129, OR = 0.46, r = .14.			
Upper range threshold			
Observed	8	104	112
Predicted	20	92	112
χ^2 (1) = 4.94, *p* = .026, FET = .025, OR = 0.36, *r* = .21.			

Table 8: Contingency tables for a 1-year RBA with any observed relevant sanctions as the outcome versus lower, median, and upper predicted reconvictions

	Recidivism	No recidivism	Total
Lower range threshold			
Observed	31	81	112
Predicted	12	100	112
χ^2 (1) = 9.32, *p* = .002, FET = .002, OR = 3.17, *r* = .29.			
Median			
Observed	31	81	112
Predicted	16	96	112
χ2 (1) = 5.28, p = .022, FET = .021, OR = 2.29, r = .21.			
Upper range threshold			
Observed	31	81	112
Predicted	20	92	112
χ^2 (1) = 2.54, *p* = .111, FET = .111, OR = 1.76, *r* = .15.			

Because we are conducting multiple statistical tests simultaneously increasing the probability of finding a "false positive" result (i.e., obtaining a statistically significant result by chance alone). To compensate, a Bonferroni correction was applied to each set of tests, which adjusts the original threshold for statistical significance ($p < 0.05$) to account for the number of simultaneous tests being run: 3 tests for each outcome (0.05 ÷ 3: $p < 0.0167$). For an observed frequency to be considered statistically significantly different from a predicted value the chi-squared test must reflect a probability lower than 0.0167.[3]

The results of the chi-squared tests are summarised in Table 9 The proportional difference between the observed and predicted values was only statistically significant for any relevant sections and when comparing the observed rate to the lower boundary of the predicted range. In short, only in a scenario where (a) proven reconvictions and breaches and recalls were combined and (b) when RM2000/s was assumed to be underestimating risk, did our observed rate fall outside of our predicted range of expectations for this sample. The odds ratio indicated that under these circumstances observed reoffending was approximately three time higher than predicted by RM2000/s. None of the remaining comparisons were statistically significant.

Table 9: Summary of outcomes from six chi-squared tests and comparisons to Bonferroni adjusted alpha levels to judge statistical significance.

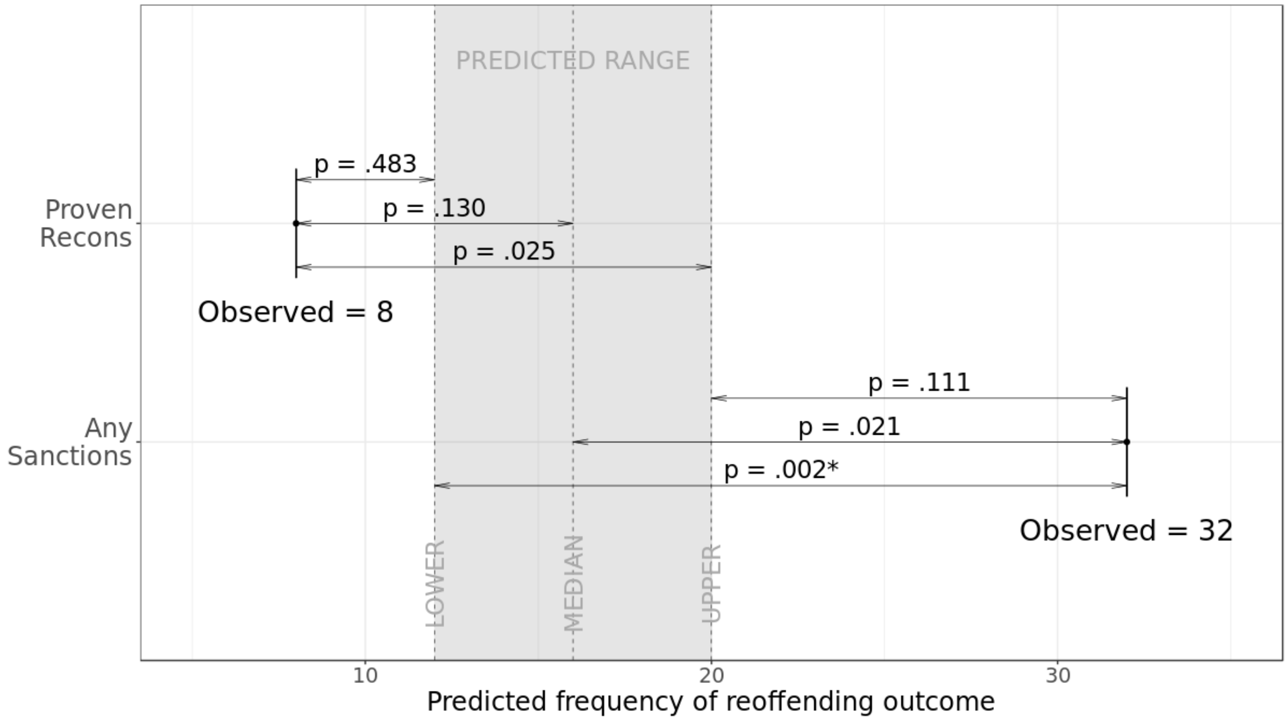
Comparison	χ^2	Odds ratio	*p* (Fisher's Exact Test)	*Bonferroni target α*	Statistically significant?	*r (effect size)*
Proven sexual reconvictions						
Lower	0.49	0.64	0.483	0.0167	No	.07
Median	2.29	0.43	0.130	0.0167	No	.14
Upper	4.94	0.36	0.025	0.0167	No	.21
Any relevant sanctions						
Lower	9.32	3.17	0.002	0.0167	Yes	.29
Median	5.28	2.29	0.021	0.0167	No	.21
Upper	2.54	1.76	0.111	0.0167	No	.15

This indicates that in all other scenarios apart from where proven reconvictions and breaches and recalls were combined (regardless of whether we included or excluded other relevant sanctions or if we assumed RM2000/s is under- or over-estimating risk) the observed reoffending rates fell within the range of expected values: no change, neither

[3] The smaller the *p*-value of a statistical test the smaller the likelihood that the researcher would see an outcome as large as the one observed if the null hypothesis that there is no real effect is true.

higher than expected nor lower than predicted. In simple terms, as illustrated in Figure 4, the observed number or proven reconvictions was lower than the lowest edge of the predicted range, but not by far enough to be considered statistically significant. At the other end, observed number of relevant sanctions was higher than the upper edge of the predicted range, but not by far enough to be considered statistically significant.

Figure 4: An illustration of outcomes from the six chi-squared analyses, indicating the respective disparity between the observed and RM2000/s predicted values.



* Statistically significant to $p < 0.0167$.

4.3 Other relevant variables

This section describes the findings from the three other sources of relevant variables that were explored for whether they were predictive of all relevant sanctions: (1) other criminogenic variables via cumulative scores for OASys categories; (2) paedophilic/hebephilic interests via SSPI-2 scores; (3) learning disabilities and challenges via scores on an LDC screening test. According to the LST, 30.4% of the sample were judged to have a score that indicated potential for LDC (but which does not confirm it) with an average score of 2.0 (SD = 2.4) out of a possible maximum of 8.

Average score on the SSPI-2 was 3.2 (SD = 1.4) out of a possible maximum of 5, with 48.2% judged to have “high” scores (9.8% low and 42.0% medium) based on trichotomous cut-score ranges established by Helmus et al. (2015), which differentiate between groups with: (1) relatively low scores and a reconviction rate of 2.6%; (2) relatively higher scores and a reconviction rate of 8.5%; and (3) the highest scores and a recidivism rate of 19.3%.

The resulting logistic regression model for the five OASys sections as predictors of any relevant sanctions (not proven reconvictions) is presented in Table 10. Note that mean scores on these sections can be found in Table 3. Only a lack of permanent accommodation was statistically significantly predictive of recidivism. Increased scores on the other OASys sections or additional measures were not predictive of relevant sanctions in this model.

Table 10: Logistic regression coefficients and odds ratios for OASys sections as predictors of known recidivism

			95% confidence intervals for odds ratio (OR)		
	***B* (SE)**	***p***	**Lower**	**OR**	**Upper**
Constant	-2.47 (0.94)				
Accommodation	0.18 (0.08)	.020*	1.03	1.19	1.39
Relationships	0.03 (0.17)	.877	0.74	1.03	1.42
Think/behaviour	0.17 (0.14)	.234	0.90	1.18	1.57
Attitudes	0.13 (0.13)	.315	0.88	1.14	1.47
Paedophilia	0.06 (0.11)	.580	0.85	1.06	1.32
LDC	-0.18 (0.16)	.267	0.61	0.84	1.15

* Statistically significant, $p < .05$

Note: AIC = 133.3. R^2 = .13 (Hosmer-Lemeshow), .15 (Cox-Snell), .21 (Nagelkerke). Model X^2 (5) = 17.53, p = .008. AUC = .76.

5. Discussion

This study sought to explore the criminal justice outcomes experienced by a sample of individuals with a sexual paraphilia, taken from the HSP programme caseload, after release from prison. The study explored how the actual frequency of recidivistic outcomes (reconvictions or other relevant sanctions) compared to predicted frequencies, and whether relevant variables could broadly predict recidivistic outcomes. It is important to reiterate from the outset that this constitutes an exploratory study. It is intended to provide descriptive and indicative data on the post-release outcomes for a small subset of participants convicted of sexual offences who also have paraphilia. This study should not be considered as an evaluation of the effectiveness of HSP, nor should the findings be used to draw conclusions about any association between completing HSP and subsequent reconviction or recall rates.

5.1 Results

The results indicated that of this sample of individuals who have been assessed as meeting the criteria for a paraphilic disorder, almost one-third (30.4%) of those released into the community after sentence for an average of 16 months, received some form of criminal justice penalty. The proven sexual reconviction rate was 7.1% (including one conviction not technically classified as sexual but that included public masturbation). Another 20% of the sample was recalled to prison for breaches of conditions of release. Broad qualitative analysis of these recalls indicate that most recalls were for possessing undisclosed electronic devices (many of which were found to contain concerning material), unsupervised access to children, possession of pornographic material (legal-but-prohibited and/or potentially illegal), and breaking travel or curfew rules.

When we compared the observed number of proven HSP reconvictions to a range of plausible predicted frequencies, we found that proven reconvictions alone were not statistically significantly lower than the lowermost value in that range. We also found that the combined number of reconvictions and breaches/recalls were not significantly higher than the uppermost value in that range. Because the “true” predicted value could lie anywhere within that range it is plausible to conclude that the observed frequency of post-release criminal sanctions did not appear to fall outside of predictions by RM2000/s (i.e. no change from the predicted rate). Therefore, we might reasonably conclude that our paraphilic sample reoffended at a rate that would be expected from the wider population of individuals with sexual convictions (including those without paraphilia). The only scenario in which observed frequencies fell outside of this range is when we combined all relevant criminal sanctions, which we know is likely to be an overestimation of true reoffending, and if RM2000/s was underestimating risk to high degree. Both are “worst case” scenarios.

As noted in the methods section, it was difficult to establish the known rate of criminal justice sanctions due to the short-term follow-up (less than two years). Firstly, the known time lag in PNC data means that we can be relatively confident that proven reconvictions is an under-estimation of the true number of reconvictions. Secondly, the OASys records relating to some, but not all, of the cases of recall to prison indicated that further investigations into potential new crimes were also being undertaken. This means we can be relatively confident that (a) the numbers of proven reconvictions in this sample would increase given time and (b) the number of recalls is an over-estimation of the true rate of reconvictions since not all of them truly represent new criminal behaviour. Consequently, it may not be unreasonable to suggest that proven sexual reconviction rate will regress towards the RM2000/s expected frequency. This could later be found to represent a relatively accurate 1-year frequency once all the data is available.

The study also found that a substantial minority of the sample (30%) received a high score on a screen for LDC – albeit, who may not necessarily have been found to demonstrate LDC – and that almost half scored highly on a screen for paedophilic or hebephilic interests. Although LDC and paedo/hebephilia were not found to be statistically significantly associated with recidivism (reconviction and recall combined), these findings suggest LDC may be an area of future interest. It is notable given the potential for LDC in this sample that HSP has been redeveloped to cater to the LDC population to an even greater degree than the first iteration. Given that paedophilic interest is typically shown to be a good predictor of sexual recidivism for this population (see McPhail et al., 2018), it is notable that neither SSPI scores or the trichotomous cut-off groups suggested by Helmus et al. (2017) were found to be predictive of known recidivism in this HSP sample, despite 50% of the sample being found to be in a high-scoring category.

There are several reasons why this might be case. Firstly, due to the small number of occurrences of proven sexual reconvictions we were only able to explore whether they were predictive of any criminal justice outcome. This included recalls to prison and breaches of licence condition, not new contact sexual reconvictions. The data presented here is instead more likely to represent an indication of the individual's ability to adhere to the conditions of release (i.e., their general criminality), rather than to new instances of sexual contact with children. The existing research has focused on paedophilic interest as a predictor of sexual reconvictions, not recall or breach, so it is perhaps unsurprising that we did not find an association. Secondly, although the SSPI-2 has been found to be a good proxy of physical measures of sexual interest, such as penile plethysmography (PPG), it remains a proxy for alternative more robust measures. It has also not yet been validated with a U.K. sample. However, it is important to note that any lack of findings might well also be simply explained by the short follow-ups, roughly estimated normative rates, and small sample, not the relative predictive ability of a sexual interest in children.

The one positive predictor we identified for relevant sanctions in our sample was the OASys section related to problems with access to permanent accommodation. One potential reason for this may be that issues like LDC or lack of accommodation – above others – can make it more difficult for those individuals to adhere to their conditions of release. For example, licence conditions can be extensive and written in legal prose and typically include restrictions on location, movement, and association, which may increase the likelihood that they are violated by individuals with LDC and/or permanent accommodation issues. This notion should, however, be clearly caveated with a note that this was an analysis of a small dataset (albeit with sufficient EPVs) and consequently there may have been over-fitting of the regression models that should be considered for further research attention. It would be beneficial to explore the relationships between LDC, accommodation, and reconviction/recall further in larger paraphilic samples.

5.2 Conclusions

In conclusion, these data indicate that the levels of recidivism exhibited by this sample of individuals assessed as having paraphilic interests does not fall outside of the bounds of plausible expectation given their profile on risk assessments. Differences between the observed and expected numbers of recidivists cannot be causally attributed to participation on the treatment programme they were sampled from (HSP), due to the methodological limits of the analyses conducted here. However, this study found that this sample did not appear to commit further sexual crimes at a rate above that which would be expected from a normative non-paraphilic population.

This study effectively compared this highly paraphilic sample with the expected rates in the wider population of individuals with sexual convictions. The finding that this paraphilic sample did not appear to reoffend at a higher rate than would be expected from a non-paraphilic population is notable. Desistance is typically viewed as a gradual journey (see, for example, Paternoster & Bushway, 2008; Serin & Lloyd, 2009) and these are individuals who face ongoing challenges to living a non-criminal life due to their interest in harmful sex. Although lapses are expected, some that may result in new convictions or sanctions, success in staying crime-free is a continual effort requiring considerable application of the forms of sexual self-regulation and social support that are the focus of HSP.

Nonetheless, these findings should be viewed considering the short follow-up times, small sample, the exploratory nature of the methods used, and the possibility that paraphilic interests may have been present in the Thornton (2007) normative sample. We reiterate that the methods used in this study do not represent a high level of statistical rigour and used a variety of hypothetical and estimated values to provide some form of wider context for the observed recidivism statistics.

Finally, again despite methodological limitations clearly preventing us from over-selling their utility, the findings that issues relating to post-release accommodation might predict who is at greater risk of breaching their licence conditions could provide insights that can improve the future post-release management in this sub-population. Further analysis should be considered once larger samples become available.

References

Andrews, D. A., & Bonta, J. (2010). The psychology of criminal conduct (5th ed.). Cincinnati, OH: Anderson.

Barnett, G. D., Wakeling, H. C., & Howard, P. D. (2010). An examination of the predictive validity of the Risk Matrix 2000 in England and Wales. Sexual Abuse: A Journal of Research & Treatment, 22(4), 443–470.

Berlin, F. S. (2019). Paraphilic disorders: A better understanding. Current Psychiatry, 16(4), 22–28.

Berrar, D., & Flach, P. (2012). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Briefings in Bioinformatics, 13(1), 83–97.

Carter, A., & Mann, R. E. (2016). Organizing principles for an integrated model of change for the treatment of sexual offending. In D. Boer (Ed.), The Wiley handbook on the theories, assessment and treatment of sexual offending (pp. 359–381). London, UK: Wiley-Blackwell.

Cooke, D. J., & Michie, C. (2014). The generalizability of the Risk Matrix 2000: On model shrinkage and the misinterpretation of the area under the curve. Journal of Threat Assessment and Management, 1(1), 42.

Drummond, C., Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65, 95–130.

Finkelhor, D., & Jones, L. (2006). Why have child maltreatment and child victimization declined? Journal of social issues, 62(4), 685–716.

Finkelhor, D., Saito, K., & Jones, L. (2015). Updated Trends in Child Maltreatment, 2013. Durham, NH: Crimes against Children Research Center.

Greenland S. (1989). Modeling and variable selection in epidemiologic analysis. American Journal of Public Health, 79(3), 340–349.

Grubin, D. (2008). Validation of Risk Matrix 2000 for use in Scotland. Report prepared for the Risk Management Authority. Available from http://www.nomsintranet.org.uk/roh/official-documents/Grubin_2008.pdf

Hand, D. J., & Anagnostopoulos, C. (2014). A better beta for the H-measure of classification performance. Pattern Recognition Letters, 40, 41–46.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. Psychological assessment, 21(1), 1–21.

Harrell, F. E., Lee, K. L., Mark, D. B. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4), 361–87.

Helmus, M. L., Babchishin, K. M., & Hanson, R. K. (2013). The predictive accuracy of the Risk Matrix 2000: A meta-analysis. Sexual Offender Treatment, 8(2), 1–20.

Helmus, M. L., Ó Ciardha, C., & Seto, M. C. (2015). The Screening Scale for Pedophilic Interests (SSPI): construct, predictive, and incremental validity. Law and Human Behavior, 39(1), 35.

Holoyda B. J., & Kellaher, D. C. (2016). The biological treatment of paraphilic disorders: An updated review. Current Psychiatry Reports, 18–19.

Howard, P. D., & Barnett, G. D. (2015). Development of a new sexual reoffending predictor. In R. Moore (Ed.), A compendium of research into the Offender Assessment System (OASys): 2009-2013 (Ministry of Justice Analytical Series) (pp. 209–238).

Laws, D. R. (2009). Penile plethysmography: Strengths, limitations, innovations. In D. Thornton & D. R. Laws (Eds.), Cognitive approaches to the assessment of sexual interest in sexual offenders (pp. 7–29). London: Wiley-Blackwell.

Lindsay, W. R. (2011). People with intellectual disability who offend or are involved with the criminal justice system. Current Opinion in Psychiatry, 24(5), 377–381.

Lindsay, W. R., Hastings, R. P., Beech, A. R. (2011). Forensic research in offenders with intellectual & developmental disabilities 1: prevalence and risk assessment. Psychology, Crime & Law, 17(1), 3–7.

Mann, R.E., Hanson, R.K., & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. Sexual Abuse: A Journal of Research & Treatment, 22, 172–190.

McClish, D.K. (1989). Analyzing a portion of the ROC curve. Medical Decision Making, 9(3), 190–195.

McPhail, I. V., Olver, M. E., Brouillette-Alarie, S., Looman, J. (2018). Taxometric analysis of the latent structure of pedophilic interest. Archives of Sexual Behavior, 47(12), 2223–2240.

Ministry of Justice (2020) HMPPS Annual Digest: April 2019 to March 2020. Ministry of Justice. https://www.gov.uk/government/statistics/hmpps-annual-digest-april-2019-to-march-2020

Moore, R. (2015). A compendium of research into the Offender Assessment System (OASys): 2009-2013. London, U.K. Ministry of Justice Analytical Series.

Paternoster, R., & Bushway, S. (2008). Desistance and the feared self: Toward an identity theory of criminal desistance. J. Crim. L. & Criminology, 99, 1103.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), 1373–1379.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law and Human Behavior, 29(5), 615–620.

Serin, R. C., & Lloyd, C. D. (2009). Examining the process of offender change: The transition to crime desistance. Psychology, Crime & Law, 15(4), 347–364.

Seto, M. C., Stephens, S., Lalumière, M. L., & Cantor, J. M. (2015). The revised Screening Scale for Pedophilic Interests (SSPI–2): Development and criterion-related validation. Sexual Abuse, 29(7), 619–635.

Stephens, S., Seto, M. C., Cantor, J. M., & Lalumuiere, M. L. (2019). The revised Screening Scale for Pedophilic Interests (SSPI-2) may be a measure of pedohebephilia. The Journal of Sexual Medicine, 16(1), 1655–1663.

Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., et al. (2003). Distinguishing and combining risks for sexual and violent recidivism. Annals of New York Academy of Sciences, 989, 225–235.

Thornton, D. (2007, unpublished). Scoring guide for Risk Matrix 2000.9/SVA. Retrieved August 13, 2018, from https://www.birmingham.ac.uk/documents/college-les/psych/rm2000scoringinstructions.pdf

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology, 165(6), 710–718.

Winder, B., Lievesley, R., Elliott, H., Hocken, K., Faulkner, J., Norman, C., & Kaul, A. (2018). Evaluation of the use of pharmacological treatment with prisoners experiencing high levels of hypersexual disorder. The Journal of Forensic Psychiatry & Psychology, 29, 53–71.

Winder, B., Federoff, J. P., Grubin, D., Klapilova, K., Kamenskov, M., Tucker, D. et al. (2019). International Review of Psychiatry, 31(2), 159–168.

Woodrow, A. C., & Bright, D. A. (2011). Effectiveness of a sex offender treatment programme: A risk band analysis. International Journal of Offender Therapy and Comparative Criminology, 55(1), 43–55.

Appendix A
Methodological procedure

Estimating expected recidivism rates

Microsoft Excel contains functions that can be used to derive linear, exponential, and logarithmic trend from known values using the regression coefficient to predict growth (forward projections) or decay (backward projections) for one or more new values using the algebraic equation *y = (c * LN(x)) + b*. MS Excel can also be used to generate similar exponential (GROWTH) and linear (TREND) growth and decay using known values. The GROWTH function derives its estimated values from the equation *y = b * m ^ x*. The TREND function derives its estimated values from the equation *y = m * x + b*. Each of the MS Excel functions calculates the results of these algebraic equations automatically.

Each of these functions was used to estimate decay from published normative rates published in Thornton (2007) to estimate 1-year rates. It is important to note that these functions require multiple known values from different time points in the same dataset (e.g., Thornton presents rates at 5, 10, and 15 years) to estimate the trend lines, which means that we are not able to use figures from other, more recent RM2000/s studies where they only report rates at a single time point. Each of these MS Excel functions serves to take a column of known values (i.e., a list of recidivism rates) with a respective column of time points to which those known values refer (i.e., the number of follow-up years for which each rate applies) and to use the algebraic equations to estimate a new value, given a new time point.

Adjusting for risk assessment measurement error

Barnett et al. (2010) validated the RM2000/s using 2,755 U.K. sex offenders and reported an AUC value of .68 (95% CI = [.63, .73]). Two meta-analyses combining data from international samples (the latter including the Barnett et al. study) reported AUCs – based on Cohen’s d (see Rice & Harris, 2005) – of .68 (95% CI = [.65, .71]) (Hanson & Morton-Bourgon, 2009: n = 2,755) and .74 (95% CI = [.67, .81]) (Helmus et al., 2013: n = 10,644).

For RM2000/s a bootstrapping process was created that simulates that measurement error and its effect on the distribution of predicted recidivists. This process was designed to simulate what happens if we re-assign individuals in the HSP sample to a risk band with a reasonable amount of measurement error accounted for to produce a range of values for the total number of individuals with reoffences that RM2000/s would predict, rather than one single value. This range of values represents an estimate of the extent to which that predicted total might increase or decrease when measurement error[4] is present.

The bootstrapping process that was created does not change the raw content of the data. It generates a pair of new artificial samples, each the same size as the HSP sample: (1) Sample A, which assumes the probability of being assigned to each risk level is the same as it was in the HSP sample (e.g., the probability of being assigned to the "Low" risk category is 5 ÷ 112 = 0.045 or 4.5%); and (2) Sample B, which assumes accuracy was based on chance effects alone (e.g., the probability of being assigned to any of the four risk categories is the same: 1 ÷ 4 categories = 0.25 or 25%).

The function then chooses to draw, for a sample of a given n, either from Sample A or B based on the probability reflected by the AUC value of the RM2000/s (a .68 probability of drawing from Sample A rather than Sample B). This function was replicated for 50,000 samples to provide an estimate of the underlying distribution of classifications and the effect of that distribution to generate a plausible range of potential estimates for comparison.

Other predictors

The potential predictive validity of other relevant assessments was explored with the aim to provide some context for the RBA findings, via a binary logistic regression model. It has been noted that logistic regression models can generate biased estimates, unreliable confidence intervals, and convergence problems as the number of predictor variables in the model approaches the number of events (Peduzzi et al., 1996). Some have suggested an event-per-variable (EPV) ratio of 10 for regression models to correct for the potential for type II errors (Greenland, 1989; Harrell et al., 1996). Vittinghoff and McCulloch (2007), however, found that the rule of thumb of an EPV of 10 or more is not "a well-defined bright line" and that models with EPVs between 5–9 were comparable with those with EPVs between 10–16. They concluded that for significant findings with 5–9 EPV, "only a minor

[4] A test of the bootstrap function found that AUC values of 0.68 and 0.74 both generate the same broad median and upper/lower 1.5 inter-quartile range (IQR) estimates.

degree of extra caution is warranted, in particular for plausible and highly significant associations hypothesized a priori" (p. 717).

Given 31 occurrences of reconvictions and breaches combined, the remaining 8 variables would generate an event-per-variable ratio (EPV) of only 3.88. This clearly limited our ability to also include contrasts in any regression model. In total, 6 variables were entered into the binary logistic regression model, generating a final EPV of 5.17.

OASys is a national automated system for assessing the risk and needs of a prisoner or probationer at various points throughout their sentence. The main body of an OASys assessment consists of 12 sections, 10 of which are dynamic criminogenic variables. It was decided to include in the model any section for which the mean value in the sample exceeded the stated threshold that would "require attention" in an OASys assessment and with the largest standard deviation relative to the scoring range.[5] Five variables exceeded their threshold (see Table A1) and with the ability to choose a maximum of four OASys predictors to maintain an EPV greater than five, the four with the largest standard deviation were included in the model.

Table A1: Mean and standard deviations for included OASys section scores for the sample and the corresponding OASys threshold for concern

	Mean (SD)	OASys threshold
Accommodation	3.5 (3.4)	2
Relationships	3.6 (1.6)	2
Thinking/behaviour	5.2 (1.9)	4
Attitudes	2.9 (2.0)	2

Two researchers calculated scores for each individual in the sample on the Screening Scale for Pedophilic Interests Version 2 (SSPI-2: Seto et al., 2015). The SSPI-2 is the most recent version of a tool described as a brief actuarial measure of pedohebephilic sexual interest, a sexual arousal to children as is assessed by phallometric testing[6] (Seto et al., 2015; Stevens et al., 2019). Scores have been found to positively correlate with

[5] It's unlikely that a variable with a small standard deviation will be able to predict category membership since it suggests that many of the sample obtained a very similar score.

[6] Phallometric tests measure changes in penile circumference or volume while test participants are presented with stimuli depicting different ages, sexes, and sexual activities (Laws, 2009).

phallometric scores and the SSPI-2 represents a useful proxy where phallometric tests are not possible as a structured method of assessing sexual interest in children based on offence-related variables (Seto et al., 2015).

The HSP eligibility assessment also included the administration of the Learning Screen Tool (LST). The LST is described in the guidance notes as a measure to identify individuals with sexual convictions who require further assessment of IQ for allocation onto an adapted suite of programmes for individuals with sexual convictions. As such, it is not an IQ test, but a low score is considered indicative of low IQ and, in practice, can be followed up by a validated test of cognitive ability depending on risk level or custodial/community context. HSP accommodates individuals with a range of learning abilities including those with LDC.

Inter-rater agreement was assessed for the scoring of these measures for the purposes of this analysis. However, it should be noted that this did not account for inter-rater reliability of scoring for the original individual OASys reports. The first author scored a random selection of 10% (n = 12) of the cases. This generated an agreement score of 100% on the LST, likely because those overall scores are the product of a series of single numerical values taken directly from OASys reports. Scoring the SSPI-2 involved more subjective judgments based using OASys file information, and generated an agreement score of 66.7% (8 out of 12 cases) for raw scores. Of the four cases where disagreements occurred, two cases differed by only 1 point and two differed by 2 points (on a scoring range of 0–5). These disagreements, however, only resulted in different trichotomous classifications (see Helmus et al., 2015) in two cases, with 83.3% of classifications being the same. Statistical associations between proven reconvictions/other relevant sanctions and categorical classifications on the LST and the SSPI-2 were tested using chi-square analyses with Fisher’s exact test and a step-down Holm-Bonferroni adjustment.