Candidate Number: POL1535

Sheena Urwin

Wolfson College

Supervisor: Geoffrey Barnes

ALGORITHMIC FORECASTING OF OFFENDER DANGEROUSNESS FOR POLICE CUSTODY OFFICERS: AN ASSESSMENT OF ACCURACY FOR THE DURHAM CONSTABULARY MODEL

Submitted in part fulfilment of the requirements for the Master's Degree in Applied Criminology and Police Management

December 2016

Abstract

Algorithmic forecasts of future dangerousness, conducted at the gateway to the Criminal Justice System (CJS), could minimise harm in communities by providing evidence-based decision support to UK police custody officers. Forecasting dangerousness is not only forecasting crime will re-occur but what kinds of crime will occur. Rarely, is a structured, consistent, assessment of the potential future harm considered. Indeed, rarer still is such an assessment made at the gateway to the CJS, yet this may be pivotal to consistent decision making. Assessing future harm within an evidence based framework can promote consistency of decision making, enabling targeted interventions to rigorously test what works in preventing harm and reducing reoffending (Berk et al., 2009; Barnes and Hyatt 2012; Sherman 2012; Sherman 2013; Hyatt and Barnes 2014; Neyroud, 2015).

This research was conducted in Durham Constabulary and reviews the algorithmic forecasting of offender dangerousness utilising a forecasting model called the Harm Assessment Risk Tool (HART) used at the entry point to the CJS. The model was constructed using relatively new machine learning techniques. The model is built using a random forest machine learning technique which offers features such as an ability to balances different types of errors which is desirable in areas of consequential decision making. No such model, however, can be trusted until it is validated with data that were not used in the construction of the model and this research will present the necessary validation of HAT. Accuracy level fall in validation, and the findings in this research present no exception to that rule.

HART forecasts, over a two-year period, whether a suspect is likely to commit a serious offence over the next two years (High risk), any offence (Moderate risk) or no offences (Low risk). Utilising an independent dataset, the 2013 population of 14,882 custody events in Durham Constabulary were used in this validation study. The random

forest procedure has an ability to balance different types of errors, by using cost ratios decided by the organisation.

The worst-case scenario in using such a decision support tool is that a suspect, forecast as low risk, turns out to be high risk; referred to as a high-risk false negative error. HART, in balancing the errors, was able to minimise harm. The model ensured that the error rate for a high-risk false negative result was 2%. Therefore, Durham Constabulary can be 98% sure the worst-case scenario will not happen, should custody officers follow the decision support tool forecast. The model became more conservative in its forecast of high harm in order to minimise the worst-case scenario of serious harm in our communities.

Acknowledgements

I would like to thank Chief Constable Mick Barton, QPM who tapped me on the shoulder one day and offered me the most amazing opportunity. I am truly grateful for the experience and learning I have taken from Cambridge University and I thank you wholeheartedly for your support.

I am indebted to Cambridge University staff for their guidance and support. In year 1 Dr Heather Strang provided me with help and advice which kept me focussed for which I am grateful. My warmest thanks are reserved for Dr Geoffrey Barnes, my thesis supervisor, for his support, guidance, and encouragement. Above all though, his humour which made the process a lot easier - is now the right time for a mariachi band? Thanks, are also reserved for Sainsbury's Monument Trust whose funding enabled the building of HART.

I extend thanks to colleagues Gillian Routledge, John Cooper, Paul Guy, the Checkpoint team, and staff within the CET Case building team without whose support implementation of HART, and this research, would have been impossible.

I am deeply grateful to colleagues who have supported me through the process and have tolerated my distractions and absences, special thanks to Gillian Chambers, Lisa Holliday, Terri Raine and Paula Barrasford.

My gratitude is most strong for my family, who have endured 18 months of absences and my preoccupation with this study. Thank you for giving me the space and time to complete this study. Thank you to Rose and Dave McGee who taught me how to work hard and equipped me with the willpower and determination to succeed. Michael, my husband, whose belief in me has carried me through and shown me how lucky I am.

Lastly, but certainly not least, Charlotte and Emily, your turn next ☺ - hard work pays off.

Table of Contents

3.0 List of Figures

4.0 List of Tables

5.0 List of Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
ANOVA	Analysis of Variance
CJS	Criminal Justice System
CPS	Crown Prosecution Service
CART	Classification and Regression Trees
HART	Harm Assessment Risk Tool
HMIC	Her Majesty's Inspectorate of Constabulary
IT	Information Technology
LSI-R	Level of Service Inventory – Revised
OOB	Out of Bag
PNC	Police National Computer
PND	Police National Database
Tukey HSD	Tukey Honestly Significantly Different
UK	United Kingdom
US	United States

6.0 Introduction

The most dangerous offenders could be treated differently from lower risk offenders to promote their desistance from crime (Sherman, 2012). If offenders are to be treated differently based on their perceived dangerousness however, how can the police and the public be sure that offender dangerousness is being identified in the most accurate, and consistent manner? Identifying offenders and ensuring that appropriate treatment is targeted to the appropriate risk level in order to reduce reoffending and prevent further crime is in line with Peelian principles (Home Office, 2012).

Policing in England and Wales over the last six years has experienced substantial reductions in central government funding, totalling £1.7 billion (HMIC, 2016). This has been a difficult time for policing, with police forces initially focused on personnel levels as a means, in the short term, of meeting financial challenges. The continual need to improve, in terms of effectiveness and efficiency, remains of paramount importance to all police forces and criminal justice agencies.

Having a grasp of future demand and risk will ensure a force is prepared for the challenges ahead. Appropriate evidence-based decisions are made when demand is understood, particularly in relation to future offending. Crime and its harms are concentrated among a small number of the most dangerous offenders which Sherman (2007) refers to as the 'Power Few'. Targeting those high harm offenders will assist a police force to understand where the limited police resources could be deployed and how, going forward, the capabilities of the force may need to change (HMIC, 2016).

A large amount of research exists asserting that statistical forecasting is more generally accurate than intuitive clinical judgement (Meehl, 1954; Dawes et al., 1989; Grove et al., 2000; Ægisdóttir et al., 2006; Kahneman and Klein, 2009; Kahneman, 2011). Despite this, Neyroud (2015) highlights that the police entry point to the CJS

generally uses a clinical model of forecasting risk rather than modern actuarial or machine learning methods. In England and Wales, that decision to bail (conditionally or unconditionally), hold in custody, prosecute, or divert from the Criminal Justice System (CJS) with an out of court disposal, all rests with the custody officer. The custody officer in making the decision takes account of the evidence against the suspect and the public interest (Neyroud, 2015). Clearly the decision can have a large impact on the suspect in terms of prosecution, punishment and the effects on future employment, but also on the ability of the police to properly prevent crime and minimise harm in communities. The decision support provided to custody officers in England and Wales relies on a police-developed gravity matrix which is linked to the Magistrates Sentencing guidelines. The matrix does not focus on an assessment of future dangerousness but looks only at the nature of the current, presenting offence. Accurately forecasting future dangerousness would allow police forces to consistently make appropriate decisions at the gateway to the CJS (Sherman, 2011; Neyroud and Sherman, 2012; Neyroud, 2015).

Advanced machine learning statistical methods can be utilised to take account of large amounts of data readily available in police forces. Random forest forecasting is a machine learning statistical method that can account for errors in a way that traditional forecasting methods do not. This makes the method desirable in a police setting. Particularly, where the consequences of errors can be very serious indeed (Berk et al., 2009; Barnes and Hyatt, 2012). Random forest forecasting in a CJS setting has not taken place in a police post arrest environment which sits at the entry point to the CJS. This research will contribute to the growing body of research in this area (Berk et al., 2009; Sherman, 2011; Sherman and Neyroud, 2012; Barnes and Hyatt, 2012; Berk et al., 2016).

6.1 The Model

Working with Dr Barnes at the University of Cambridge, Durham Constabulary has developed a random forest forecasting harm assessment risk tool (HART). The model predicts offender dangerousness and classifies offenders into high, moderate, or low risk groups. The groups are forecast in terms of their likelihood over the next 24 months of committing a serious offence (High), non-serious offence (Moderate) or no offence (Low). The forecast is made when the offender is placed before the custody officer upon entry into the custody environment post arrest. Durham Constabulary would then utilise the forecast to support the decision they make concerning the suspect.

In classifying offenders at the entry point to the CJS, the police can tailor their decision to the dangerousness of the offender. A statistical forecasting tool that can quickly review and forecast dangerousness using vast amounts of data will lead to more consistent processing of offenders in an environment which can have a high turnover of custody officers (Neyroud, 2015). Effective forecasting can lead to effective triage, which in turn can lead to the right offenders receiving the appropriate custody disposal to support a desistance in committing crime, referred to by Sherman (2012) as Offender Desistance Policing.

6.2 Purpose and Structure

Sherman (2013) asserts to effectively target and test, the measures used by the police must be highly reliable. This study assesses the validated accuracy of the first police random forest model in England and Wales to forecast future dangerousness of offenders. To conduct the assessment of accuracy and error distribution, an independent dataset of 14,882 custody events for 2013 in Durham Constabulary were used. In using fresh independent data, actual forecasts can be compared to actual outcomes with data not used in the construction of the model. The research reviewed the unique way random forests models take account of differing error types to minimise high harm in the

community. The review of different types of high-risk errors summarised in case studies and the characteristics of the three forecast risk groups provide context to the validation. The research proceeds to then assess the extent to which intuitive clinical forecasts of custody officers agree with the algorithmic forecasting model.

The research questions are;

- What is the validated accuracy of the Durham Constabulary forecasting model using custody events from 2013 compared to 2008-2012 construction sample?
- What is the distribution of forecasting errors using the 2013 validation data compared to 2008-2012 construction sample?
- What are the descriptive characteristics of the three forecasted risk groups, as measured in the 2013 validation data?
- To what extent do the clinical forecasts of custody officers agree with the model-generated algorithmic forecasts?

With decreasing resources and the costs associated with placing offenders into the CJS, it is important to ensure the right offenders are targeted. Neyroud (2015) argues, an evidence based approach to the gateway to the CJS is critical to the effectiveness of the CJS and is 'urgently necessary' (Neyroud, 2015 p. 12). The prevention of crime, and apprehension of the offender, together with their rehabilitation and conviction as identified by HMIC (2016), 'are among the highest obligations of the state in the discharge of its duty to protect citizens' (HMIC, 2016 p 6). It is therefore incumbent upon the police to ensure that if a more accurate method of effectively targeting dangerous offenders exists to minimise high harm in the community, it should be fully explored within an evidence-based framework.

This thesis is presented in five chapters and commences with a review of the literature surrounding clinical and statistical decision making in a CJS setting. A focus is given to random forest modelling and the unique elements offered by such a model. A summary of the forecasting model is provided before describing the datasets and

research questions in the methodology chapter. The next chapter discusses each research question in turn before providing a discussion on the ethics of machine learning models and the limitations of the research. Lastly, the thesis provides overall conclusions together with policy implications and future research identified.

7.0 Literature Review

This chapter will review research relating to clinical and algorithmic forecasting before exploring what affects our judgement in clinical decisions making. An assessment of how statistical forecasting research has developed over time in a criminal justice system (CJS) setting is described, coupled with a brief description of the ethical issues presented by statistical forecasting in a CJS setting. A summary of modern methods of forecasting, in particular random forests, is presented highlighting the necessary validation process for such models.

7.1 Clinical versus Algorithmic forecasting

Accuracy of clinical decision making versus statistical decision making, has been researched over many years (Meehl, 1954; Grove et al., 2000; Ægisdóttir et al., 2006; Harcourt, 2007; Harcourt, 2010). Within the literature on the subject, many different terms are used to describe clinical judgement; a hunch, gut instinct, intuitive, clinical prediction - essentially a human forecast based on skills, knowledge, expertise and experience (Meehl, 1954; Harcourt, 2010; Kahneman, 2011). Different terms extend to those describing statistical forecasts too, utilising a statistical black box, algorithmic, actuarial, mechanical process with available data – essentially mathematically arriving at a forecast (Meehl, 1954; Dawes et al., 1989; Grove et al., 2000; Harcourt, 2007; Berk et al., 2009; Berk, 2012). The terms used to describe the outcome of the process, be it clinical or statistical, are described differently also; prediction, forecast, judgement, decision support to name a few. The process for both clinical and algorithmic decision making are arguably the same; the outcome is produced once 'data' is collected and interpreted (Dawes et al., 1989).

The clinical and statistical debate started arguably several decades ago with Meehl's work (1954). Although the forecasting methods available at that time were

relatively primitive compared to today's techniques, they demonstrated that fewer errors occur when statistical methods are used. Meehl was an American psychology professor and of the 20 studies analysed by him, 19 were found to support the mechanical method.

In contrast, other scholars believe there is room for both, dependent upon the setting and circumstances in which the prediction is made (Holt, 1958). The combination of both the data and experience of the clinician together would enable more refined judgement (Holt, 1958; Goldkamp, 1987). Further studies since this time have reinforced the view that statistical methods perform more accurately than clinical judgement.

A meta-analysis of 136 studies across many different fields (Grove et al., 2000) confirmed that statistical forecasting, of various kinds was generally as accurate, or more accurate than clinical prediction. There were just 8 studies within the analysis where clinical prediction outperformed the statistical method, which was generally due to the clinician having access to more data than the statistical model (Grove et al., 2000). A further meta-analysis relating to mental health practitioners, involving 67 studies over 56 years of research, again found a greater degree of accuracy in favour of statistical prediction (Ægisdóttir et al., 2006).

In summary, there is a wealth of research since 1954 across many and varied fields which supports Meehl's (1954) view that statistical prediction is generally more accurate (Dawes et al., 1989; Grove et al., 2000; Ægisdóttir et al., 2006; Kahneman, 2011) than the clinical predictions of experienced practitioners. Despite the strength of evidence that exists generally, it is often very difficult for some to accept that a statistical process can outperform their own clinical judgement, such is the power of our own intuitive judgement.

7.1.1 Clinical Judgement

Even when individuals accept, in an aggregate sense, that statistical forecasting generally produces better results than clinical judgements, they often fail to accept that their own decisions suffer from the same deficiencies described in these larger studies. Kahneman (2011) proposed in making general judgements and decisions that there are two processes that operate in the mind - System 1 and System 2, and goes onto describe the role of heuristics in judgement, that lead to bias in our decision making which in turn can affect the accuracy of the decision. System 1 is fast, automatic thinking, requiring no effort; System 2 on the other hand, is more thoughtful and requires effort. Conversely, others believe that too much information could cause an individual to make the wrong decision and that quick decisions can be good decisions (Gladwell, 2005). Two heuristics that can affect our judgement will be discussed here - availability and representativeness, providing an example of how this relates to custody officer's decision making.

Firstly, the heuristic of representativeness generally concerns how typical a case, person or outcome is. Decisions are then made using a quick System 1 assessment about an individual based on their similarities to a group. There are some difficulties however, generally the decision maker can disregard the base rates or the 'frequency of outcomes' (Tversky and Kahneman, 1975 p.1124), concerning how representative an individual is of a particular group. Even when presented with base rate information, often the decision maker will believe the base rate or general information to be irrelevant. They believe they have individual information and their own stereotypes start to govern their judgement, without paying any attention to the base rates or the veracity of the information they have of the individual (Kahneman, 2011).

When applying the above to the custody decision that is at issue here, there is a risk that in a custody suite, the custody officer could use System 1 thinking without any

reference to base rate information. Furthermore, there is a danger that in forming his or her judgement, stereotypes formed by the officer may relate solely to experience he/she has. Therefore, the custody officers background, length of police service, length of time as a custody officer may all have a bearing on the actual decision made.

The second heuristic of availability contributes to and shapes our thinking and decisions (Tversky and Kahneman, 1975). Availability concerns how readily you are able to recall information, for example, in your experience what specific instances can you recall that may shape your judgement about a case, individual, or outcome (Kahneman, 2011). To illustrate this in context; if the police decision maker releases a suspect onto unconditional bail and the suspect goes on to commit a serious assault there would be a media furore. The available information can be recalled with ease both by the public and police colleagues. The ability to recall such an event, would likely affect the decision making of the police custody officer subsequently.

In highlighting, how heuristics affect our judgement, it is little surprise that the belief in our abilities to predict the future can be overconfident (Kahneman, 2011). The skill and expertise with which a custody officer makes decisions means they can believe they have developed expert intuition. Klein and Kahneman (2009), considered different professions and whilst policing was not one of them, the findings suggest that intuitive judgement can be predicated on both genuine skill and learning from mistakes brought about by heuristic biases. It is also suggested that experts know the limitations of their knowledge (Klein and Kahneman, 2009). That research concludes, the need to determine whether expert intuition, 'can be trusted requires an examination of the environment in which the judgement is made and of the opportunity that the judge has had to learn the regularities of that environment'. (Klein and Kahneman, 2009, p.524).

To summarise this section, decision making and the role of heuristics in judgement or expert intuition can cause errors. When a clinical prediction is made, it is

likely to be overconfident and extreme. If expert intuition is considered a more accurate clinical prediction as pointed out by Klein and Kahneman (2009), this should be tested in the environment; testing clinical and statistical forecasting in a policing environment is commenced as part of this thesis. It is natural for a human being to generate overconfident judgements, as they piece together the readily available evidence and produce a coherent story which may be plausible but not probable. Kahneman (2011) cautions, 'Be warned: your intuitions will deliver predictions that are too extreme and you will be inclined to put far too much faith in them' (Kahneman, 2011, p.194).

7.2 Statistical Forecasts in a Criminal Justice Setting

In life, many predictions and forecasts are made routinely, produced by personal judgement. Heuristics and expert intuition contribute to how one judges and goes on to predict, forecast or decide a course of action particularly in uncertain areas. In the CJS various methods of statistical prediction have been used. Statistical predictions and guidelines derived from them are there to provide a structure to points in the CJS where discretion is used (Goldkamp, 1987). Judgements and discretion are routinely used at various points in the CJS i.e. arrest, bail or custody, prosecution, conviction, sentencing, prisons and probation (Berk and De Leeuw, 1999; Berk et al., 2009; Barnes and Hyatt, 2012). In the UK, research relating to police forces has demonstrated differences in decision making for out of court disposal of offences for a suspect over decades, highlighting the disparity between UK police forces (Slothower 2014; Neyroud 2015).

The right statistical method should be used (Gottfredson and Moriarty, 2006) and many of the older statistical tools have been rigorously studied however newer machine learning methods have less so. The development of statistical decision support tools began due to the discretion and disparity in decision making. Questions surrounding fairness and whether decisions were appropriate lead to guidelines and a framework to provide structure to the decision-making processes being established (Goldkamp, 1987).

In the US research relating to parole decision guidelines commenced in the 1920s (Harcourt, 2007; Berk, 2012). Despite research in the arena of parole decisions, prediction models differed and very rarely was new research implemented (Harcourt, 2007).

Research in parole prediction methods highlights between 1923 – 1978, 24 academic studies with only 4 or 17% implemented (Harcourt, 2007). The most popular model used in relation to parole is the Level of Service Inventory-Revised (LSI-R), these guidelines are applied post-conviction (Harcourt, 2007). The LSI-R guidelines in some US states is used alongside other information to assist a decision maker in determining parole. The LSI-R provides a means of assessing information such as socio demographic and offending history through interview which, when complete, produces a score which provides decision support (Harcourt, 2007). Research into the LSI-R has suggested there are concerns over the levels of disagreement in LSI-R assessments to the extent the method should not continue (Austin et al., 2003). Likewise, in the UK regarding the same LSI-R used in probation, Home Office research concluded that the LSI-R did not predict serious offences to a high enough standard to make its use appropriate for assessing dangerousness (Raynor et al., 2000).

Other statistical procedures have been used, all arguably endeavouring to achieve higher degrees of accuracy in the decision-making environment to support criminal justice agencies. Post arrest decisions in the custody environment are no exception. In the UK police decision support post arrest consists of a gravity factor matrix document. The matrix, adopted by the Association of Chief Police Officers, was created at the request of the UK government in response to criticism of a lack of consistency in out of court disposals. The gravity matrix was based on a legal framework with the Magistrates Court Sentencing handbook. Local police forces were to implement the matrix in their local police force areas. There is no evidence to suggest that forces across the country are routinely using the matrix (P Neyroud 2016, personal

communication, 29 September) and if they are, arguably, whether individual officers interpret and apply the matrix in the same way is questionable.

The gravity factor matrix looked at factors such as previous offending history together with mitigating or aggravating factors of the offence for which they are arrested however does allow some discretion. In practice, the gravity matrix document can be found in some custody suites - whether a custody officer looks at this document when dealing with suspects or whether they consider themselves knowledgeable enough in the environment to not refer to the matrix is again open for debate.

There are other approaches to statistical forecasting such as various types of regression analysis focusing on sentencing, parole, probation, prison inmate classification (Berk and De Leew, 1999; Paternoster, et al., 2003; Berk et al., 2005; Harcourt, 2007; Berk, 2012; Berk and Bleich, 2013). Within a CJS setting accuracy with which these models can predict can often be a deciding factor for an organisation (Berk et al., 2005; Berk and Bleich, 2013). There is, however, a fine balance between accuracy and ethics when building statistical models and considering the variables to be used within the models (Berk, 2016).

7.2.1 Ethics

Forecasting can present ethical issues surrounding racial bias, socioeconomic bias, demographic bias and gender bias (Harcourt, 2007; Harcourt, 2010; Starr, 2014; Berk and Hyatt, 2015; Harcourt, 2015). Issues can centre on post code and gender; should one's gender contribute to a risk outcome. One could argue that if the police target a neighbourhood/postcode then more people may be likely to be arrested in the area, conversely living in an affluent or rural area that does not see much police activity the likelihood of arrest is less – should that affect the risk outcome of a model?

Variables and the impact of bias can be dependent upon the frequency with which such variables are encountered in descriptive modelling of decisions (Goldkamp,

1987). The forecasting model itself is not the problem, the outcomes used to build the model can be. If outcomes in the data reflect racial bias, the model utilises the outcome data to make forecasts. Therefore, criminal history, which is a strong predictor, could be an indicator of race by proxy particularly in the US (Harcourt, 2007; Harcourt, 2010; Harcourt, 2015; Starr, 2015). There are examples in the US media which demonstrate race may have been a factor in forecasting, leading to particularly uncomfortable consequences in terms of life chances when such forecasts are used at the sentencing point in the CJS (Angwin et al., 2016). Although this racial bias may not necessarily apply in Durham whose arrestees are 97% white, readers in other locations may want to explore this literature further (Harcourt, 2007; Harcourt, 2010; Hannah-Moffatt, 2013; Berk and Hyatt, 2015; Harcourt, 2015; Starr, 2015; Skeem and Lowenkamp, 2015; Berk, 2016).

There are two key points to remember concerning forecasting, firstly the facts of reality are what is used to build the forecasting model - the outcomes of a CJS. If the forecasting model did not reflect reality it would be wrong. There are several points in the CJS where clinical discretion based decisions are made, arrest, bail, charge, sentence – there is a build-up of biases throughout the process, a forecasting model reflects that reality. If we change the status quo in reality, the model will adapt and change with it.

Secondly, an organisation decides which variables are included in the model, balancing fairness with accuracy. Any variable can be omitted from the forecasting model; this may affect the accuracy of the model but these are the finely balanced decisions that need to be considered. In some areas of the US, the main cause of death for young black men is homicide perpetrated by young black men. Therefore, leaving a variable out of the model could potentially change the risk level of a suspect which may mean victims and communities are not provided with the protection they should be able to expect (Berk and Hyatt, 2015). The variables to be included in forecasting models

need to be carefully considered by the organisation seeking to utilise the forecasts (Reyes, 2016).

A final note on statistical forecasts; if forecasts of future dangerousness are biased, surely the question should be, is clinical judgement any less or more biased (Berk and Hyatt, 2015). Harcourt (2015, p.1) argues forecasting tools are, 'simply the wrong way forward', if the goal is to reduce the numbers imprisoned in a country, forecasts will exacerbate the difficulties of gender, or race imbalance. Hannah-Moffatt (2013) argues that perhaps there is a better way with forecasting and using these types of forecasts at the front end of the CJS post arrest, as Durham Constabulary are in this thesis, rather than to inform sentencing decisions may be a more constructive way forward.

7.3 Machine Learning

Statistical methods should be employed, particularly in an environment where accuracy is important and the errors can be costly (Ægisdóttir et al., 2006), the CJS is such an environment. Research conducted by Berk et al., (2009) into forecasting homicide and attempted homicide post-conviction, within probation and parole arena, highlighted that logistic regression, when classifying the offenders, produced a 99.7% false negative result. That is, when the model forecasts whether an offender would commit a homicide or attempted homicide offence it was correct less than 1% of the time. The development of forecasting models with big data sets readily available in the public sector, improvements in IT and more complex modern machine learning tools available enables more accurate forecasts to be made (Berk et al., 2005; Berk et al., 2009; Barnes and Hyatt, 2012; Berk, 2012; Berk and Hyatt, 2016).

Logistic regression has largely been the method of choice for many years however this uses linear decision boundaries which can lead to inaccurate forecasting (Berk et al., 2009). Research exists to compare different statistical approaches

(Breiman, 2001; Berk et al., 2005; Berk et al., 2005; Berk et al., 2009; Berk, 2012; Berk and Bleich, 2013; Ridegway, 2013) which set out benefits of machine learning approaches. Such comparisons in summary point to non-linearity of machine learning being a key component to improving accuracy (Berk et al., 2009) but also the ability to deal with more than two outcome categories (Berk and Bleich, 2013). Machine learning has the benefit when making decisions in such a complex area as the CJS, to allow different types of errors and cost ratios to be built in to the model at the outset (Berk and Bleich, 2013). In machine learning stochastic gradient boosting, Bayesian trees and random forests are largely the competitors when it comes to applying the appropriate machine learning techniques (Berk, 2012). The method of choice in this research is random forest modelling.

While machine learning is a developing field, and multiple different approaches continue to compete with one another, one certain conclusion can be drawn from this literature; once a prediction is made, the accuracy of the prediction is what determines its success. The complexity of machine learning means it is increasingly difficult to explain to non-computer scientists and non-statisticians how a machine learning forecasting tool arrives at its outcome other than a basic explanation. Commercial companies in the US have set up to provide forecasts to public agencies at a cost however have not fully explained how the forecasts are arrived at (Angwin et al., 2016). It is difficult to understand therefore which model is being used, whether it has been validated and whether, before implementation, one can have confidence in the model. It is critical, in light of scepticism, to explain how accurate the predictions are, by comparing actual forecasts with actual outcomes in the real world, a validation can be undertaken with data that has not been used as part of the construction of the forecasting model, to establish the accuracy (Monahan and Skeem, 2013).

7.3.1 Random Forests

Using and testing statistical random forest forecasts within a CJS setting has occurred in the United States. Random forest methodology allows hundreds of thousands of criminal records to generate forecasts of dangerous offending over a specific time period. Efforts have primarily focused on predicting those offenders who are likely to commit offences in the future in a probation setting (Berk et al., 2009; Barnes et al., 2010; Barnes and Hyatt, 2012; Hyatt and Barnes, 2014). More specifically, at the post-arrest stage in the UK or arraignment (pre-trial) in the US, a random forest model has been developed in Philadelphia and has, to date, not yet been deployed largely due to ethical concerns (Reyes, 2016). That said, research conducted by, Berk et al. (2016), provides a random forest forecasting model for domestic violence, however again this has not been deployed. This study has been unable to identify any other random forest forecasting model at the post arrest stage.

Random forest modelling is a statistical technique that can use large datasets from different data sources and can take account of many predictor variables. The technique uses many classification and regression trees (CART). Each tree is unique and ultimately produces a vote in relation to one of three classifications (High, Moderate, or Low in the Durham Constabulary model). The votes are totalled enabling the classification forecast to be the outcome which receives the most votes. (Barnes and Hyatt, 2012). There are two key features that set random forest modelling apart from other statistical processes – the use of non-linear relationships and cost ratios (Barnes and Hyatt, 2012).

Firstly, non-linearity in the model is the ability to take account of relationships that are not linear, often there is a curved relationship. Linear regression by its nature assumes a linear relationship which may lead to errors. (Berk, et al., 2009) Partial dependence plots are a way of looking at the association of an input variable on the

likelihood of being classified into a risk group. Figures 1 and 2 show partial dependence plots for the Durham Constabulary model subject of this thesis (G Barnes 2016, personal communication, 19 April).

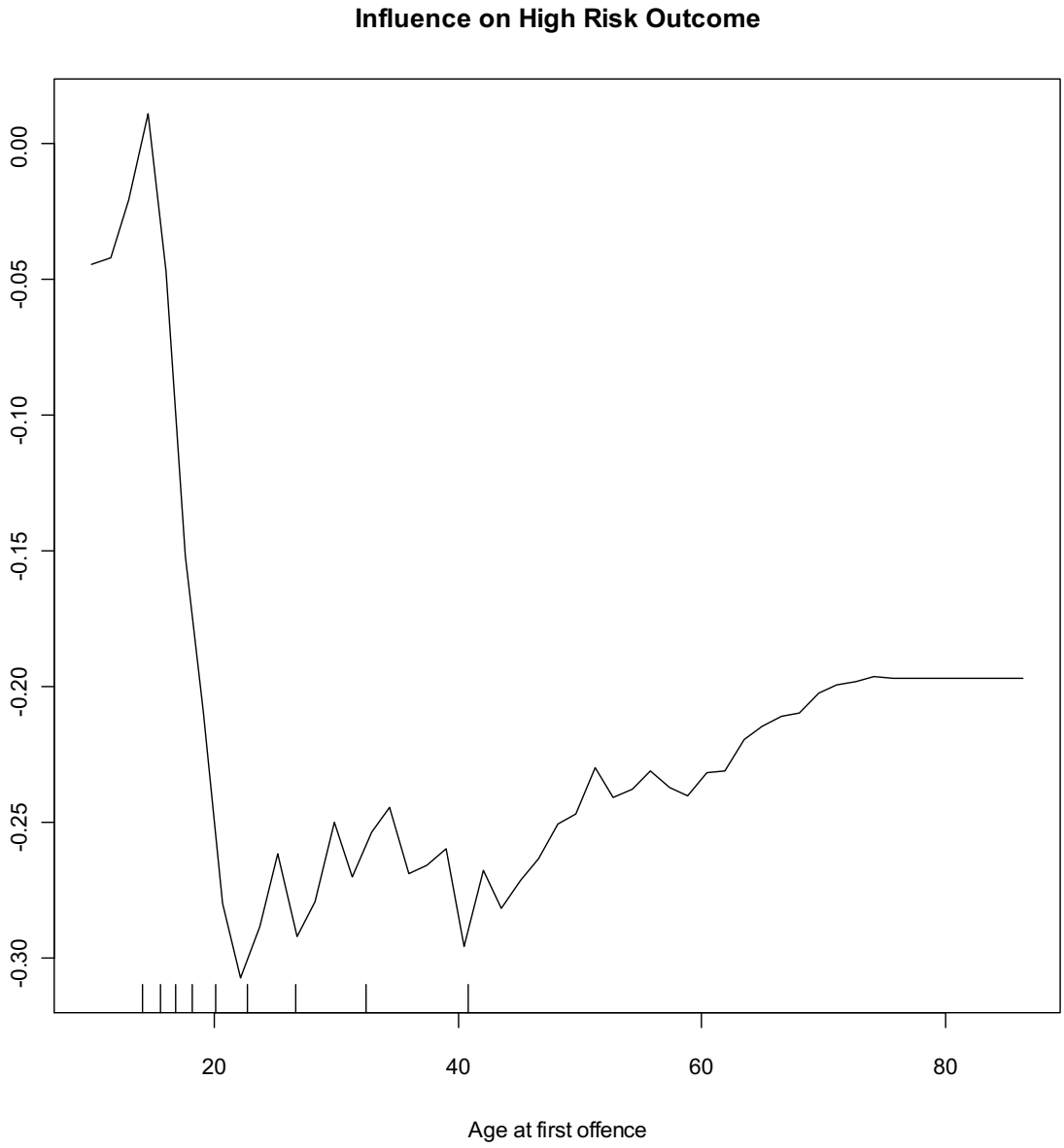

Figure 1: Dr Barnes Partial Dependence Plot – age at first offence – Construction 2008-2012

The X axis in both Figures 1 and 2 illustrate the data distribution in deciles, note that the closer the markers are the greater the density of data. Both figures illustrate the non-linear curved nature of the relationship; a straight line through these may lead to

greater error. (Berk et al., 2005; Berk et al. 2009; Barnes and Hyatt, 2012; Berk, 2012; Berk and Bleich, 2013; Berk et al., 2016).

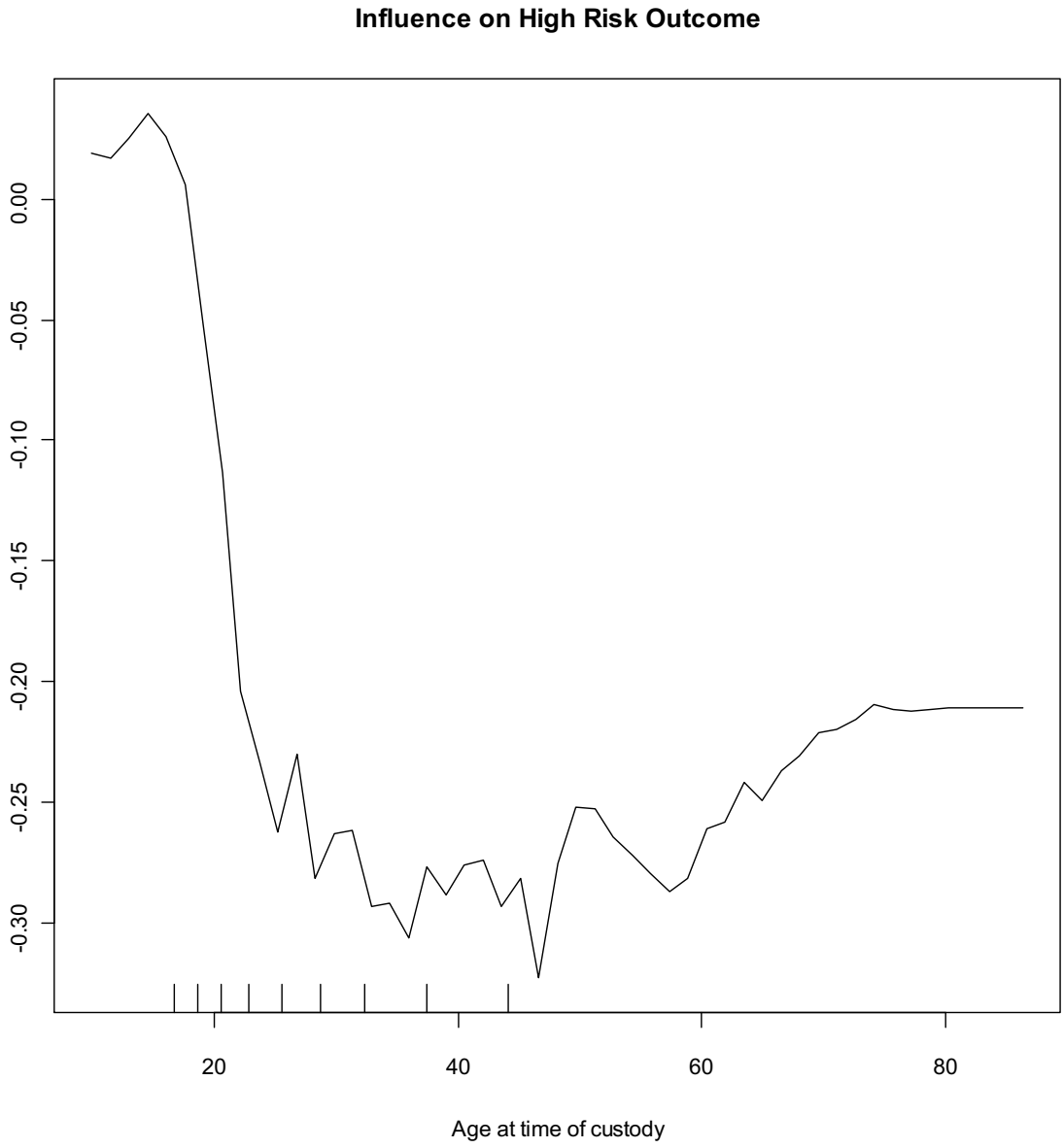

Figure 2: Dr Barnes Partial Dependence Plot – Age at presenting offence - Construction 2008 - 2012

Secondly, random forest modelling has the capacity to consider different types of errors and the costs associated with them (Berk et al., 2009; Barnes and Hyatt, 2012; Berk, 2012; Berk et al., 2016). It is important to note that forecasting models are not error free but also that not all errors are equally problematic. Random forest methods can consider cost ratios of errors, such as, an offender who is predicted to be relatively

safe, but who then goes on to commit a serious violent offence (high-risk false negative), is likely to be costlier an error than an offender who is predicted to be at high risk of committing a serious offence however turns out to be low risk (high-risk false positive). These considerations are accounted for within the building process of a random forest forecasting model, this is unique to random forest modelling and makes utilisation an attractive one for a CJS setting (Berk et al., 2009; Barnes and Hyatt, 2012; Berk, 2012; Berk et al., 2016).

Random forests predict the accuracy of the forecast based upon a unique sub set of the data held in reserve, with a different 'out of bag'(OOB) sample set aside for each individual tree within the model. Barnes and Hyatt (2012) point out this is not a fully independent validation sample, but it does provide the models' best estimate of accuracy. Similarly, another way to validate the accuracy of the model would be to utilise an independent sample, with actual outcomes; to validate the model with data not used to construct the model.

Barnes and Hyatt (2012) provide the only validation research this literature review could identify of a random forest analysis in a CJS setting. However, this is in a probation setting and is post-conviction. To obtain data to enable validation of the model, Barnes and Hyatt (2012) used data just under a decade old before the model was constructed. In validation, the Barnes and Hyatt (2012) research demonstrates a drop in overall accuracy which the research highlights is entirely normal in a validation exercise but also the validation data had different characteristics than the construction data. The data, whilst being a decade old, enabled analysis of how accurate the forecasts were and for how long. In contrast, this thesis seeks to validate Durham Constabulary model post arrest in policing, using data which brings the research up to date, through to April 2016

7.4 Summary

This literature review has explained clinical versus statistical forecasting, provided some explanation as to why clinical judgement may not be as accurate. The development of guidelines as decision support with examples both in the US and UK has been provided together with a summary of modern methods of statistical forecasts in a CJS setting. A focus on random forest forecasting is presented which highlights two key benefits of using such a tool, the consideration of non-linear relationships together with an ability to consider cost ratios.

Random forests have been used in the US to forecast offender behaviour (Berk et al., 2009; Barnes et al., 2010; Barnes and Hyatt, 2012; Berk, 2012; Hyatt and Barnes, 2014), but, perhaps due to difficulties with implementation, available data and IT, scepticism around the method used, its accuracy, and our own intuitive judgement- it seems implementation of this type of forecasting continues to provoke controversy (Harcourt, 2007, 2010, 2015; Starr, 2014, 2015; Angwin et al., 2016; Reyes, 2016). Media in the UK are starting to explore what is happening with algorithms post-conviction in the US and it is important that in the UK context, which has a very different policing history to that of the US, that we have the debate about random forest forecasting (Naughton, 2016).

The literature review has been unable to identify post initial arrest in the UK or pre-trial arraignment in the US, live deployed random forest forecasting research, albeit research in the US exists at pre-trial stage live deployment has been a challenge. Arguably there is an obligation on policing to understand the most accurate and efficient method of decision support (Neyroud, 2015). The newer methods of forecasting should be tested and validated to enable a transparent understanding.

The method for determining the accuracy of a statistical prediction model and validating it to ensure we are using the best means of structured decision support has

important implications for agencies whose decision have consequences not only for a suspect but also for the communities served by the police

To validate the model and establish in the future, for the first time in a UK post arrest environment, in the CJS, will add to a growing body of research. This research contributes to the debate of whether this type of forecasting model can be expanded for use at the initial arrest disposal point in a UK police setting and enable rigorous testing of interventions to reduce reoffending and to support an offender to desist in an evidence based way (Sherman, 2012; Sherman, 2013; Neyroud, 2015).

8.0 Methodology

The aim of this descriptive analysis is to examine algorithmic forecasting of offender dangerousness for use by police custody officers, by assessing the accuracy of the Harm Assessment Risk Tool (HART), created by Dr Barnes and Durham Constabulary, at the gateway to Criminal Justice System (CJS).

There is little research concerning the validation of such random forest models, other than Barnes and Hyatt (2012). The Barnes and Hyatt (2012) study was used at the probation decision point in the CJS rather than the gateway to the CJS which is subject of this study. There is no UK research this study has could identify of a random forest model being deployed in the CJS and certainly not in policing. Furthermore, utilising such a model at the gateway to the CJS; only Berk et al., (2016) offers a model such as this, however this relates to domestic violence, and this model has yet to be deployed in Philadelphia (US). A programme to introduce a forecasting process at arraignment more generally in Philadelphia (US) has currently stalled (Reyes, 2016).

This section will explain the methods used to answer the research questions;

- What is the validated accuracy of the Durham Constabulary forecasting model using custody events from 2013 compared to 2008-2012 construction sample?

- What is the distribution of forecasting errors using the 2013 validation data compared to 2008-2012 construction sample?

- What are the descriptive characteristics of the three forecasted risk groups, as measured in the 2013 validation data?

- To what extent do the clinical forecasts of custody officers agree with the model-generated algorithmic forecasts?

In assessing the accuracy of HART, a framework set out by Barnes & Hyatt (2012) is utilised throughout to ensure comparison and consistency of analysis. A short

summary of the random forest forecasting modelling of HART is outlined before providing a description of the two distinct datasets used to answer the research questions. Limitations of the second dataset are set out before finally taking each research question in turn and describing the method considered appropriate for analysis.

8.1 Random Forest

A random forest is a statistical process which can handle large data and can cater for many predictor variables. The random forest risk model can deal with different types of errors and the costs associated with them. Other models are not able to do this and all errors created using those methods are treated as equal. This type of statistical process has been shown to allow risk predictions to be made with ever-increasing accuracy (Berk et al., 2009; Barnes and Hyatt, 2012).

The bespoke random forest HART model predicts the risk of future dangerousness of an offender over a 24-month time horizon has three classification outcomes;

- High Risk – a new serious offence within the next 24 months
- Moderate risk – any non-serious offending within the next 24 months
- Low risk – no offence within the next 24 months

Examples of offences that constitute 'serious' are murder, attempted murder, wounding, robbery, sexual offences and firearms offences.

HART was built using custody event data for the period 1st January 2008 to 31st December 2012, which constituted approximately 104,000 custody events. The custody event data is known as the construction sample. The model has 34 predictor variables (see Appendix B). The technique uses many classification and regression trees (CART) – there are 509 in the HART model. Each tree is produced using both a randomly-selected subset of cases and a randomly selected pattern of predictor variables. Each

tree is a model and makes a prediction, which is then used as one vote out of 509 total votes. The votes are counted, and the classification of the overall model becomes the outcome which receives the most votes (Breiman, 2001; Berk et al., 2009; Barnes and Hyatt, 2012; Berk and Bleich, 2013; Berk et al., 2016).

The model can determine its accuracy using a process called 'Out of Bag' (OOB) sampling (Breiman, 2001; Berk et al., 2009; Berk, 2012; Barnes and Hyatt, 2012). OOB sampling allows the model to provide an approximation of its own accuracy. Prior to constructing a decision tree, a random sample is drawn from the construction data, and a smaller subset of that sample is held in reserve, known as OOB. Once a tree is built, the OOB sample, is then used for each tree to validate itself. It is the OOB sampling that enables the random forest model to produce the models best estimate of accuracy and to produce an approximated forecasted outcome for each case in the construction sample. A table explaining the forecasts and the actual outcomes broken down into risk levels is presented in Table 1.

Construction data 2008-2012		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	8.12%	B	6.80%	C	1.80%	16.72%
Forecast Moderate	D	2.25%	E	34.09%	F	12.19%	48.53%
Forecast Low	G	0.82%	H	7.65%	I	26.28%	34.75%
Total		11.18%		48.54%		40.27%	100.00%

Table 1: Construction matrix 2008-2012

Whilst the random forest procedure provides estimates of the accuracy of the model by creating the OOB sample, these are not truly independent data as they have also been used in the construction of the model. This is one reason why a true validation, using fresh and fully independent data would be likely to show a reduction in accuracy. Nevertheless, OOB does provide a good estimate of the accuracy of the model. The

construction matrix information enables analysis of the estimated accuracy overall, individually across the three forecasted outcomes, together with the distribution of errors, immediately after it is built. The information provided by the construction matrix can then be compared to the validation data set for 2013.

8.2 Data

There are two very distinct data sets for this research, a 2013 validation dataset and an agreement dataset for 2016. The 2013 validation dataset was created to make an independent assessment of the accuracy of the forecasting model, which will inform the answers to the first three research questions. The agreement dataset seeks to review to what extent the custody officer and HART agree.

8.2.1 Validation 2013 Dataset

An independent dataset is used to validate the risk model and compare actual forecast and actual outcome accuracy to HART's estimated accuracy. The construction data contained cases from 2008-2012, while the independent validation data contained cases from 2013 and can therefore be used as an independent validation dataset. Both the construction dataset and the 2013 validation dataset are drawn from Durham Constabulary's case and custody management IT systems. The case and custody IT system is utilised by custody officers to record all detained individuals onto a custody record upon their entry into the police station.

The dataset is the whole population of custody event data 1st January 2013 to 31st December, 2013 for Durham Constabulary. A custody event is the disposal decision taken by the custody officer following arrest at the end of the first custody period. The nature of the custody decision can be to bail (conditionally or unconditionally), remand in custody, taken no further action, administer an out of court disposal/diversion scheme or prosecute the suspect again with a decision to bail (conditionally or unconditionally). The dataset consists of 14,882 custody events from 2013, with 95 variables (see

Appendix A). The 95 variables include the 34 predictors for the model, along with other demographic data, details about the predicted outcome for each case, and the actual offences committed by the offender during their two-year follow-up period. The custody event data for 2013 has been 'dropped down' the forecasting model to establish what predictions would have been generated had the forecasting model been live in 2013. Having established the forecasts for all custody events for 2013, the forecast has been compared to the following 24 months of actual offending data up to 31st December 2015.

What is the purpose of using 2013 data rather than 2014 or 2015? To review the accuracy of the forecast made in 2013, 24 months of data following the date of the forecast are needed to establish whether the forecast was accurate in accordance with the definitions of high, moderate and low risk. Therefore, it is crucial that to assess the accuracy of the forecast, knowledge of what happened during the 24 months in terms of offending enables the validation of the forecast. The last date for a potential forecast in 2013 would be 31st December 2013. Each custody event in the dataset has its own 24-month time horizon, for example, if a forecast is made on 3rd March 2013 the check for future offending will be for 24 months after that date and not up to 31st December 2015.

The time horizon of 24 months could mean that for a custody event, a suspect could present more than once during the period – repeated in the data. Each time a suspect is before the custody officer with a new presenting offence it is essentially a new custody event. Therefore, as the model was built using the custody event as the unit of prediction, each time a suspect is brought back into custody their prediction may change. The suspect may be older, have a different postcode, or different intelligence count etc., therefore each time a custody event is recorded for a suspect – whether or not there is more than one instance of the suspect in the validation dataset - they will be included in the data.

8.2.2 Agreement 2016 Dataset

To answer research question four, a separate 2016 dataset was used to assess the extent to which the custody officer clinical forecasts agree with HART forecasts. The agreement dataset period was 20 September 2016 – 9 November 2016 and consists of 888 custody officer and model forecasts for comparison.

The process comprised of the custody officer making their own prediction of the suspects future behaviour. To conduct the process and capture the data, a front-end IT user interface, was created for the custody officer to complete. The custody officer would make their own prediction of future dangerousness by completing the forecasting element of the interface for each custody event. To make their own prediction the custody officer would provide the answer to two questions.

Question 1 Do you think that this suspect will be arrested for a new offence in the Durham force area within the next 2 years? Yes, or No

Question 2 Do you think that any of these offences will be serious, such as (a) murder (b) attempted murder (c) grievous bodily harm (d) robbery (e) sexual offence or (f) firearms offence? Yes, or No

The two clinical questions when answered by the custody officer effectively provide a forecast of high, moderate or low risk as defined by the model. The HART forecast, whilst still being created, was deliberately not revealed to the custody officers. In so doing the custody officer's prediction cannot be tainted by knowing the model forecast, or trying to match the model forecast. In deliberately withholding the model forecast from the custody officer, the custody officer was therefore blind to the model forecast. This exercise is described as wilful blindness. Using the HART and custody officer forecasts for the same 888 suspects enabled a comparison of the extent to which

statistical forecasts and clinical judgement of custody officers in the three forecast groups agree.

8.3 Data Limitations

The code used by Dr Barnes to generate the forecasting model construction data 2008-2012 was also used to generate the 2013 validation dataset, therefore the validation data is of equivalent cleanliness to the construction data used to build the forecasting model (G Barnes 2016, personal communication, 29 June).

The agreement dataset has limitations in terms of the number of eligible forecasts available for analysis and compliance with the clinical process. Firstly, in cleaning the data, a number of forecasts from the custody event data were not included. Figure 3 below details the forecasts excluded from the dataset and the reason. The number within the dataset ultimately is 888 custody events that were available for comparison.

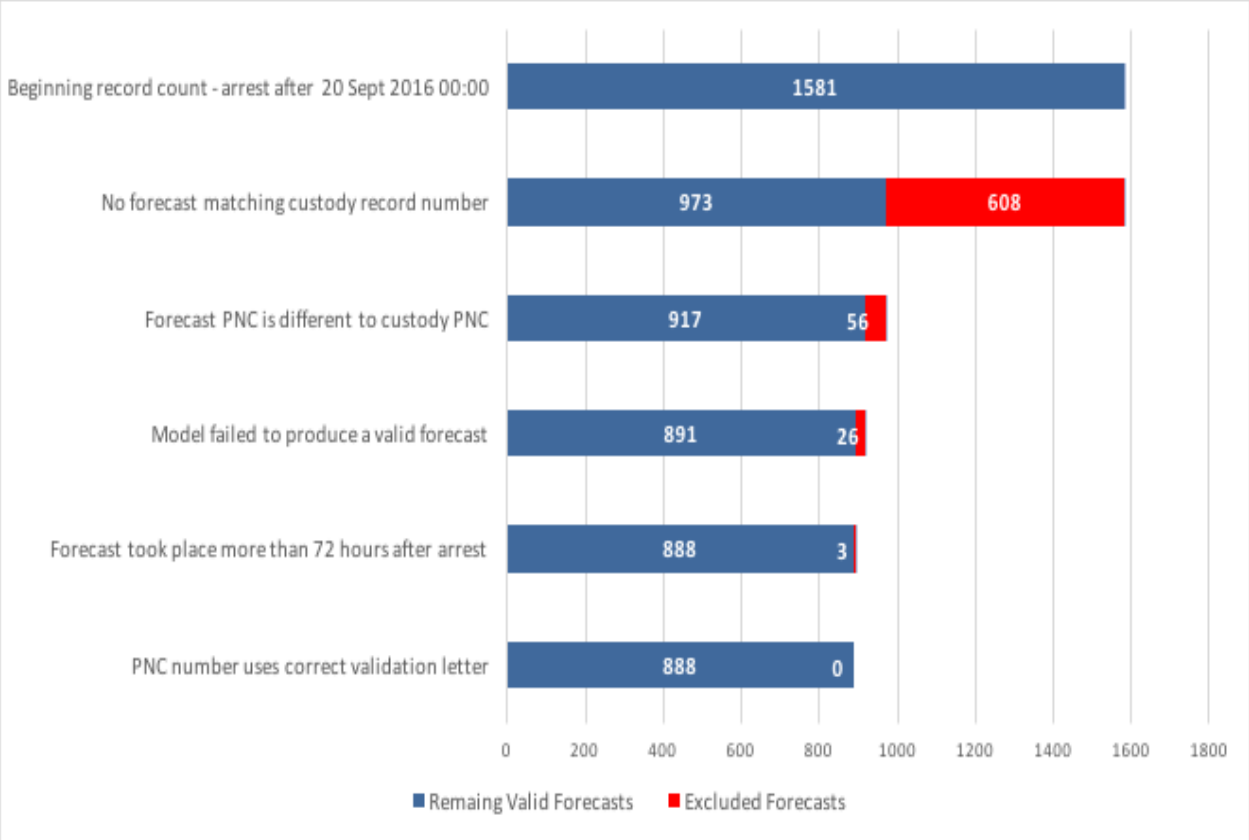

Figure 3: Excluded forecast opportunities for comparison dataset

Secondly, in terms of compliance, the implementation of the wilful blindness exercise relies on custody officers answering the wilful blindness questions. The graph at Figure 4 shows the compliance figures over the period 20 September 2016 – 9 November 2016 and demonstrates the number of opportunities to forecast versus the actual number of forecasts completed by custody officers. On average Figure 4 illustrates compliance was 80%, it is therefore reasonable to draw conclusions from this data.

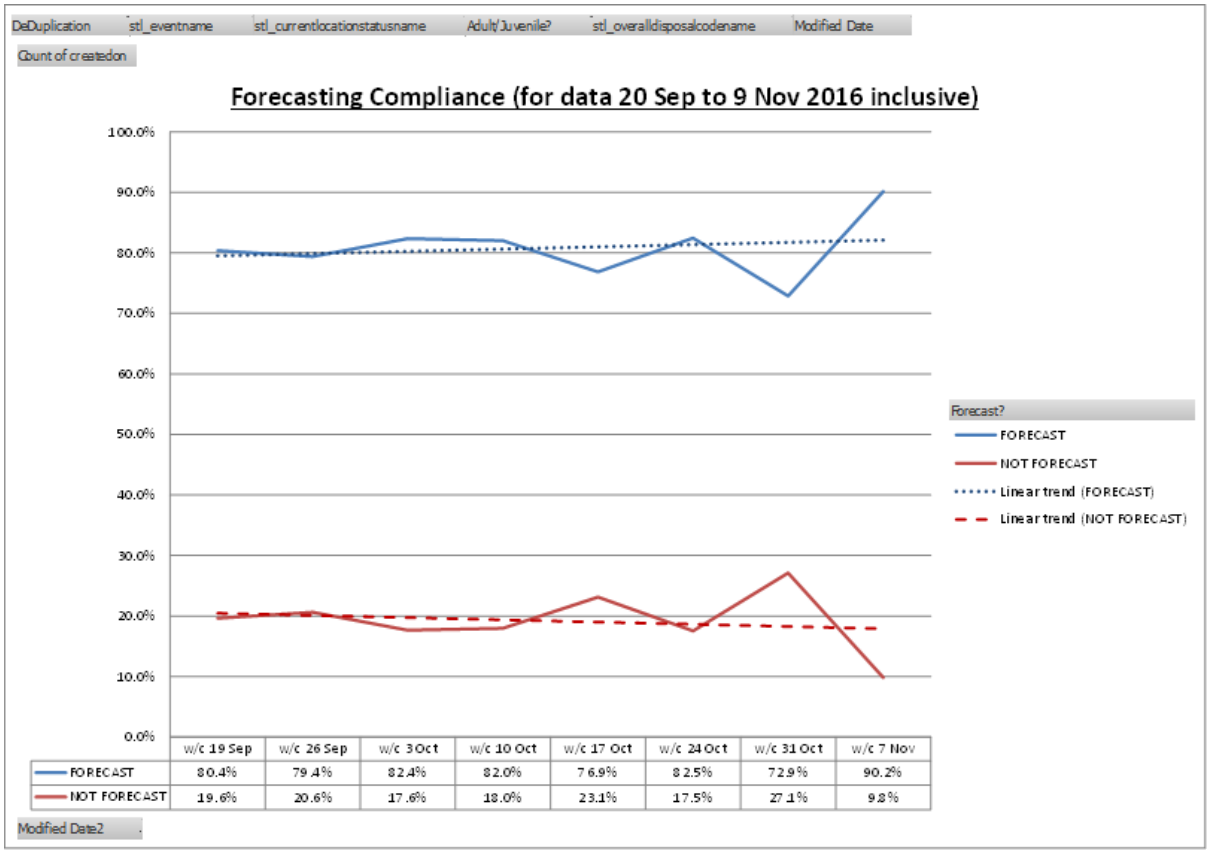

Figure 4: Forecasting compliance 20 September – 9 November 2016

8.4 What is the validated accuracy of the Durham Constabulary forecasting model using custody events from 2013 compared to 2008-2012 construction sample?

The first research question examines the accuracy of the HART model's OOB construction data compared to the population of custody events from the 2013 validation dataset. The HART OOB construction confusion matrix will be compared to the 2013 validation confusion matrix.

Construction data 2008-2012		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	8.12%	B	6.80%	C	1.80%	16.72%
Forecast Moderate	D	2.25%	E	34.09%	F	12.19%	48.53%
Forecast Low	G	0.82%	H	7.65%	I	26.28%	34.75%
Total		11.18%		48.54%		40.27%	100.00%

Table 2: Construction matrix - Overall accuracy

Table 2 illustrates the accuracy of the HART construction matrix. The green cells show the overall accuracy of the model by calculating when the model made its forecast what percentage were accurate within the overall population of forecasts made. A confusion matrix for the 2013 validation dataset has been created for comparison. Further analysis of accuracy within risk group (Barnes and Hyatt, 2012) is presented in a table using basic formula to answer this research question.

8.5 What is the distribution of forecasting errors using the 2013 validation data compared to 2008-2012 construction sample?

Unrealistic expectations that forecasting models will be error free in their predictions are misguided. It is likely they will predict with greater accuracy but certainly will not be error free (Ridgeway, 2013). The second research question will review the forecasting error distribution in the 2013 validation dataset. The 2013 validation confusion matrix created as part of the first research question will be used and compared with the construction confusion matrix. The error distribution of the construction data is

highlighted in Table 3, with the intensity of the colour indicating errors which are likely to be more costly errors.

Construction data 2008-2012		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	8.12%	B	6.80%	C	1.80%	16.72%
Forecast Moderate	D	2.25%	E	34.09%	F	12.19%	48.53%
Forecast Low	G	0.82%	H	7.65%	I	26.28%	34.75%
Total		11.18%		48.54%		40.27%	100.00%

Table 3: Construction matrix – error distribution

Dangerous errors are coloured red and cautious errors are coloured amber. The high-risk false negative, lower left (G) with the most intense red colour is a very dangerous error. The high-risk false negative can mean that a suspect forecast as low risk goes on to commit a serious offence. The high-risk false positive error in the upper right (C) with most intense amber colour is a very cautious error. The high-risk false positive error involves a suspect being forecast high-risk which could mean the suspect receives intensive police attention and that police resources are being inefficiently used. Clearly the false negative error involves a member of the community becoming the victim of a serious offence which following a low risk forecast would be the most undesirable outcome.

There are different ways to review the error distribution in a 3 x 3 confusion matrix such as Table 3. An emphasis will be placed on the type of errors and cost ratios and a table of results will be presented highlighting how the calculations are arrived at.

To provide context to the type of errors, nine short case studies have been conducted to understand, when the model gets the high-risk category wrong, what these kinds of errors look like, other case studies then examine what correct high risk forecasts look like. The case studies were drawn from the 2013 validation dataset. The data consists of three categories, high-risk false positive (n = 332), high-risk false negative (n

= 108) and high-risk true positive (n = 931). Two cases studies from each of the three categories were randomly selected. A third case study was also deliberately chosen (without random selection) from each of the three categories, that were thought to illustrate particular aspects of these kinds of cases. Although an equal number of cases from each of these three lists are highlighted here, these small sample sizes are not at all proportionate to the actual distribution of cases within the confusion matrix.

8.6 What are the descriptive characteristics of the three forecasted risk groups, as measured in the 2013 validation data?

The third research question relates to describing the characteristics of the three forecast risk groups. The predictor variables (see Appendix B) will be used to establish the characteristics of the risk groups examining the differences between the three forecasted risk groups. One way ANOVA tests will be conducted to analyse the characteristics of each forecast risk group. The test will show whether the model correctly separates different suspects into the three different risk groups. An example can be age, in general, younger offenders commit a large number and more serious crimes, therefore the high-risk group would be expected on average to have younger suspects than those who are forecasted as moderate or low.

8.7 To what extent do the clinical forecasts of custody officers agree with the model-generated algorithmic forecasts?

The fourth and final question relates to what extent the clinical forecasts of custody officers agree with HART generated algorithmic forecasts regarding future dangerousness of offenders. The desire here is not to see who is more accurate that is for future research in the policing environment – this research question is to review to what extent there is agreement. Comparisons between clinical forecasts and the algorithmic forecasting for the three forecasted groups are made. A table setting out the agreement is presented to show how agreement differs between the forecast groups.

9.0 Results

This chapter presents information relating to the 2013 validation cohort to provide context. The results of each research question are presented in turn. To aid comparison with HART the format of the structure and presentation of results is similar to that utilised in Barnes and Hyatt (2012).

9.1 2013 Cohort

This section will present data from 2013 in Durham Constabulary policing area to provide context in assessing the accuracy of the 2013 validation data against the forecasting model. HART was built using data from 2008 – 2012, its accuracy and functioning are being tested with independent 2013 validation dataset. It is right predictive models should be tested using independent datasets however to is important to assess whether 2013 conditions were different from 2008 - 2012.

The percentage of recorded crime data that were theft offences covering before, during and after 2013 are shown in Figure 5. The chart indicates that theft offences remained constant during 2013.

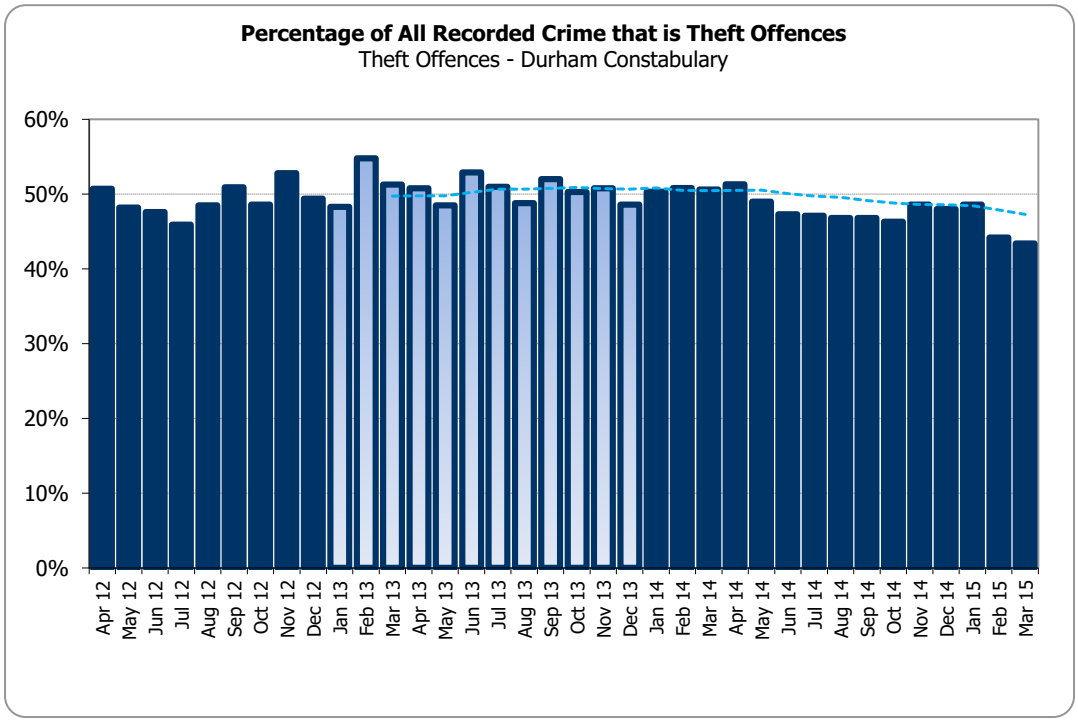

Figure 5: Percentage of all recorded crime that are theft offences

Figure 6, illustrates recorded crime data covering before, during and after 2013. The chart indicates the start of an increase of violence against the person which continued with an increasing trend into 2014 and 2015. The increasing trend is relevant to the analysis as HART is forecasting behaviour for what appears to be a somewhat different actual offending behaviour. Not all violence against the person is classified in HART as a serious offence, however, a proportion of these will be serious offences.

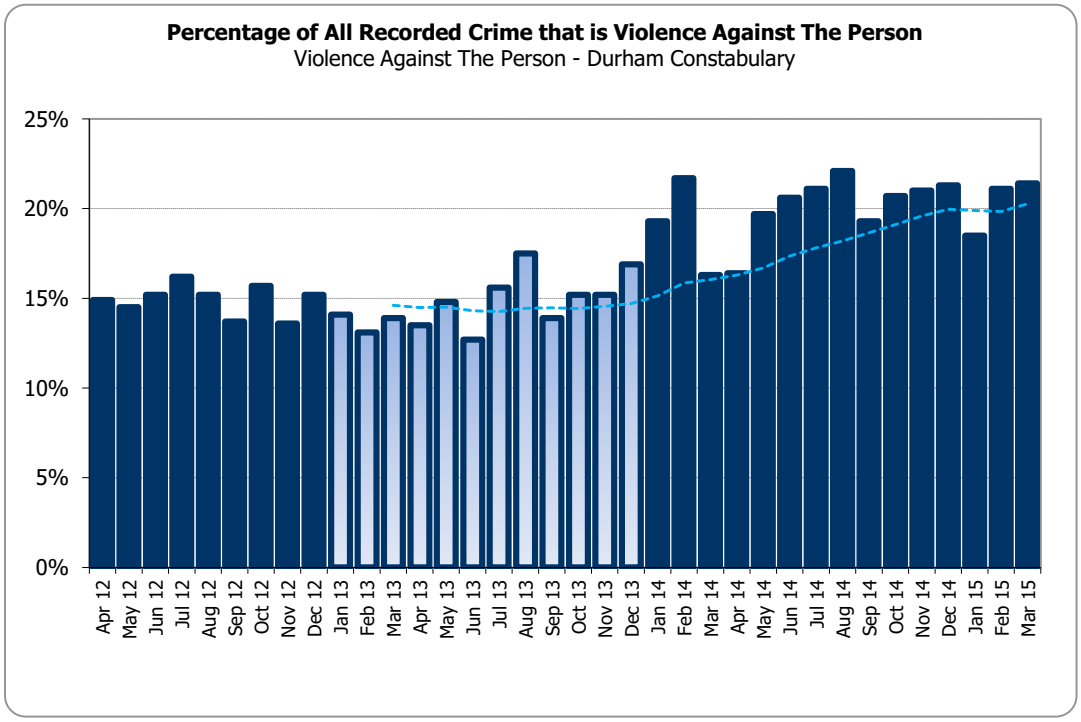

Figure 6: Percentage of all recorded crime that are violence against the person offences

Figure 7 illustrates recorded crime data covering before, during and after 2013, indicates a steep rise in the percentage of all recorded crimes that are sexual offences which are almost invariably classified as serious within the model.

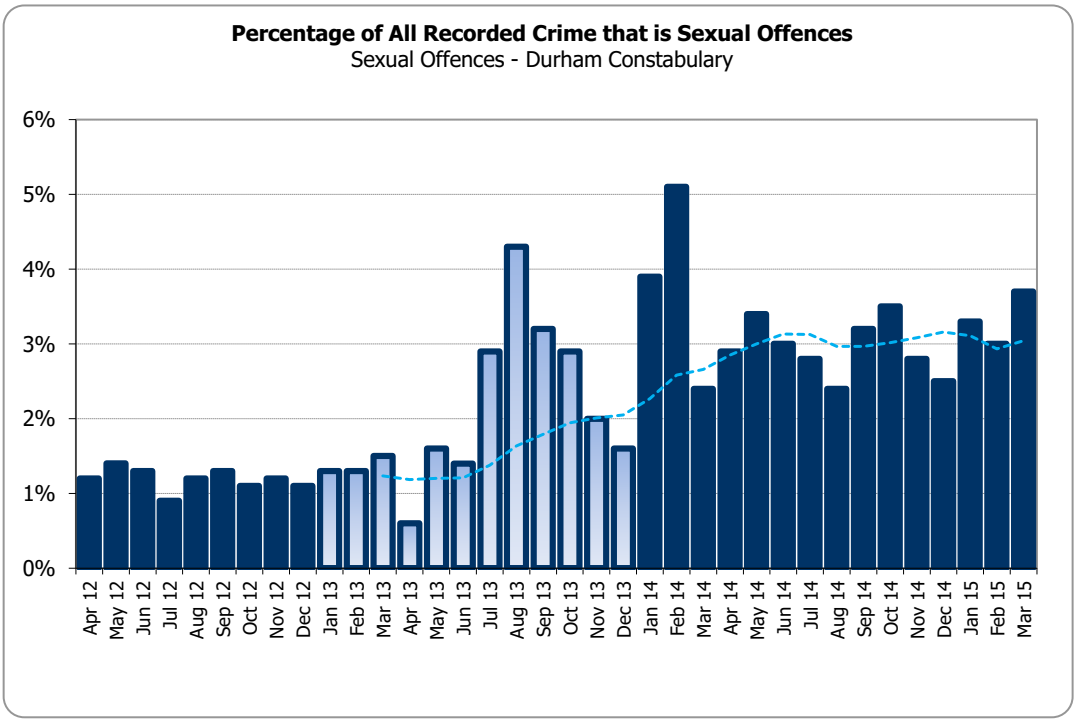

Figure 7: Percentage of all recorded crime that are sexual offences

The values of recorded crime data suggest that offending for the 2013 cohort during the 24-month follow up period was different from earlier years. This may mean that HART is overlaying trends from an earlier period onto a reoffending environment that was rather different.

Figure 8 depicts the prevalence of serious offending within 24 months for all custody events between 2008 - 2013. The chart indicates that 2013 was in keeping with historical trends, although 2010 – 2012 illustrates prevalence was slightly lower.

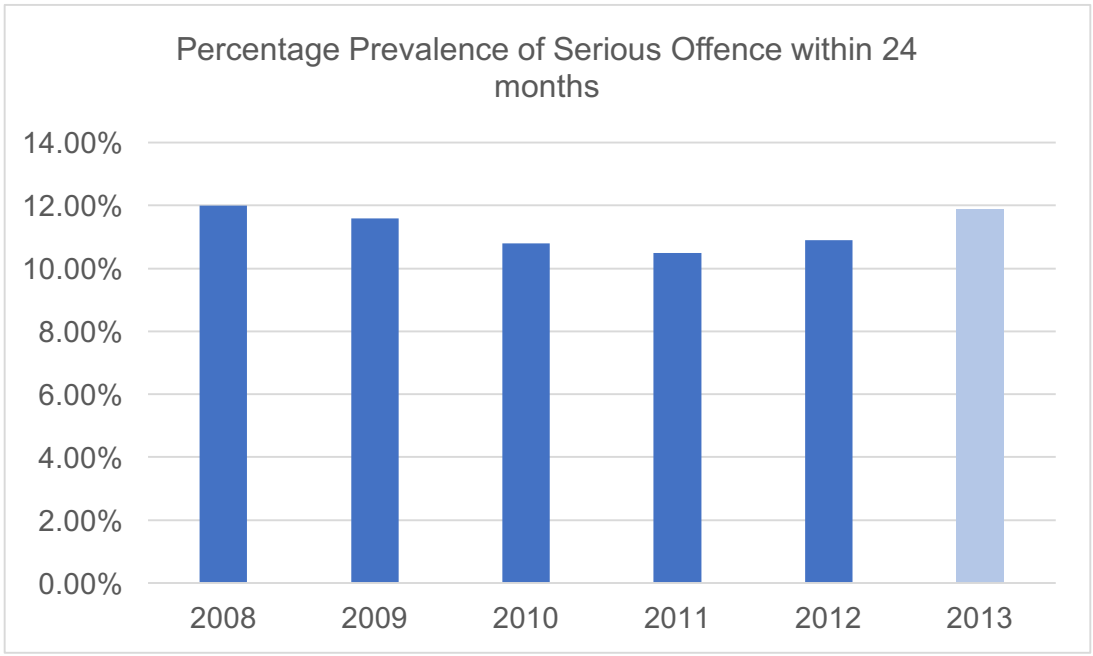

Figure 8: Prevalence of serious offences within 24 months of arrest

Figure 9 displays the mean frequency of serious offending within 24 months for all custody events between 2008 – 2013. The chart indicates that 2013 saw a higher frequency of serious offences within 24 months at 0.159. Three of the five years of construction data were lower than 2013, with 2008 being the highest at 0.171.

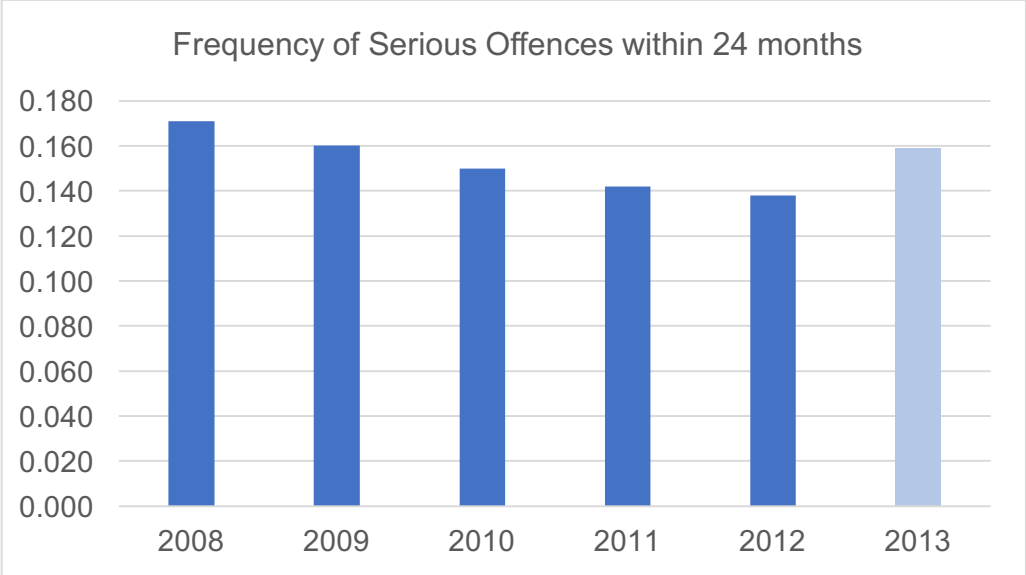

Figure 9: Frequency of serious offences within 24 month of arrest

In summary, the charts displayed in this section show that, the 2013 validation cohort may have been a somewhat different year from that with which the HART model was built to expect.

9.2 What is the validated accuracy of the Durham Constabulary forecasting model using custody events from 2013 compared to 2008-2012 construction sample?

The overall accuracy of both construction and validation datasets are illustrated in Figure 10. The degree of overall accuracy dropped from 68.50% to 62.80% in the validation dataset. It can be anticipated, when validating a model with fresh, independent data that accuracy will fall, it is therefore expected that there would be a difference between construction and validation, in this study that reduction is 5.7% points. In the only other validation research of this type found by this study, (Barnes and Hyatt, 2012), overall accuracy also decreased however by greater margin at 8.3%.

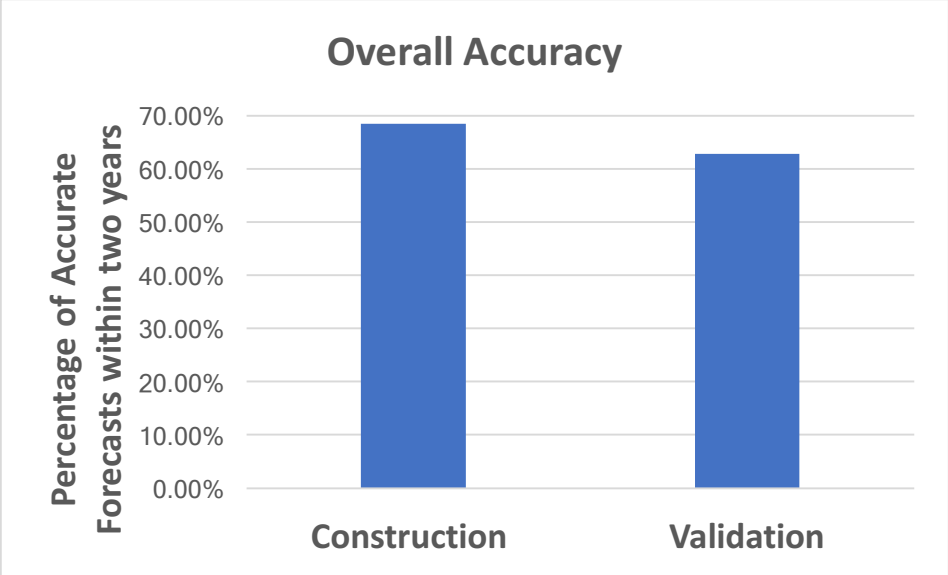

Figure 10: Overall accuracy of construction data and 2013 validation dataset

To compare the overall accuracy a confusion matrix for the 2013 validation dataset is presented to aid comparison with the construction matrix. Tables 4 and 5 provide more detail of the overall accuracy, the cells coloured green (labelled A, E, and I) highlight the overall accuracy. Tables 4 and 5 show a 2%-point reduction in high-risk, a

2%-point reduction in moderate risk, and a 2% points reduction in low risk of all observations. However, the confusion matrices provide the accuracy as a percentage of all cases. To assess within risk group accuracy one would be better served by reviewing the number of those forecast as high-risk (moderate or low) and how many were accurately forecast. Reviewing the degree of forecasting accuracy within risk group is highlighted in Table 6.

Construction data 2008-2012		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	8.12%	B	6.80%	C	1.80%	16.72%
Forecast Moderate	D	2.25%	E	34.09%	F	12.19%	48.53%
Forecast Low	G	0.82%	H	7.65%	I	26.28%	34.75%
Total		11.18%		48.54%		40.27%	100.00%

Table 4: Construction matrix - overall accuracy

2013 Validation		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	6.26%	B	10.01%	C	2.23%	18.49%
Forecast Moderate	D	4.88%	E	32.53%	F	13.55%	50.95%
Forecast Low	G	0.73%	H	5.81%	I	24.02%	30.55%
Total		11.86%		48.35%		39.79%	100.00%

Table 5: Validation 2013 matrix - overall accuracy

Table 6 provides the formula for determining the accuracy results using the labels assigned in the validation confusion matrix in Table 4 and 5. The light blue rows of information provided in Table 6 show the denominators that are used to produce the accuracy calculations in the darker blue rows. The darker blue rows detail more specific accuracy, which is arguably a more reasonable method for reviewing accuracy across risk groups. The table presents results of the construction model and the 2013 validation dataset.

Description	Formula	HART Construction	Validation 2013
Overall Accuracy	(A + E + I) / Total events	68.50%	62.80%
Percent actually high risk	(A + D + G) / Total events	11.20%	11.86%
Percent actualy moderate risk	(B + E + H) / Total events	48.50%	48.35%
Percent actually low risk	(C + F + I) / Total events	40.30%	39.79%
Of those actually high risk, percent forecast correctly:	A / (A +D + G)	72.60%	52.75%
Of those actually moderate risk, percent forecast correctly:	E / (B + E + H)	70.20%	67.28%
Of those actually low risk, percent forecast correctly:	I / C + F + !)	65.30%	60.35%
Percent forecasted high risk:	(A + B + C) / Total events	16.70%	18.49%
Percent forecasted moderate risk:	(D + E + F) / Total events	48.50%	50.95%
Percent forecasted low risk:	(G + H + I) / Total events	34.80%	30.55%
Of those forecast high risk, percent forecast correctly:	A / (A + B + C)	48.50%	33.83%
Of those forecast moderate risk, percent forecast correctly:	E / (D + E + F)	70.20%	63.84%
Of those forecast low risk, percent forecast correctly:	I / (G + H + I)	75.60%	78.60%

Table 6: Risk group accuracy comparison

There are seven different accuracy figures highlighted in the darker blue colour on Table 6, the first is overall accuracy followed by six specifically relating to accuracy within each risk group. Five of the risk group accuracy figures show a reduction in accuracy, with only those forecast correctly as low risk showing an increase from 75.60% to 78.60%.

Focusing on the high-risk group, Table 6 shows a drop of 20% points in those actually high-risk who were forecast correctly, with construction data correctly forecasting 72.60% and validation data 52.75%. This is a large drop in accuracy and is undesirable.

Comparing this to the Barnes and Hyatt (2012) validation, a fall in the degree of accuracy of 26.4% was evident from 61.4% to 35.0% in the same area. A drop of accuracy in high-risk forecasts is also shown in Table 6 for those forecast high-risk; the construction model forecast correctly 48.5% compared to validation 33. 8% a drop of just under 15%-point drop in forecasting accuracy.

The Table 6 results illustrate an increase in actual high-risk outcomes from 11% (construction) to 12% (validation) and the model responded by increasing the forecasts of high risk from 16.7% in construction to 18.5% in validation. Moderate risk actual outcomes remain virtually constant at 48.5% in construction and 48.35% in validation however the model increased moderate risk forecasting slightly which indicates the population of 2013 validation dataset may have been different in terms of offending behaviour to that of the construction data.

In summary, overall accuracy fell as expected and more specifically accuracy in the high-risk group fell substantially. The model responded to this by increasing high and moderate risk forecasts. The only area that accuracy increased was in the low risk group. This is an important observation in that by forecasting accurately in the low risk group area minimises the error rate and the worst kind of error, of a suspect forecast as low risk who commits a serious offence.

Some consider that overall accuracy is not good indicator of this type of model, and that the ability of the model to avoid more costly error is key (Berk, 2012). Random forest modelling provides an ability to build in cost ratios and therefore to build errors into the model to minimize undesirable outcomes. Not all errors are equal. The error distribution is examined in the next section.

9.3 What is the distribution of forecasting errors using the 2013 validation data compared to 2008-2012 construction sample?

To compare the distribution of errors overall a confusion matrix for the 2013 validation dataset is created to aid comparison with HART construction matrix. Both matrices are presented below;

Construction data 2008-2012		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	8.12%	B	6.80%	C	1.80%	16.72%
Forecast Moderate	D	2.25%	E	34.09%	F	12.19%	48.53%
Forecast Low	G	0.82%	H	7.65%	I	26.28%	34.75%
Total		11.18%		48.54%		40.27%	100.00%

Table 7: Construction matrix - error distribution

2013 Validation		Actual High		Actual Moderate		Actual Low	Total
Forecast High	A	6.26%	B	10.01%	C	2.23%	18.49%
Forecast Moderate	D	4.88%	E	32.53%	F	13.55%	50.95%
Forecast Low	G	0.73%	H	5.81%	I	24.02%	30.55%
Total		11.86%		48.35%		39.79%	100.00%

Table 8: 2013 Validation matrix - error distribution

Tables 7 and 8 illustrate in the coloured boxes the different types of forecasting error for each of the forecast risk groups, labelled (B, C, D, F, G, and H). The lower left corner of the matrices (D, G and H) indicate dangerous errors, the intensity of the red colour reflects the very dangerous errors (G) and less dangerous errors (D and H). The very dangerous (G) or high-risk false negative error, indicates the model forecast an outcome would not take place but it did, for example, the offender was forecast as low risk but went on to commit a high-risk offence. Whilst cells D and H are less dangerous errors as the level of risk was forecast lower than the actual offending behaviour

however the error was not so costly. The errors in cells D and G were not considered as costly due to the difference in forecast outcome and actual offending outcome not being as great as the very dangerous error (G).

The upper right corner of the matrices (B, C and F) are cautious errors with the intensity of the amber colour reflecting the very cautious error (C) and the less cautious errors (B and F). A very cautious error (C) or high-risk false positive, indicates the model has forecast an outcome however the outcome does not occur for example, the suspect being forecast as high-risk however does not go on to commit an offence. Cells B and F are less cautious errors as the level of risk was forecast higher than the actual offending behaviour however the error was not so very cautious.

The matrices show not all errors are equal. The very dangerous errors remained relatively unchanged from construction to validation at 0.82% to 0.73% respectively. The high-risk false negative rate of 0.73% represents 108 very dangerous errors out of the 14,882 custody events in the validation cohort. The very cautious errors also remained relatively unchanged in construction to validation at 1.80% to 2.23% respectively. The high-risk false positive rate of 2.23% represents 332 very cautious errors out of the 14,882 custody events in the validation cohort. These types of errors are identified by the organisation as requiring the most weighting therefore it would appear the model is functioning as one would expect.

Tables 7 and 8 show that cautious errors taken together (B + C + F) increased from 20.79% to 25.79% of all observations in construction and validation respectively, which equate to an increase of 5% points of cases that fall into this type of error. Dangerous errors taken together (D + G + H) increased also, however by a much smaller margin from 10.72% to 11.42% which equates to 0.7%point increase in cases falling into this type of error. Nevertheless, in the dangerous error cell D, for those offenders who were forecast to be moderate and were actually high risk increased by 2.63% points of

all observations however this equates to a 53.89% increase in the proportion of case that fall into this error type. This is perhaps a reflection of the somewhat different cohort in the 2013 follow up period.

9.3.1 Cost Ratios

This section sets out, in Table 9, for both HART and the 2013 validation dataset, the calculations for how the high-risk false positive and false negative statistics are calculated, and the cost ratios using the labels assigned in Table 8.

Description	Formula	HART Construction	Validation 2013
Of those forecasted high risk, percent that were actually low risk:	C / (A + B + C)	10.80%	12.06%
Of those forecasted low risk, percent that were actually high risk:	G / (G + H + I)	2.40%	2.38%
Falase Positive to False Negative			
High Risk:	(B + C) / (D + G)	2.803	2.183
Moderate Risk:	(D + F) / (B + H)	0.999	1.165
Low Risk:	(G + H) / (C + F)	0.605	0.414
False Negative to False Positive Ratio			
High Risk:	(D + G) / (B + C)	0.357	0.458
Moderate Risk:	(B + H) / (D + F)	1.001	0.858
Low Risk:	(C + I) / (G + H)	1.652	2.413
Percent of cases that are cautious	(B + C + F) / Total events	20.80%	25.78%
Percent of cases that are dangerous errors	(D + G + H) / Total events	10.70%	11.41%
Cautious errors to dangerous errors	(B + C + F) / (D + G + H)	1.94	2.258
Percent of cases that are very	C / Total events	1.80%	2.23%
Percent of cases that are very dangerous errors	G / Total events	0.82%	0.73%
Very Cautious errors ro very	C / G	2.192	3.074

Table 9: Error distribution and cost ratios comparison

The forecasting model was built to minimise high-risk false negatives; this is the most dangerous error. The results show the model minimised the high-risk false negative which remained largely unchanged at 2.40% and 2.38% respectively. To achieve accuracy in the high-risk false negative area, in the building of the model the prevalence of other error types are set higher to minimise the harm caused by high-risk false negative outcomes. The results show that there was an increase from construction to validation for high-risk false positive outcomes; from 10.80% to 12.06% respectively. The low-risk false negative to false positive ratio rose from 1.65 to 2.41 indicating an increasing preference towards more cautious errors. Lastly the very cautious to very dangerous error or high-risk false positive to false negative ratio increased from 2.19 to 3.07 respectively.

In summary, the model appeared to become more cautious in light of the prevalence and frequency of high risk outcomes it was presented with in 2013 follow up period. The error distribution was more cautious, and very dangerous errors remain the same in validation as in construction. A selection of cases provide context to the high-risk category in the next section.

9.3.2 Case Studies

Nine case studies are completed to provide context to the errors, from the 14,882 custody events in the 2013 cohort, with a focus on high risk outcomes. As described in the earlier methodology chapter, the forecasting model has 509 decision trees, each tree produces a risk outcome, which means there are 509 risk outcomes or votes. The votes are calculated and the overall risk outcome has the most votes. The cases studies highlight the distribution of votes in 'error type' order followed by a summary of the case studies with further information contained in Appendix C.

High Risk Error Type	Case Study	Votes		
		High	Moderate	Low
False Negative	1	18	37	454
	2	115	196	198
	3	114	78	317
False Positive	4	308	165	36
	5	264	213	32
	6	248	242	19
True Positive	7	228	217	64
	8	279	217	13
	9	414	87	8

Table 10: Case Study forecast vote distribution

Table 10 shows the distribution of votes across the case studies. It can be seen, in some instances, that the votes were very close regarding which forecasted risk category the suspect was placed in (case studies 2, 6, and 7). Conversely there are others case studies that show the model was very confident in its forecast (case studies 1, 3, 4 and 9).

The case studies firstly were examined to determine what the final disposal of the offence was, that led to the offender's presenting arrest in 2013, which would have triggered the forecast had the model been in operation at that time. In five instances, no further action was taken in relation to the presenting offence following investigation (Case studies 2, 3, 4, 5, and 7), An out of court disposal was administered for the presenting offence in case study 1, The suspects in case studies 6 and 9 were charged and convicted following the presenting offence and finally in case study 8 the suspect was charged however the charges were withdrawn at court.

Secondly, the case studies were examined to determine the final outcome of the offence that led to the high-risk error for the six high-risk error case studies (1- 6). Case Study 1 relates to an arrest for the offence of rape for which no further action was taken. Case study 2 relates to arrests for arson with intent to endanger life and grievous bodily harm for which the suspect was charged however was found not guilty at court. Case

study 3 relates to an arrest for sexual touching and sexual activity with a child for which no further action was taken. Finally, case studies 3 to 6 (High-risk false positive errors) no arrests were made, all suspects had not been subsequently arrested in the two year follow up period.

Thirdly, the case studies indicate that there is information that is unknown to the model however is known to the police. Information is available in national IT systems such as the Police National Computer (PNC) which is a full criminal record of the suspect and Police National Database (PND) and provides an intelligence picture. It is difficult to know whether the information would have altered the level of risk forecast by the model. There were three areas in which the model was unaware of information the police were aware of, firstly Case studies 1 and 7 highlighted intelligence available via PND on the suspect. Secondly, in case studies 1 and 4 more criminal history information was available from PNC. Finally, in five cases studies (1, 2, 5, 8, and 9) warning markers from PNC were available indicating warnings such as violent, mental health, suicidal and ADHD.

Whilst the case studies provide some context in relation to high risk offenders, practically it is difficult in a random forest model to establish with just nine studies any patterns or characteristics within the groups. With 14,882 custody events across the 2013 validation dataset the results in the next section provide more detail of the characteristics of each forecast risk group.

9.4 What are the descriptive characteristics of the three forecasted risk groups, as measured in the 2013 validation data?

This section presents the descriptive characteristics of the three forecast groups and compares them to one another. Firstly, a statistical procedure called a one-way ANOVA is used to establish the mean values and whether the mean values in the three forecast risk groups indicate the groups are generally different. The ANOVA, however,

does not provide analysis to establish which specific risk groups are different from each other therefore a further test, Tukey HSD, has been used to establish the difference between the risk groups.

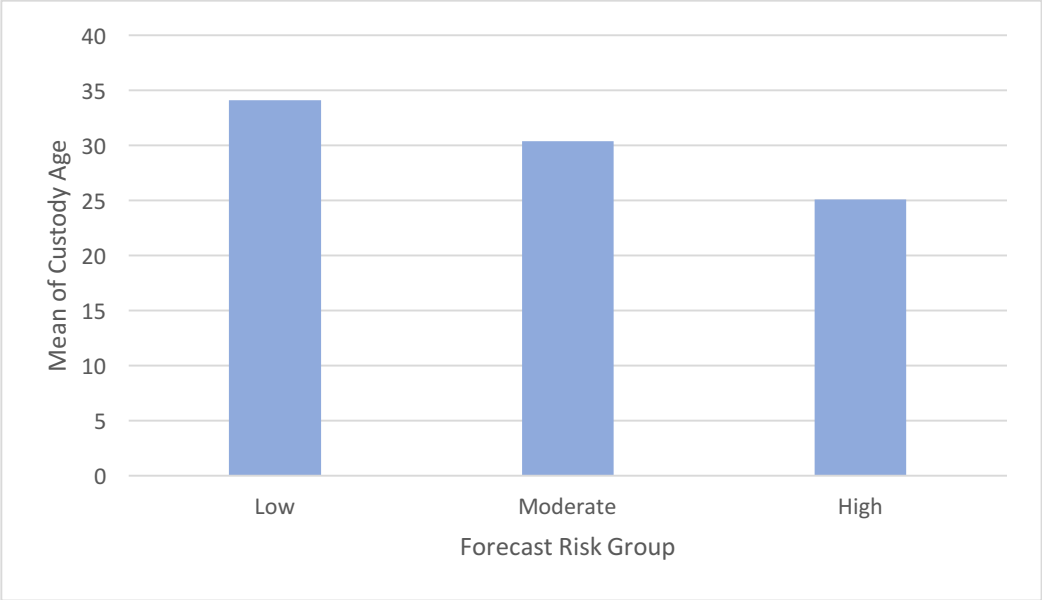

Figure 11: Mean custody age for 2013 validation dataset

Figure 11 presents the mean age of the suspect at the point of the presenting custody event, and indicates the forecast high risk group is on average younger in age. This is not an unusual finding, and gives confidence that the model is identifying the correct people into appropriate risk groups (Farringdon, 2006). Furthermore, Table 11 shows the custody age at the time of the presenting offence and the results highlight that the difference generally between groups is statistically significant and more specifically between groups is statistically significant.

Custody age at presenting offence									
	Mean value				ANOVA Sig.		Significance		
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Custody Age	34.08	30.34	25.09		0.000		0.000	0.000	0.000

Table 11: Age at the time of the presenting offence (Mean value, ANOVA, Tukey Test HSD)

Figure 12 displays the gender split within the forecast risk groups at the time of the custody event. The results show there are much less female suspects in all three categories and the proportion of female offenders are greatest in the low risk category. Conversely the male proportion of suspects is highest in the high-risk group indicating males exhibit more risky behaviour. Additionally, Table 12 shows that the difference generally between groups is statistically significant and more specifically between groups is statistically significant.

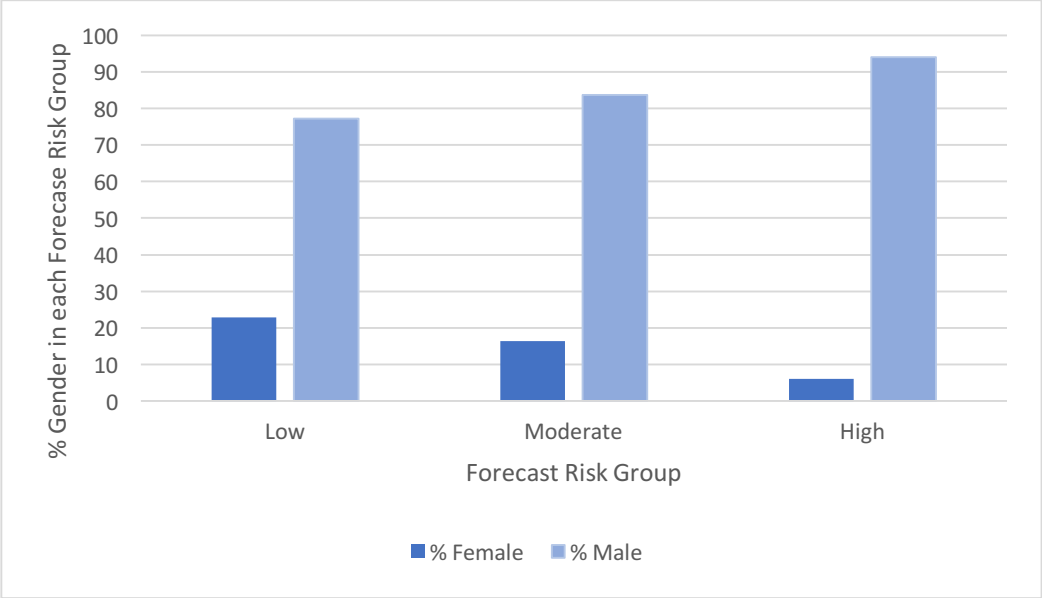

Figure 12: Percentage of gender within forecast risk groups for 2013 validation dataset

Gender									
	Mean value				ANOVA Sig.		Significance		
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Gender	0.77	0.84	0.94		0.000		0.000	0.000	0.000

Table 12: Gender across risk groups (Mean values, ANOVA, and Tukey HSD test)

Table 13 shows the frequency of previous offences and the results highlight that the difference generally between groups is statistically significant in all offences types. Between group tests show all forecast risk groups are statistically significantly different for nearly every category of prior offence counts, other than the previous sexual registration offence category. The table indicates the forecasting model generally places suspects with more prior offending into more serious risk groups.

Frequency of Prior Offences								
	Mean value				ANOVA Sig.		Significance	
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Any offence	0.57	20.24	32.44		0.000		0.000	0.000	0.000
Murder	0.00	0.06	0.02		0.000		0.000	0.000	0.000
Serious	0.05	0.48	1.57		0.000		0.000	0.000	0.000
Violence	0.23	3.34	7.63		0.000		0.000	0.000	0.000
Sexual	0.03	0.06	0.34		0.000		0.000	0.000	0.000
Sexual Reg	0.00	0.00	0.00		0.046		0.052	0.997	0.144
Weapon	0.01	0.41	0.72		0.000		0.000	0.000	0.000
Firearms	0.00	0.04	0.10		0.000		0.000	0.000	0.000
Drug	0.03	1.2	0.83		0.000		0.000	0.000	0.000
Drug Dist	0.01	0.16	0.07		0.000		0.000	0.000	0.000
Property	0.11	9.53	14.52		0.000		0.000	0.000	0.000

Table 13: Number of prior offences (Mean values, ANOVA, and Tukey HSD test)

Figure 13 presents the mean count of prior offending at the time of the presenting custody event. Figure 13 illustrates the higher the prior offence count is, the more likely the suspect would be in the high-risk group.

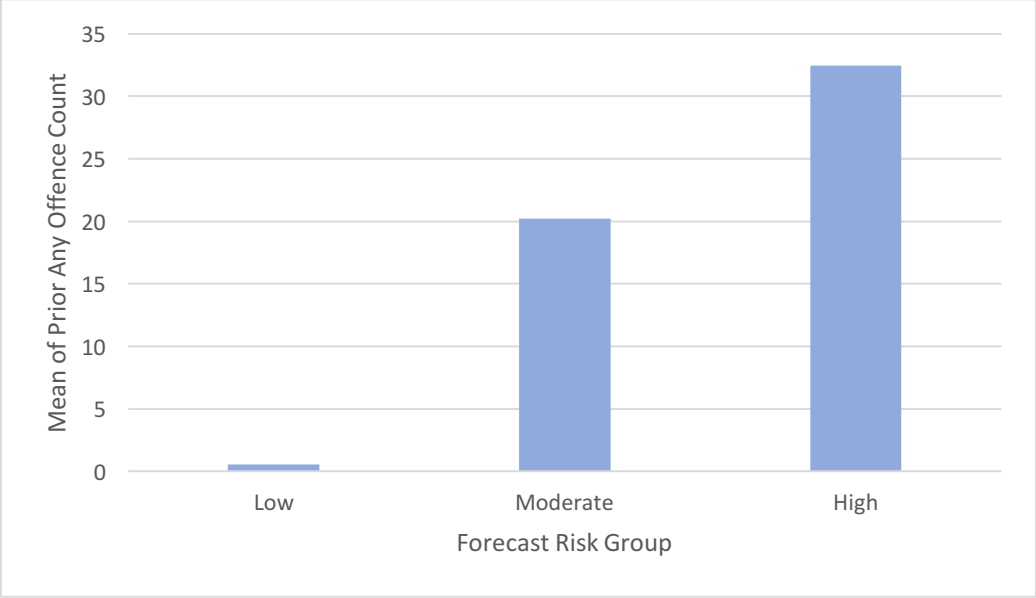

Figure 13: Mean count of any prior offending

Previous murder offences (which includes attempted murder) follow a different directional path to others along with previous drug and drug distribution offences when reviewing mean values. The prior murder offence count, however, constitute a very small number. Prior drug offences and drug distribution offences show that suspects are more likely to be classified as moderate risk based on previous drug and drug distribution offence count. The different directional path followed for drug offences may be an indication of those offenders who are addicted to drugs offending more frequently than other types of offenders but perhaps typically in a non-violent and non-serious manner. Figure 14 overleaf illustrates the drug offence mean count.

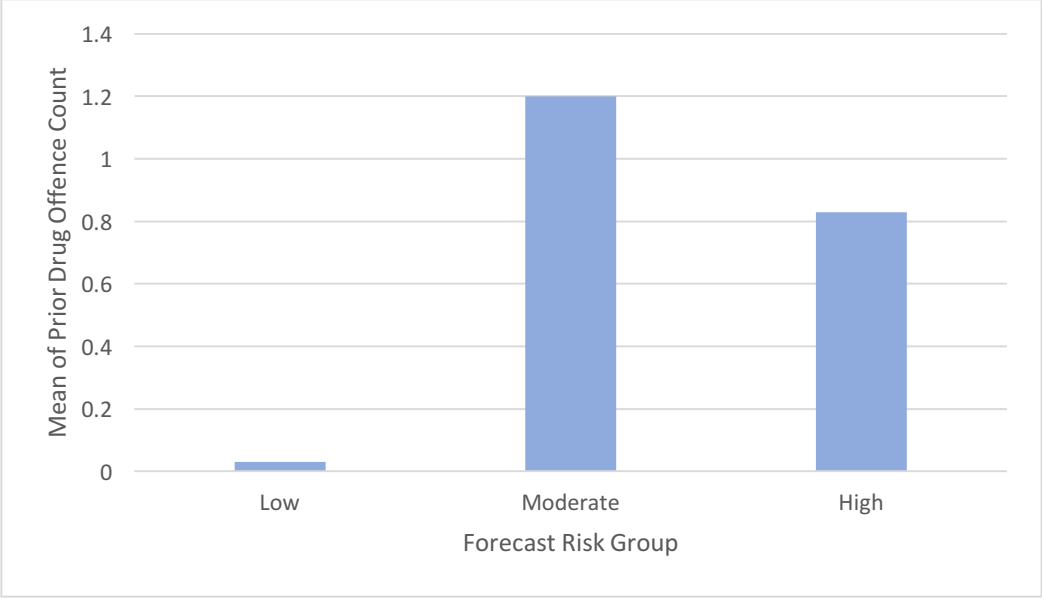

Figure 14: Mean count of prior drug offences

The age at which offenders begin offending, referred to within criminological literature as age of onset (Farringdon, 2006), is significantly different across the three forecasted risk groups. Table 14 shows the difference generally between groups is statistically significant in all categories of offending types. Between group tests show all forecast risk groups are statistically significantly different from one another. The mean values for each offending category all follow the same directional path, indicating that suspects who commence their offending at an older age are generally not as dangerous as those who commence their offending at a younger age. Figure 15 illustrates the directional path for age of onset of offending for any offence, all of which conform to that which would be expected (Farringdon, 2006).

Age at Onset of Offending								
	Mean value				ANOVA Sig.		Significance	
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Any offence	32.225	22.503	17.538		0.000		0.000	0.000	0.000
Violence	32.800	24.365	18.888		0.000		0.000	0.000	0.000
Sexual	36.399	26.964	21.305		0.000		0.000	0.000	0.000
Weapon	30.911	26.161	21.575		0.000		0.000	0.000	0.000
Drug	31.672	25.858	22.735		0.000		0.000	0.000	0.000
Property	30.736	23.016	18.014		0.000		0.000	0.000	0.000

Table 14: Age of onset of offending (Mean values, ANOVA, and Tukey HSD test)

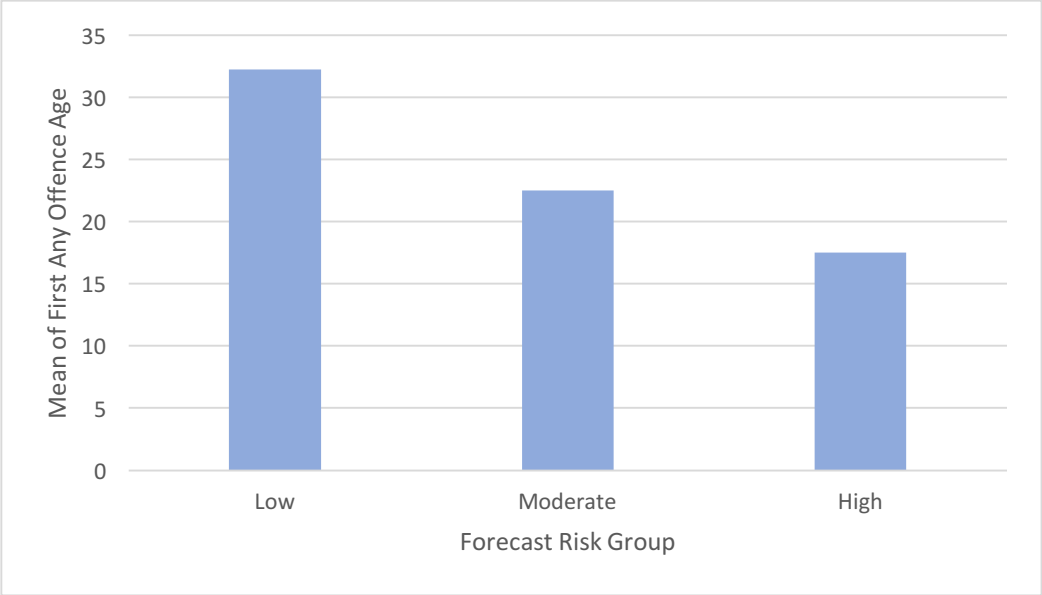

Figure 15: Meant count of the age for the first offence of any type

Table 15 presents the results for the number of years since the most recent offence. Should the offender have had no previous history of any offences, which will apply in many cases, a null return is provided in the data and they are left out from the analysis. The difference generally between groups is statistically significant in all offence categories. Between group tests show all forecast risk groups are statistically significantly different from one another, other than, for serious offences and sexual offences. In both serious and sexual offence categories, the difference between low risk and moderate risk forecast groups is not statistically significant.

Time since last offence								
	Mean value				ANOVA Sig.		Significance	
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Any Offence	4.6918	1.0258	0.4967		0.000		0.000	0.000	0.000
Serious	4.9205	4.793	2.5154		0.000		0.893	0.000	0.000
Violence	4.9203	2.5952	1.1424		0.000		0.000	0.000	0.000
Sexual	3.7174	4.0109	2.9293		0.000		0.723	0.000	0.000
Weapon	6.91430	4.1442	3.0464		0.000		0.000	0.000	0.000
Drug	5.8808	3.1496	2.6496		0.000		0.000	0.000	0.000
Property	5.6898	1.8226	1.0987		0.000		0.000	0.000	0.000

Table 15: Time elapsed since most recent offence (Mean values, ANOVA, and Tukey HSD

In Table 15, the mean values show that, in terms of the time elapsed in years since the most recent offence, that offenders with more recent prior offending generally fall into high risk categories. Figure 16 illustrates the directional path the results follow, other than sexual offences for which there is a marginal difference in mean value.

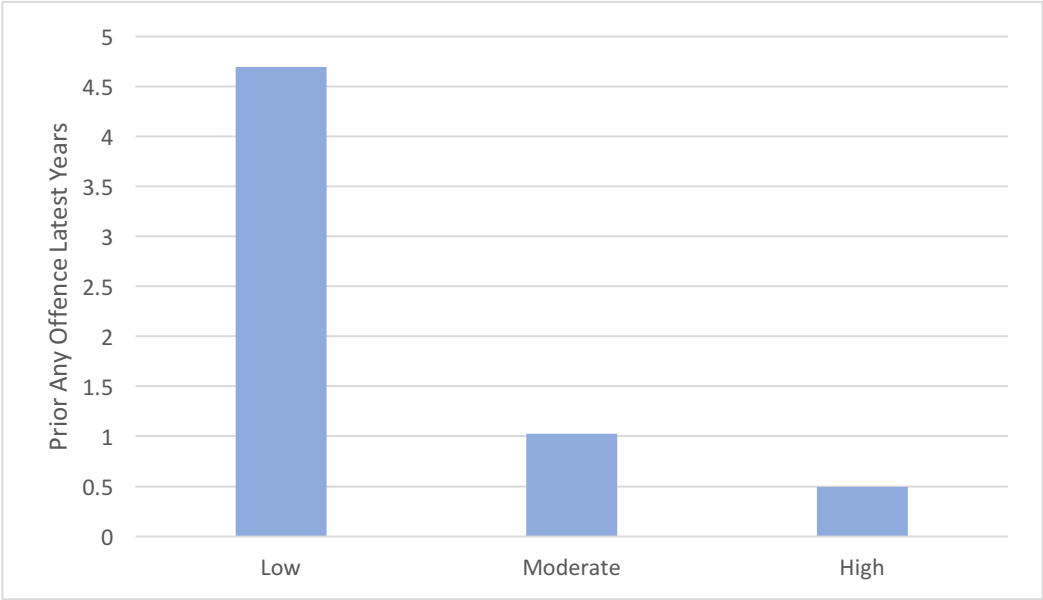

Figure 16: Mean number of years elapsed since last offence

Table 16 provides the findings for the forecast risk group relating to the count of previous custody events and the time elapsed since the last custody event. As was the case with more recent prior offending, the higher forecasted risk groups generally feature offenders who have more recent experience with arrest and detention in Durham's custody suites. The mean values for both prior custody events and time since last custody event are not particularly unusual and follow the expected directional pathway.

Frequency of Custody Events and Time since last Custody Event									
	Mean value				ANOVA Sig.		Significance		
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Prior custody event	0.43	13.88	21.93		0.000		0.000	0.000	0.000
Time since last custody	4.5886	1.0049	0.478		0.000		0.000	0.000	0.000

Table 16: Frequency of prior custody events and time elapsed in years since last custody event (Mean values, ANOVA, Tukey HSD test)

Table 17 shows analyses for the presenting offence relating to violence and property, the mean values indicate the percentage of suspects brought into custody with a violence or property offence. The two measures are collected as a binary outcome of yes or no, therefore the results are presented as prevalence. The results show, in the low risk category, that 41% of suspects present with a violent offence are forecast in the low risk group, this could account for low level assault (common) cases which can often be domestic abuse related. In relation to property, moderate and high risk groups appear to have the same percentage of suspects presenting with a property offence.

Presenting Offence									
	Mean value				ANOVA Sig.		Significance		
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Violence	0.41	0.27	0.34		0.000		0.000	0.000	0.000
Property	0.27	0.47	0.49		0.000		0.000	0.050	0.000

Table 17: Presenting offence (Mean values, ANOVA, and Tukey HSD test)

The number of intelligence record submissions held by the police in relation to a suspect are included in the model as a predictor. The content of intelligence records is a controversial area with the police highlighting such information can speed up investigations and help to identify patterns of crime however others argue intelligence is composed of guesswork, speculation and hearsay. The HART model takes no account of the quality or veracity of the intelligence but includes the count. Table 18 shows for intelligence count, the difference generally between groups is statistically significant. Between group tests show all forecast risk groups are statistically significantly different from one another. Figure 17 presents the mean values which indicate the forecast high risk group has the highest number of intelligence reports.

Intelligence Count									
	Mean value				ANOVA Sig.		Significance		
	Low	Moderate	High				Low vs Moderate	Moderate vs High	Low vs High
Intel Count	1.64	46.58	70.93		0.000		0.000	0.000	0.000

Table 18:Intelligence submissions (Mean value, ANOVA, and Tukey HSD test)

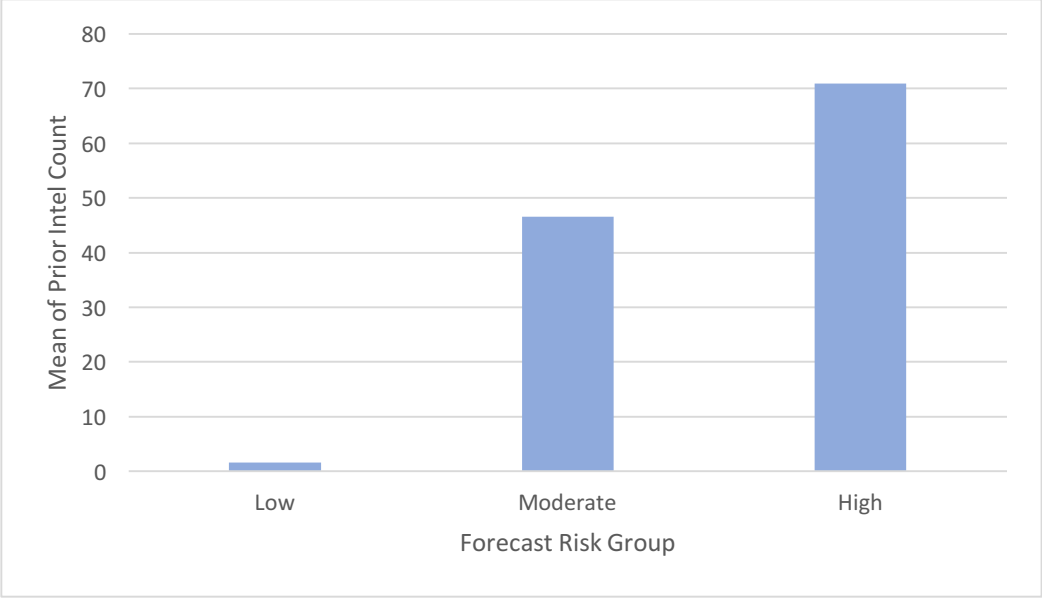

Figure 17: Mean count of prior intelligence record submissions

Finally, the variable relating to the top 25 postcodes is used as a predictor in the model. Figure 18 depicts in descending order the top 25 most common outward postcodes in County Durham which have the highest percentage of forecast high risk offenders within the postcode. Separately, a chi-squared test was conducted to establish whether the outward postcode is significantly related to forecast outcome. The results show chi-squared = 1611.0, d.f.= 54, p=.000 and is therefore statistically significant.

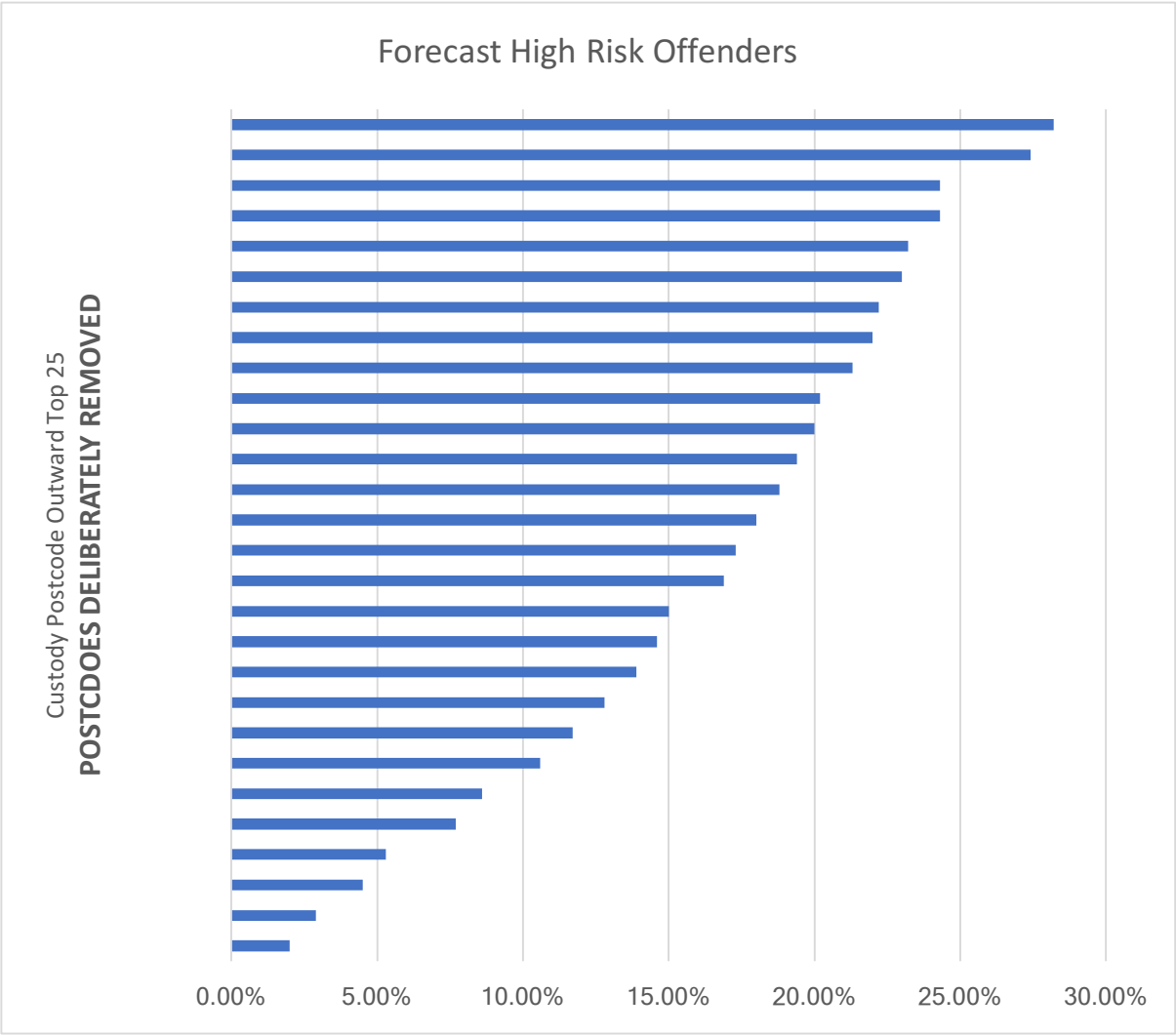

Figure 18: Top 25 postcodes of forecast high-risk offenders for 2013 validation dataset

To summarise this section, the characteristics presented for each forecast risk group are generally consistent with the criminological theory. All ANOVA tests for all 34 predictor variables have not been presented here, as they can be largely summarised into the fact that all groups are statistically significantly different from one another. The mean values within group show that those forecast high risk correspond to expectations of serious offending. The graphs of mean values for the different characteristics of the three forecasted risk groups are presented in Appendix D.

9.5 To what extent do the clinical forecasts of custody officers agree with the model-generated algorithmic forecasts?

This section analyses the extent to which clinical police custody officers agree with the algorithmic forecasting model and vice versa. A matrix of agreement is provided in Table 19 which details overall the percentage of clinical police forecasts in the three risks groups and the percentage of model forecasts within each risk group. There were 888 custody events in this sample.

Forecast Agreement Matrix		Police High Risk		Police Moderate Risk		Police Low Risk	Total
Model High Risk	A	1.58%	B	11.49%	C	2.03%	15.09%
Model Moderate Risk	D	3.49%	E	39.86%	F	13.29%	56.64%
Model Low Risk	G	1.35%	H	12.16%	I	14.75%	28.27%
Total		6.42%		63.51%		30.07%	100.00%

Table 19: Agreement Matrix 2016

Table 19 shows overall levels of agreement coloured in green – the total overall agreement is (A + E + I) 56.19%. The highest level of agreement is in the moderate category, at 39.86%, with the least amount of agreement between the two forecasts being in the high-risk category which is 1.58%. In the low risk category, the agreement overall is at 14.73%. Police custody officers generally appear to be reticent to use the high-risk category. The percentage of high risk forecasts for each risk group illustrates the police forecasting high risk 6.42% of the time with the model predicting high risk 15.09% of the time. The police forecast a higher proportion of moderate and low risk arrestees than the model, and in the high-risk area that the model forecasts more frequently than the police. Having summarised Table 19 using overall forecasts, the model and the police appear to forecast low risk approximately the same number of times, however, Table 20 over leaf shows the extent to which there is agreement in each individual risk group rather than the overall figures.

Description	Percentage
Of those forecast by the model as High risk percent of occasions the police agreeed	10.45%
Of those forecast by the model as Moderate risk percent of occasions the police agreed	70.38%
Of those forecast by the model as Low risk percent of occasions the police agreed	52.19%
Of those forecast by the police as High risk percent of occasions the model agreed	24.56%
Of those forecast by the police as Moderate risk percent of occasions the model agreed	62.77%
Of those forecast by the police as Low risk percent of occasions the model agreed	49.06%

Table 20: Extent of agreement within forecast risk groups

Table 20 shows the agreement within each risk group both for the model and the police forecasts. When the model forecasts high risk, the police agree with this forecast 10.45% of the time. Conversely, when the police forecast high risk the model agrees with the forecast 24.56% of the time. There is clear disagreement between the model and the police in the high-risk category both in terms of the volume of forecasts in Table 19 and in terms of the agreement levels in Table 20.

The agreement in the low risk category is virtually the same whether between model and police or vice versa at 52.19% and 49.06% respectively. In the low risk category, there is effectively agreement only half of the time, remembering that, in the overall agreement matrix in Table 19, both police and the model forecast low risk at approximately the same rate: 30.07% and 28.27% overall.

The highest levels of agreement exist in the moderate risk category group. Table 20 illustrates when the model forecasts moderate risk the police agree with this forecast 70.38% of the time. However, the level of agreement falls when the police forecast moderate. The model agrees with the assessment 62.77% of the time. Whether this enhanced agreement is caused by a better understanding among custody officers of moderate risk offenders or whether it is largely driven by their clear preference for moderate forecasts, remains to be seen.

Due to the number of custody officers over a 24-hour period and the number of custody suites a force has, coupled with the turnover of custody officers in the environment, there is a high likelihood that the decision making of custody officers is inconsistent, the forecasting model conversely will consistently make forecasts based on the previous decisions of over 104,000 custody officer's decisions used to build HART. The next chapter discusses the results further.

10.0 Discussion

The results presented in the previous chapter provided interesting findings in the context of custody officer decision making consistency and the model. Accuracy and error distribution of the model in validation was also presented. The validation year was a somewhat different type of offending cohort, with increased serious offending observed in the two year follow up period. The model was handling an offending cohort different to that with which it had been constructed. Accuracy fell in validation and whilst these results are concerning initially, when taking account of the unique way error types are distributed, one can see the model became more cautious in its forecasts. Furthermore, the model continued to avoid the most dangerous error at nearly an identical rate seen in the construction data. The structure of this chapter will firstly take the research questions in turn before discussing fairness and research limitations.

10.1 What is the validated accuracy of the Durham Constabulary forecasting model using custody events from 2013 compared to 2008-2012 construction sample?

The 2013 validation dataset was dropped down the forecasting model to establish what the forecast of dangerousness would have been had the forecasting been live at that time. The 2013 data was not used in the construction of the model. Sufficient time has elapsed to allow for the two year follow up period, which is in accordance with the time horizon of the forecast. Therefore, the actual outcomes for the forecasts could be established and compared to the outcomes predicted by the model.

The 2013 validated accuracy overall of the model was 62.80%, which constitutes a drop from the construction data of 5.7% points. In any validation exercise of a forecasting method, there is always a reduction in accuracy in a fresh and independent validation sample. The fall occurs across all forecasting methods, including both traditional regression methods and machine learning. There is no research comparison

for a police arrestee setting of a random forest forecast model or indeed a validation. The validation study of a random forest machine learning approach in a US probation setting however, offers an opportunity to compare due to the same forecasting method being used (Barnes and Hyatt, 2012). The Barnes and Hyatt (2012) validation study shows an overall accuracy of 57.8%, which represents a reduction in accuracy of 8.3% points, this is greater than the reduction seen in the Durham model. The validation dataset used by Barnes and Hyatt (2012) was with an older validation sample which, at the time, was just under a decade old; in this research, we have used the full year of data where the presenting events occurred immediately after the construction data.

There are two ways to review the accuracy of the model (Table 6), firstly those forecast as high risk who were forecast correctly, however, this may not be the appropriate way to assess accuracy. The model is specifically built to avoid a very dangerous error with cost ratios. In doing so the model will place more offenders into the high-risk group to avoid an error. Therefore, the accuracy will be affected due to this overestimation of risk as desired by the organization. Another way and arguably a more reasonable way to review accuracy is to review those who were actually high risk that were forecast correctly by the model. In doing so, the accuracy in this validation presents the greatest reduction, in the high-risk category of those actually high risk who were forecast correctly, with a 20%-point drop from 72.60% to 52.75%. The cohort of suspects in 2013 was different from construction with higher prevalence and frequency of serious offending from that which it had been constructed. The differing cohort of offending that was presented to the model is likely to have affected the forecasting accuracy. Many may consider the reduction of high risk accuracy to be unacceptable. Alternatively, others would perhaps point to whether the level of accuracy is better or worse than current clinical judgement (Neyroud, 2015).

In light of the reduction seen in accuracy, consideration of the frequency of refreshing and rebuilding the forecasting model to reflect the differing conditions may be

prudent, together with the associated costs. The follow up period after 2013 appears to be changing in relation to recorded crime, and the prevalence and frequency of serious offending behaviour is starting to increase (Figures 5 – 9). The model is forecasting behaviour based on construction data (2008-2012) that may have exhibited less serious criminal behaviour. That said, as highlighted earlier, had the model been presented with very similar data to that of the construction data, decreases in accuracy would still have been seen in the validation the extent may or may not have been greater (Barnes and Hyatt, 2012).

10.2 What is the distribution of forecasting errors using the 2013 validation data compared to 2008-2012 construction sample?

The random forest model takes account of the weighting of different types of errors (Berk et al., 2009; Berk, 2012; Barnes & Hyatt, 2012). The error distribution shows that in validation overall, cautious errors had a larger increase than dangerous errors. In relation to cautious errors, the increase was 5.00% points higher in validation compared to the construction sample and in relation to dangerous errors whilst there was an increase this was by a much smaller margin of 0.7%. The model was more cautious in validation when presented with the different conditions to that with which it had been constructed.

Overall the validation presented 1% very dangerous errors and very cautious errors were 2%. This represents no change when comparing construction with validation. The model was built to avoid very dangerous errors and the model achieved that whilst overall accuracy declined. As highlighted within the literature review, the organisation ultimately decides on the cost ratios built into the forecasting model (Berk et al., 2009; Barnes and Hyatt, 2012; Berk, 2016; Berk et al., 2016). Despite the accuracy falling, the model was able to respond as desired. This is due to the cost ratios built in to the random forest model, to ensure the worst errors were minimized, and this was achieved to the extent that the figures remained unchanged.

It is the cost ratios, in large part, enable the model to respond in a way that minimizes the least desirable, very dangerous error. Cautious errors indicate a forecast higher than the actual offender behaviour and dangerous errors indicate forecasts that were lower than the actual offender behaviour. The organisation would prefer cautious errors over dangerous errors. The validation assessment has shown the cautious to dangerous error ratio increased from 1.94 to 2.56, therefore the model was 2.56 times more likely to be cautious. The very cautious to very dangerous error ratio also increased from 2.19 to 3.07 demonstrating the model is now 3.07 times more likely to be very cautious. The increases in cost ratios show the model became much more cautious and conservative in its forecasts.

The cost ratios meant in relation to very dangerous errors, in validation, that we can be confident - with 98% accuracy - they will not occur. The model was able, despite the changing conditions in 2013 follow up period, to ensure those forecast low risk whose behaviour eventually revealed them to be actually high risk remained at 2%, which offers a great deal of reassurance.

Whilst the model became more cautious, for some this raises ethical questions over deliberately overestimating the forecasting of individuals as high risk, and the impact that may have on a suspect when the organisation is aware that a proportion are not, based on their actual behaviour. For others, however, protecting the public from the risk of high harm by minimising very dangerous errors is a priority and is ethically the appropriate route to take. It would be difficult to believe one human being could be 98% accurate in forecasting that the worst kind of error would not occur. Indeed, it is difficult to believe that a number of different custody officers, over the course of a year, could adjust their own intuitive decision making to the aggregated changing patterns of offending behaviour. Therefore, the forecasting model offers consistency of decision making.

It is useful to know when it goes wrong, what those kinds of errors look like, and the case studies provide that context. The case studies demonstrated there is information that is unknown to the current model and the inclusion, or exclusion, of that information needs to be fully understood. An issue arising from the case studies is that the model is limited to Durham Constabulary data. Therefore, if the suspect is from a neighbouring force area, or travels around the country, the model will be blind to the information, as only the individual force records have been used. This is one reason why the forecasting model is a decision support tool; the model cannot know every piece of information. Police may have more information from national criminal records systems or local partner agencies that may change a decision.

10.3 What are the descriptive characteristics of the three forecasted risk groups, as measured in the 2013 validation data?

The descriptive characteristics of the three forecast risks groups largely follows existing evidence when considering criminal behaviour and the suspects past offending history (Berk, 2012; Barnes and Hyatt, 2012). Although the characteristics do not generally represent anything out of the ordinary, it is helpful to consider, from a policing perspective, the characteristics associated with each group to satisfy a desire to ascertain whether they reflect the organisational perception of offending (Berk, 2016).

Generally, the age of onset in all offence types show that those in the high-risk group tend to have a younger age of onset. Those who are high risk also generally have a higher previous offending history across offence types. Those in the low risk group have a longer period of time between offences than those in high risk, and so on. Essentially, the analysis has shown that the characteristics of the three forecast risk groups are statistically significantly different from one and another. It is again reassuring that the descriptive characteristics of the forecast risk group reflect what the police organisation would expect to see regarding offending behaviour.

Further predictors can be added to random forest models, and there is no limit to the number of variables that can be added to such models (Berk et al., 2009; Berk, 2012; Barnes and Hyatt, 2012). It is worth noting here that the addition of a predictor variable is likely to improve accuracy. The degree of improvement may be small (Barnes and Hyatt, 2012; Berk, 2012), set against the effort and cost to obtain the data quality necessary to add the predictor to the model. Therefore, as with cost ratios above, an organisation would need to balance the degree of effort needed to obtain the new predictor variable data to a sufficient data quality standard against the potential improvement in accuracy that would be achieved.

10.4 To what extent do the clinical forecasts of custody officers agree with the model-generated algorithmic forecasts?

The risk forecasts made by the custody officers are important, as they are decision makers at a key stage in the Criminal Justice System (CJS). The results of comparing custody officer and HART forecasts totalled 888 custody events available for comparison. The literature review highlighted that statistical forecasts generally outperform human judgement (Meehl, 1954; Dawes et al., 1989; Grove et al., 2000; Ægisdóttir et al., 2006; Kahneman, 2011). It should be noted this type of decision making concerning offender dangerousness over a two period in a police custody environment has not been tested. The limit of the comparison, at this stage, is purely an assessment of the extent to which the forecasts agree; future research will establish whether the forecasting model or the police are more accurate.

There was clear disagreement in all risk areas forecast when comparing the clinical judgement and algorithmic forecast. The model forecast high risk more than double the number of times the police did. In the high-risk area, the levels of agreement were between 10% and 24%. Conversely, in the low risk area, the police and the model forecasted a similar volume of suspects overall. Even here, however, the agreement was generally 50%. In an area, such as low risk, where we would seek to minimize the

worst type of very dangerous error, to observe such disagreement is concerning. Similarly, when future research concentrating on the accuracy of such forecasts becomes apparent, the information may present some uncomfortable reading. The model makes double the number of high risk forecasts, which appears to suggest the model is tougher. However, the model is also designed to over-compensate and will predict high risk more often, thereby minimizing the worst outcome – high-risk false negative or very dangerous error.

The area of moderate risk had the highest levels of agreement between clinical and algorithmic forecasting. Notwithstanding the levels of agreement in the moderate risk group the police forecast of moderate had the highest volume of all categories at 63%. Which could suggest that when officers assess dangerousness, cost ratios do not seem to be operating in the same way as the forecasting model operates. Put in simple terms officers simply do not wish to make a mistake in such a consequential arena of decision making, by making any kind of error, and therefore may seek to avoid mistakes at either extreme end of the criminal behaviour spectrum. The moderate risk category is therefore chosen over the option of suggesting the offender will be unlikely to offend in the next two years (low risk), with the consequences to future victims. Equally the consequences of saying a suspect is likely commit a serious offence in the next two years (high risk) on the suspect are such that the custody officer, in being unsure, will stay with a moderate risk forecast.

The discussion here merely sets the scene for future research in this arena. As the results have illustrated, we are not aware of the boundaries of the custody officer's skills in this area, but place consequential decisions in their hands (Kahneman, 2011). Whilst forecasting models should never remove the discretion police custody officers have, at the very least they may offer better bounded decision making than current guidelines – read or unread – that currently exist (Goldkamp, 1987; Raynor et al., 2000; Austin et al., 2003; Harcourt, 2007; Berk, 2012; Slothower, 2014; Neyroud, 2015; P

Neyroud 2016, personal communication, 29 September). Discretion must exist for police officers as the forecasting model will never be 100% accurate and officers may be aware of information the model is not.

10.5 Fairness

There are concerns regarding algorithmic models such as random forest forecasting model in a CJS setting, as highlighted in the literature review (Harcourt, 2007; Harcourt, 2010; Hannah – Moffatt, 2013; Harcourt, 2015; Starr, 2014; Angwin et al., 2016; Reyes, 2016; Naughton, 2016). The theme of the arguments for and against such algorithmic forecasting centres generally around fairness. An ideology of what fairness should look like in the future, for example treating people fairly and equally, as opposed to recognising the difficulty the current real world practice, such as biased intuitive judgement, that exists within decision making (Berk, 2016).

Fairness in the future, some would argue, concerns ensuring variables in the model do not mean unequal treatment as highlighted in the literature review. By conducting validation studies such as this thesis and building forecasting models in a transparent way an organisation can see the impact of particular variables and make an informed decision about what is most important. Arguments concerning ideological fairness exist within the current human processes however human judgement may be no less unfair (Kahneman, 2011). No model will be perfect, but comparing it to ideal perfection is unfair when current practice may perform worse.

Since, assessing offender dangerousness concerns protecting the public, the central question is whether a predictor produces sufficient benefit that it outweighs the bias that might result. An organisation is in control and can decide not to have such information in a forecasting model of dangerousness. In doing so it is critical to understand that the decision to exclude data may lead to errors and reductions in accuracy elsewhere, which of course may be a risk an organisation chooses to take in

order to achieve what it perceives is fair and equal treatment to all suspects (Berk, 2016).

10.6 Research Limitations

A limitation of the forecasting model was highlighted with the results of the case studies, which demonstrated the model is blind to certain information. The model is built utilising only Durham Constabulary data from its own recording systems. Clearly data on suspects can be available in neighbouring police force areas which may affect the outcome classification of the suspect, had the model been aware of the data. If such models had access to national (or even regional) offending data or intelligence such as PNC or PND information in the UK, this issue could be largely overcome. The lack of available information, in large part, is the reason why the forecast model should not remove discretion from the police in this key decision making area. The model may also be unaware of information in other public sector organisations than the police.

The fourth research question enables an understanding of the levels of agreement between the police and the forecasting model, based on a separate exercise to have a sample from which conclusions could be reasonably drawn. This research question required the compliance of custody officers, both to generate the forecast by the model (although they were not shown the results), and in producing their own clinical forecasts of the offenders' future behaviour. Compliance with this process was uneven. This thesis does not provide room to discuss all implementation difficulties in obtaining the sample. The problems encountered involved both the computer interface that the custody officers were asked to use, and their willingness to add this task to an already long list of duties. These problems were solved over time but not until late in the thesis process. If compliance was not relatively high clearly there is potential for missing clinical data. There is also potential for custody officers, because of their own biases or

being uncomfortable with forecasting, choosing not to forecast certain types of offender or risk category.

Other evidenced based policing research in the constabulary meant that the sample size was substantially reduced for question four by not including juveniles in the sample. Albeit, the forecasting model was built to enable juvenile forecasts. Future research would benefit from ensuring that juveniles are included in a comparison of clinical and algorithmic decision making in the custody environment. Evidence and theory point to the fact younger offenders are more likely to see their offending escalate which makes it particularly relevant to include juveniles. Research concerning whether custody officers make different forecasts for juveniles than they do for adults would contribute to whether juveniles are treated fairly and consistently at the gateway to the CJS.

10.7 Summary

In summary, whilst the results show accuracy fell in validation, the model adjusted to a change in the offender cohort to that which it had been constructed and the reaction was to become more cautious and conservative in forecasting dangerousness. The reason for the caution is based on Durham Constabulary's desire to minimise the worst type of error likely to cause high harm in the communities Durham Constabulary serve. Therefore, the model carried out its role well in validation, showing a maintenance of 98% accuracy that the high risk of high harm would not occur, whether a human being could forecast with such precision when presented with an aggregate pattern of changing offender behaviour is unlikely.

The clinical and algorithmic forecasts show a large disparity in terms of the volume of forecasts for different risk groups but also show large disparity in terms of agreement on a risk between the two. It will be interesting in future research to see which produced more accurate forecasts. The next chapter provides concluding

comments and summarises the policy implications and future research which build upon this this study.

11.0 Conclusions

In returning to the central topic of this research, the literature review highlighted that statistical methods have been used over decades, and have consistently demonstrated more accuracy than clinical methods (Meehl, 1954; Dawes et al., 1989; Grove et al., 2000; Ægisdóttir et al., 2006; Kahneman, 2011). The random forest, machine learning method of forecasting risk used in this study offers some unique features. The method identifies and makes use of non-linear relationships which help to improve accuracy. The method also has the ability to factor differing types of errors and cost ratios. Such cost ratios clearly appeal to those responsible for consequential decisions in the Criminal Justice System (CJS), (Berk, 2012).

The study assessed the accuracy of a random forest forecasting model with an independent validation dataset. The analysis demonstrated that overall accuracy fell when faced with changing offender behaviour patterns from 2013. Nevertheless, the random forest model, in taking account of cost ratios, ensured the errors were of a more cautious nature. Essentially, the model became more conservative in its forecasts of risk to minimise the potential of a very dangerous error. Despite the changing offending behaviour patterns during the 2013 follow up period, due to the unique balancing of cost ratios and error types, the model was able to ensure the likelihood of very dangerous errors occurring was just 2%. Therefore, the organisation can be 98% sure that a very dangerous error will not occur.

Despite the strength of evidence relating to clinical versus statistical decision making, no research at the gateway to the CJS with such a forecasting model has been conducted in the dynamic environment of a custody suite in UK policing. Therefore, this study sought to commence that research and implement a wilful blindness exercise, with custody officers producing forecasts whilst the model forecast was not revealed, that

may assist such research in the future. At this stage, this study can merely show the levels of agreement that exist between the forecasting model and the police.

Wilful blindness demonstrated clear differences in forecast risk levels between the police and the forecasting model, which were stark in the high risk and low risk areas. This difference is concerning particularly in light of previous research surrounding clinical and statistical forecasting. The analysis suggests there may not be any cost ratios operating when the police make decisions. Without cost ratios, future research may identify the police are more accurate at identifying high risk suspects however proportionately may make more very dangerous high harm errors.

11.1 Policy Implications

The levels of disparity between the model and the police highlight a primary focus centres on consistency. Decisions at the gateway to the CJS allow for discretion in relation to bail (conditional or unconditional), out of court disposal, and charging a suspect to appear at court (Neyroud, 2015). This study has presented literature regarding the accuracy of clinically biased decision making. Policing is an around the clock job, with more than one individual making clinical decisions, therefore the collective biased judgements of not one but many custody officers contribute to decision making across any given year at the gateway. The forecasting model provides decision support around the clock at the gateway to the CJS without biased heuristic judgement influencing the forecast.

This study has presented the stark disparity in agreement levels, suggesting a potential risk aversion of custody officers who when deciding the dangerousness of suspects forecasted moderate more frequently when asked their opinion. Clearly, in this decision-making environment the police view can result in an out of court disposal for a suspect that is potentially dangerous or charging an offender to court that is unlikely over

the next two years to be rearrested. A forecasting model being adopted in a custody environment will offer consistent decision support in a consequential environment.

The forecasting model is still wilfully blind, with custody staff still unable to see the direct forecasts produced by the model. Therefore, it is not known how the custody officers will react when they are presented with one of three risk groups. Custody officers may choose to use it when it confirms their own preconceived notions or disregard the forecast when it fails to confirm their biases. The outcome remains to be seen and is for future research. Nevertheless, the author suspects the presentation of a forecast risk group will attract more police and CJS attention to those that affect community safety the most.

The final decision must be ultimately made by the custody officer, who may be in possession of more information than the model. This research has demonstrated, through the case studies, suspects can cross police force boundaries and other agencies can hold information the model is unaware of. Such information may alter the decision and it will, if added to the model, improve the accuracy of the forecasts. Exploring what readily available diverse data sources exist, with potential to include such information in a model, will improve accuracy and contribute to minimizing dangerous errors to support critical decisions.

By making a policy decision to utilise a forecasting model, there is potential for demand reduction. This must be balanced, however, against the costs of the creation of the model, monitoring offending patterns and refreshing the model, together with the implementation costs to the organisation. If custody officers do not disregard the model forecasts, and use the forecast to support their final disposal decision, there will be improved consistency of out of court disposals. This has the potential to reduce the number of cases in the CJS. Other agencies such as CPS, Courts, Probation and the Legal Aid budget may also benefit from a reduction in demand.

Often, when demand reduction is discussed with the introduction of such models, critics can centre their attention on a distaste for using such models, suggesting the aim is solely to achieve demand reduction and reduce costs. In UK policing, there continues to be austerity measures which have meant a reduction in resources. By introducing such a model, there may be a reduction in demand by ensuring the right cases are taken to court and the most dangerous offenders are targeted to reduce reoffending and minimise harm. These models have the potential to reduce demand, which is desperately needed in UK policing and will enable the police to deploy resources more efficiently which will ultimately benefit our victims and communities.

Within the policing environment, there are various points at which the police are assessing risk to predict where harm is likely to occur in order to prevent crime with the aim to reduce harm. Similar random forest forecasting models used could benefit other areas in policing where risk is assessed, such as, sexual offender assessments and domestic abuse. If we focus police resources on those who are most likely to not only reoffend but reoffend in a serious or dangerous way, then we are adhering to the Peelian principles of preventing the worst amount of harm to society (Home Office, 2012).

11.2 Future Research

Whilst this is the first time a random forest model has been used in the CJS in the UK, it joins a growing list of areas that are using random forest forecasting. The work of Berk, et al. (2009); Sherman, et al. (2012); Neyroud, (2015), and Barnes and Hyatt, (2012) set the scene for this type of research and future research should build upon this study by tracking the validation cohort and the wilful blindness cohort of suspects in terms of accuracy and reoffending.

Forecasting decision support is not a new concept, if one accepts that decision support has existed for decades. There must also be an acceptance that, as time passes, more innovative modern decision support tools will become available. In

forecasting risk of harm in a more accurate and consistent way, reduces the uncertainty about the future. The forecasting model will not be 100% accurate. However, by using and quantifying big data within a random forest forecasting model this author opines more certainty and consistency regarding future events. By taming uncertainty in this way, a decision is made with the support of the forecasting model and can be justified. A decision is justified and informed by a risk category that considers a vast amount of data before arriving at an outcome. With the knowledge of the risk category, a decision is justifiably made at the point of initial custody disposal.

With decreasing resources and the costs associated with placing offenders into the CJS, and having invested in new superior forecasting methods an organisation can then begin testing constructive interventions targeted at the right risk level. In doing so, the organisation can establish what works in interventions targeted at specific risk groups to reduce reoffending and encourage desistance from crime and minimize harm to future victims. (Sherman et al., 1998; Sherman et al., 2012; Hannah-Moffat, 2013; Sherman, 2013; Hyatt and Barnes, 2014; Neyroud, 2015).

Neyroud (2015) argues that an evidence-based approach to the gateway to the CJS is critical to the effectiveness of the CJS and is 'urgently necessary' (Neyroud 2015 p. 12). In reducing crime and rehabilitating offenders to desist from crime, the police work in accordance with Peelian principles to prevent crime, which in turn keeps our victims and wider communities safe (Home Office, 2012). The prevention of crime, and apprehension of the offender, together with their rehabilitation and conviction as identified by HMIC (2016), 'are among the highest obligations of the state in the discharge of its duty to protect citizens' (HMIC 2016 p 6). It is therefore incumbent upon the police to ensure that if a more accurate method of effectively targeting dangerous offenders exists to minimise high harm in the community, it should be fully explored within an evidence-based framework.

12.0 References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. May 23, 2016. 'Machine Bias – There is software used across the country to predict future criminals. And its biased against blacks'. Retrieved 11th December 2016 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Austin, J., Coleman, D., Peyton, J. and Johnson, K.D. (2003) Reliability and validity study of the LSI-R risk assessment instrument. *Washington: The Institute on Crime, Justice, and Corrections, George Washington University*.

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G. and Rush, J.D. (2006) The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), pp.341-382.

Barnes, G.C., Ahlman, L., Gill, C., Sherman, L.W., Kurtz, E. and Malvestuto, R. (2010) Low-intensity community supervision for low-risk offenders: a randomized, controlled trial. *Journal of Experimental Criminology*, *6*(2), pp.159-189.

Barnes, G. and Hyatt, J.M. (2012) *Classifying adult probationers by forecasting future offending*. BiblioGov.

Berk, R. (2012) *Criminal justice forecasts of risk: a machine learning approach*. Springer Science & Business Media.

Berk, R. (2016) 'A Primer on Fairness in Criminal Justice Risk Assessments'. (Unpublished, University of Pennsylvania).

Berk, R.A. and Bleich, J., (2013) Statistical procedures for forecasting criminal behavior. *Criminology & Public Policy*, *12*(3), pp.513-544.

Berk, R.A. and de Leeuw, J. (1999) An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association*, *94*(448), pp.1045-1052.

Berk, R. and Hyatt, J. (2015) Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, *27*(4), pp.222-228.

Berk, R., Li, A. and Hickman, L.J. (2005) Statistical difficulties in determining the role of race in capital cases: A re-analysis of data from the state of Maryland. *Journal of Quantitative Criminology*, *21*(4), pp.365-390.

Berk, R.A., Kriegler, B. and Baek, J.H. (2006) Forecasting dangerous inmate misconduct: An application of ensemble statistical procedures. *Journal of Quantitative Criminology*, *22*(2), pp.131-145.

Berk, R., Sherman, L., Barnes, G., Kurtz, E. and Ahlman, L. (2009) Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), pp.191-211.

Berk, R.A., Sorenson, S.B. and Barnes, G. (2016) Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions. *Journal of Empirical Legal Studies*, *13*(1), pp.94-115.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), pp.199-231.

Dawes, R.M., Faust, D. and Meehl, P.E. (1989) Clinical versus actuarial judgment. *Science*, *243*(4899), pp.1668-1674.

Farrington, D. P. (2010) Developmental and life-course criminology: Theories and policy implications. In DeLisi, M. J. and Beaver, K. M. (Eds.) Criminological Theory: A Life-Course Approach. *Sudbury, Massachusetts: Jones and Bartlett* (pp. 167-185).

Gladwell, M. (2005) *Blink*, London: Penguin

Goldkamp, J.S. (1987) Prediction in criminal justice policy development. *Crime and Justice*, pp.103-150.

Gottfredson, S.D. and Moriarty, L.J. (2006) Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, *52*(1), pp.178-200.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. and Nelson, C. (2000) Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1).

Hannah-Moffat, K. (2013) Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly*, *30*(2), pp.270-296.

Harcourt B.E. (2007) *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

Harcourt, B.E. (2010) Risk as a proxy for race. *Criminology and Public Policy, Forthcoming*.

Harcourt, B.E. (2015) Risk as a Proxy for Race. *Federal Sentencing Reporter*, *27*(4), pp.237-243.

Her Majesty's Inspectorate of Constabulary (2016), State of Policing: The Annual Assessment of Policing in England and Wales 2015. Online : https://www.justiceinspectorates.gov.uk/hmic/wp-content/uploads/state-of-policing-2015-double-page.pdf

Holt, R.R. (1958) Clinical and statistical prediction: A reformulation and some new data. *The Journal of Abnormal and Social Psychology*, *56*(1).

Home Office (2012), *Policing by Consent; Robert Peel's 9 Principles in Policing*. Retrieved 11th December 2016 from https://www.gov.uk/government/publications/policing-by-consent/definition-of-policing-by-consent

Hyatt, J.M. and Barnes, G.C. (2014) An experimental evaluation of the impact of intensive supervision on the recidivism of high-risk probationers. *Crime & Delinquency*, p.0011128714555757.

Kahneman, D. (2011) *Thinking, fast and slow*, Macmillan: Penguin.

Kahneman, D. and Klein, G., 2009. Conditions for intuitive expertise: a failure to disagree. *American Psychologist*, *64*(6), p.515.

Meehl, P.E. (1954) *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence,* 3rd ed., University of Minnesota: Oxford University Press

Monahan, J. and Skeem, J.L. (2013) Risk redux: The resurgence of risk assessment in criminal sanctioning. *Virginia Public Law and Legal Theory Research Paper*, (2013-36).

Naughton, J., 26 June 2016. Even algorithms are biased against black men. The Guardian newspaper UK. Retrieved 11th December 2016 from https://www.theguardian.com/commentisfree/2016/jun/26/algorithms-racial-bias-offenders-florida?CMP=fb_gu

Neyroud, P. (2015) Evidence-Based Triage in Prosecuting Arrestees Testing an Actuarial System of Selective Targeting. *International Criminal Justice Review*, p.1057567715576173.

Paternoster, R., Brame, R., Bacon, S., Ditchfield, A., Beckman, K. and Frederique, N. (2003). An empirical analysis of Maryland's death sentencing system with respect to the influence of race and legal jurisdiction. *Retrieved June*, *12*, p.2004.

Raynor, P., Kynch, J., Roberts, C. and Merrington, S. (2000) *Risk and need assessment in probation services: an evaluation*. London: Home Office.

Reyes, J., 16 September 2016. 'Philadelphia is grappling with the prospect of a racist computer algorithm'. Retrieved 11th December 2016 from http://technical.ly/philly/2016/09/16/jails-risk-assessment-richard-berk/

Ridgeway, G. (2013) The pitfalls of prediction. *NIJ Journal*, *271*, pp.34-40.

Sherman, L.W. (2007) The power few: experimental criminology and the reduction of harm. *Journal of Experimental Criminology*, *3*(4), pp.299-321.

Sherman, L.W. (2011) Al Capone, the sword of Damocles, and the police–corrections budget ratio. *Criminology & Public Policy*, *10*(1), pp.195-206.

Sherman, L.W. (2012) 'Offender Desistance Policing', in Anti-Social Behaviour and Crime. Hogrefe and Hubar, pp. 199–218.

Sherman, L.W. (2013) The rise of evidence-based policing: Targeting, testing, and tracking. *Crime and justice*, *42*(1), pp.377-451.

Sherman, L.W., Gottfredson, D.C., MacKenzie, D.L., Eck, J., Reuter, P. and Bushway, S.D. (1998) Preventing Crime: What Works, What Doesn't, What's Promising. Research in Brief. National Institute of Justice.

Sherman, L.W., Neyroud, P. and Pease, K. (2012) *Offender-desistance Policing and the Sword of Damocles*. Civitas.

Skeem, J.L. and Lowenkamp, C.T. (2015) Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. *Available at SSRN*.

Slothower, M. (2014) Strengthening police professionalism with decision support: Bounded discretion in out-of-court disposals. *Policing*, *8*(4), pp.353-367.

Starr, S.B. (2014) Evidence-based sentencing and the scientific rationalization of discrimination. *Stan. L. Rev.*, *66*, p.803.

Starr, S.B. (2015) The Risk Assessment Era. *Federal Sentencing Reporter*, *27*(4), pp.205-206.

Tversky, A. and Kahneman, D. (1975) Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making* (pp. 141-162). Springer Netherlands.

13.0 Appendix A: Data Variables

Data Variable (** = used as a predictor in the forecasting model)	**Description**
ForecastConstructionID_PK	Unique Identification in Forecast construction model
NominalID_FK	Unique identification of a suspect given by the model
PncNominal	Unique PNC identification for an individual
CroNominal	Unique CRO identification for an individual
CustodyID_FK	Unique Custody record identifier
CustodyRecord	Unique Custody record number
CustodyDateTime	Date and Time entered custody
CustodyDate	Date of initial custody event
FollowUpYears	Total time passed since first initial custody event
BirthDate	Date of Birth
CustodyAge **	Age at presenting custody event
CustodyJuvenile	Was the suspect a juvenile at time of presenting offence
Gender **	Male or Female
InstantAnyOffenceCount **	Count of any offences at presenting custody event

Data Variable (** = used as a predictor in the forecasting model)	**Description**
InstantMurderOffenceCount	Count of murder or attempted murder offences at presenting custody event
InstantSeriousOffenceCount	Count of serious offences at presenting custody event
InstantViolenceOffenceCount	Count of violent offences at presenting custody event
InstantSexualOffenceCount	Count of sexual offences at presenting custody event
InstantWeaponOffenceCount	Count of weapon offences at presenting custody event
InstantFirearmOffenceCount	Count of firearms offences at presenting custody event
InstantDrugOffenceCount	Count of drug offences at presenting custody event
InstantDrugDistOffenceCount	Count of drug offences at presenting custody event
InstantPropertyOffenceCount	Count of property offences at presenting custody event
InstantViolenceOffenceBinary **	A yes/no binary value is used to define the present offence in terms of violence
InstantPropertyOffenceBinary **	A yes/no binary value is used to define the present offence in terms of property offence.
CustodyPostcodeFull	Suspect's post code in at presenting custody event.
CustodyPostcodeOutward	Suspect's outward postcode at presenting custody event.
CustodyPostcodeOutwardTop24 **	The 25 most common 'outward' (first 3-4 characters) postcodes in County Durham. If the offenders postcode is outside of County Durham
CustodyInForceArea	A binary yes/no variable indicating whether the post code is within the force area

Data Variable (** = used as a predictor in the forecasting model)	**Description**
CustodyMosaicCode	Socio-geo -demographic code for County Durham
CustodyMosaicCodeTop28 **	The 28 most common socio-geo demographic characteristics for County Durham
FirstAnyOffenceDate	The suspect's date of first offence regardless of juvenile or adult
FirstAnyOffenceAge **	The suspect's age at first offender regardless of juvenile or adult
FirstViolenceOffenceDate	The suspect's date of first violent offence regardless of juvenile or adult
FirstViolenceOffenceAge **	The suspect's age at first violent offence regardless of juvenile or adult
FirstSexualOffenceDate	The suspect's date of first sexual offence regardless of juvenile or adult
FirstSexualOffenceAge **	The suspect's age at first sexual offence regardless of juvenile or adult
FirstWeaponOffenceDate	The suspect's date of first weapon offence regardless of juvenile or adult
FirstWeaponOffenceAge **	The suspect's age at first weapon offence regardless of juvenile or adult
FirstDrugOffenceDate	The suspect's date of first drug offence regardless of juvenile or adult
FirstDrugOffenceAge **	The suspect's age at first drug offence regardless of juvenile or adult
FirstPropertyOffenceDate	The suspect's date of first property offence regardless of juvenile or adult
FirstPropertyOffenceAge **	The suspect's age at first property offence regardless of juvenile or adult
PriorAnyOffenceCount **	The number of offences prior to the presenting offence for the suspect

Data Variable (** = used as a predictor in the forecasting model)	**Description**
PriorAnyOffenceLatestDate	The most recent date of any prior offence
PriorAnyOffenceLatestYears **	The number of years since any offence– if there is no offence history, Null value is returned.
PriorMurderOffenceCount **	The number of murder offences prior to the presenting offence for the suspect
PriorMurderOffenceLatestDate	The most recent date of any prior murder offence
PriorMurderOffenceLatestYears	The number of years since the most recent custody instance in which a murder offence was committed – if there is no murder offence history then a code of 100 years is used.
PriorSeriousOffenceCount **	The number of serious offences prior to the presenting offence for the suspect
PriorSeriousOffenceLatestDate	The most recent date of any prior serious offence
PriorSeriousOffenceLatestYears **	The number of years since the most recent custody instance in which a serious offence was committed – if there is no serious offence history then a code of 100 years is used.
PriorViolenceOffenceCount **	The number of violence offences prior to the presenting offence for the suspect
PriorViolenceOffenceLatestDate	The most recent date of any prior violent offence
PriorViolenceOffenceLatestYears **	The number of years since the most recent custody instance in which a violence offence was committed – if there is no violence offence history then a code of 100 years is used.
PriorSexualOffenceCount **	The number of sexual offences prior to the presenting offence for the suspect
PriorSexualOffenceLatestDate	The most recent date of any prior sexual offence
PriorSexualOffenceLatestYears **	The number of years since the most recent custody instance in which a sexual offence was committed – if there is no sexual offence history then a code of 100 years is used.

Data Variable (** = used as a predictor in the forecasting model)	**Description**
PriorSexRegOffenceCount **	The number of sex offender register offences prior to the presenting offence for the suspect
PriorSexRegOffenceLatestDate	The most recent date of any prior sex offender register offence
PriorSexRegOffenceLatestYears	The number of years since the most recent custody instance in which a sex offender register offence was committed – if there is no sex offender history then a code of 100 years is used.
PriorWeaponOffenceCount **	The number of weapon offences prior to the presenting offence for the suspect
PriorWeaponOffenceLatestDate	The most recent date of any prior weapon offence
PriorWeaponOffenceLatestYears **	The number of years since the most recent custody instance in which a weapon offence was committed – if there is no weapon offence history then a code of 100 years is used.
PriorFirearmOffenceCount **	The number of firearms offences prior to the presenting offence for the suspect
PriorFirearmOffenceLatestDate	The most recent date of any prior firearms offence
PriorFirearmOffenceLatestYears	The number of years since the most recent custody instance in which a firearms offence was committed – if there is no firearms offence history then a code of 100 years is used.
PriorDrugOffenceCount **	The number of drug offences prior to the presenting offence for the offender
PriorDrugOffenceLatestDate	The most recent date of any prior drug offence
PriorDrugOffenceLatestYears **	The number of years since the most recent custody instance in which a drugs offence was committed – if there is no drugs offence history then a code of 100 years is used.
PriorDrugDistOffenceCount **	The number of drug distribution offences prior to the presenting offence for the offender
PriorDrugDistOffenceLatestDate	The most recent date of any prior drug distribution offence

Data Variable (** = used as a predictor in the forecasting model)	**Description**
PriorDrugDistOffenceLatestYears	The number of years since the most recent custody instance in which a drug distribution offence was committed – if there is no drugs distribution offence history then a code of 100 years is used.
PriorPropertyOffenceCount **	The number of property offences prior to the presenting offence for the offender
PriorPropertyOffenceLatestDate	The most recent date of any prior property offence
PriorPropertyOffenceLatestYears **	The number of years since the most recent custody instance in which a property offence was committed – if there is no property offence history then a code of 100 years is used.
PriorCustodyCount **	The number of custody events prior to the presenting offence for the offender
PriorCustodyLatestDate	The most recent date of any prior custody event
PriorCustodyLatestYears **	The number of years since the most recent custody instance – if there is no custody event history then a code of 100 years is used.
PriorIntelCount **	The number of intelligence submissions at nominal level. The offender at nominal level will have a unique identifier, the submissions are counted within the forecasting model
PostAnyOffenceCount	Any offence count since forecast
PostAnyOffenceEarliestDate	The date of first offence following the presenting custody event
PostAnyOffenceEarliestDays	The number of days between the presenting custody event and first offence
PostWithin01AnyOffenceCount	Any offence count in 12 months since forecast
PostWithin02AnyOffenceCount	Any offence count in 24 months since forecast
PostSeriousOffenceCount	Serious offence count since forecast

Data Variable (** = used as a predictor in the forecasting model)	**Description**
PostSeriousOffenceEarliestDate	The date of first serious offence following the presenting custody event
PostSeriousOffenceEarliestDays	The number of days between the presenting custody event and first serious offence
PostWithin01SeriousOffenceCount	Serious offence count in 12 months since forecast
PostWithin02SeriousOffenceCount	Serious offence count in 24 months since forecast
PostWithin02ActualRiskGroup	Suspect's actual risk group within 24 months of forecast
PostWithin02ForecastRiskGroup	Suspect's forecast risk group within 24 months of forecast
PostWithin02ForecastVotesHigh	Number of votes for high risk forecast ???
PostWithin02ForecastVotesModerate	Number of votes for moderate risk forecast ???
PostWithin02ForecastVotesLow	Number of votes for low risk forecast

14.0 Appendix B: Predictor variables only

Predictor Variables Only	Description
CustodyAge	Age at presenting custody event
Gender	Male or Female
InstantAnyOffenceCount	Count of any offences at presenting custody event
InstantViolenceOffenceBinary	A yes/no binary value is used to define the present offence in terms of violence
InstantPropertyOffenceBinary	A yes/no binary value is used to define the present offence in terms of property offence.
CustodyPostcodeOutwardTop24	The 25 most common 'outward' (first 3-4 characters) postcodes in County Durham. If the offenders postcode is outside of County
CustodyMosaicCodeTop28	The 28 most common socio-geo demographic characteristics for County Durham
FirstAnyOffenceAge	The suspect's age at first offender regardless of juvenile or adult
FirstViolenceOffenceAge	The suspect's age at first violent offence regardless of juvenile or adult
FirstSexualOffenceAge	The suspect's age at first sexual offence regardless of juvenile or adult
FirstWeaponOffenceAge	The suspect's age at first weapon offence regardless of juvenile or adult
FirstDrugOffenceAge	The suspect's age at first drug offence regardless of juvenile or adult
FirstPropertyOffenceAge	The suspect's age at first property offence regardless of juvenile or adult
PriorAnyOffenceCount	The number of offences prior to the presenting offence for the suspect

Predictor Variables Only	Description
PriorAnyOffenceLatestYears	The number of years since any offence– if there is no offence history, Null value is returned.
PriorMurderOffenceCount	The number of murder offences prior to the presenting offence for the suspect
PriorSeriousOffenceCount	The number of serious offences prior to the presenting offence for the suspect
PriorSeriousOffenceLatestYears	The number of years since the most recent custody instance in which a serious offence was committed – if there is no serious offence history then a code of 100 years is used.
PriorViolenceOffenceCount	The number of violence offences prior to the presenting offence for the suspect
PriorViolenceOffenceLatestYears	The number of years since the most recent custody instance in which a violence offence was committed – if there is no violence offence history then a code of 100 years is used.
PriorSexualOffenceCount	The number of sexual offences prior to the presenting offence for the suspect
PriorSexualOffenceLatestYears	The number of years since the most recent custody instance in which a sexual offence was committed – if there is no sexual offence history then a code of 100 years is used.
PriorSexRegOffenceCount	The number of sex offender register offences prior to the presenting offence for the suspect
PriorWeaponOffenceCount	The number of weapon offences prior to the presenting offence for the suspect
PriorWeaponOffenceLatestYears	The number of years since the most recent custody instance in which a weapon offence was committed – if there is no weapon offence history then a code of 100 years is used.
PriorFirearmOffenceCount	The number of firearms offences prior to the presenting offence for the suspect
PriorDrugOffenceCount	The number of drug offences prior to the presenting offence for the offender

Predictor Variables Only	Description
PriorDrugOffenceLatestYears	The number of years since the most recent custody instance in which a drugs offence was committed – if there is no drugs offence history then a code of 100 years is used.
PriorDrugDistOffenceCount	The number of drug distribution offences prior to the presenting offence for the offender
PriorPropertyOffenceCount	The number of property offences prior to the presenting offence for the offender
PriorPropertyOffenceLatestYears	The number of years since the most recent custody instance in which a property offence was committed – if there is no property offence history then a code of 100 years is used.
PriorCustodyCount	The number of custody events prior to the presenting offence for the offender
PriorCustodyLatestYears	The number of years since the most recent custody instance – if there is no custody event history then a code of 100 years is used.
PriorIntelCount	The number of intelligence submissions at nominal level. The offender at nominal level will have a unique identifier, the submissions are counted within the forecasting model

15.0 Appendix C: Case Study Narrative

Case Study 1

Forecast low risk; Actually high risk (False Negative, random selection)

What do we know?

Male aged 32, ex-Military, lives outside of the force area

Warning markers of Mental Health (PTSD), Asthma and Suicidal.

All records are generally domestic abuse related – either against mother or different partners.

Criminal History

Up to the end of 2012 there are 6 entries on PND from two force areas regarding domestic abuse/drugs this information would not be known to the model.

2008: Criminal damage – domestic abuse related for which he received a caution, this offence occurred outside of Durham Constabulary force area.

2012: Driving with excess alcohol for which he was convicted, this offence occurred outside of Durham Constabulary force area.

Presenting Offences

2013: Assault (common) – domestic abuse related – out of court disposal

Follow up Period

2014: Assault (Battery) domestic abuse related for which no further action taken

False negative outcome 2014 Rape of Female over 16 – domestic abuse related for which no further action taken

2015: Assault (Battery) – domestic abuse related, convicted, found guilty

How confident was the model and what information did the model have?

High18; Moderate 37; Low 454

The model had only the offence for which the suspect was being forecast and no other offending history or intelligence on which to base its forecast.

Conclusions

Essentially the suspect did not appear in the data from which the model draws its predictors to the model and therefore the model predicted low risk with high confidence. Further information was available to police, via PNC and PND which consisted largely of domestic abuse related intelligence and two non-serious convictions. Arguably, had the model had this information, it may or may not have taken the offender up to moderate risk. I would not, however, have expected the forecast to become high risk. The suspect

in this case study was arrested 87 days after the forecast for the offence of rape in 2015 and it is this serious offence that brought about the false negative outcome.

Case Study 2

Forecast low risk; Actually high risk (False Negative, Hand picked)

What do we know?

Male, aged 34, lives inside of the force area

Warning markers Mental Health and Self Harm

Criminal History

1998: Handling stolen goods for which he was convicted

1999 – 2005: throughout the period the suspect was arrested for a variety of offences including assault (ABH), driving without due care and attention, driving with no insurance, assault (common), burglary, theft from a motor vehicle, and finally drunk and disorderly.

Presenting Offence

2013: Driving with excess alcohol for which no further action taken

Follow up period

False negative outcome 2015: Arson with intent to endanger life, Wounding with intent to do GBH. The offender was charged and subsequently found not guilty.

How confident was the model and what information did the model have?

High 115; Moderate 196; low 198

Age of onset (general): 19 years

… Violence: 19 years

… Property: 19 years

Prior offences (general): 5

… Violence: 2

… Property: 2

Prior custody events: 4

Prior intelligence reports: 6

Conclusions

The forecast would not appear to be a confident forecast, clearly the votes indicate the suspect would not be forecast as high risk, however just two votes determine the outcome of low risk. At the point at which the forecast was made, the data from which the model draws its predictors showed the suspect, had not been arrested for just under a decade. Although the prior offending was across a range of offences, none

would be classified as serious offences individually. The suspect, 148 days after the forecast, was arrested for a serious offence.

Case Study 3

Forecast low risk; Actually high risk (False Negative, random selection)

What do we know?

Male, aged 17

Frequent Missing Person

Criminal History

No previous arrests

Presenting Offence

2013: first came to notice of police for an offence of theft for which no further action was taken.

Follow up period

2013 -2014: Further arrests continue covering offences of theft, burglary, criminal damage – all offences were dealt with by taking no further action or an out of court disposal.

False negative outcome 2014: Sexual touching and sexual activity (familial) with a female child for which no further action was subsequently taken.

2014: Sexual touching two offences for which no further action was taken

2015: Criminal damage for which no further action was taken

How confident was the model and what information did the model have?

High 114; Moderate 78; Low 317

The model did not have any previous offending behaviour history or intelligence count.

Conclusions

The model had 317 votes of low risk out of 509 potential votes, therefore the model was confident in the forecast of low risk. Essentially the model had no offending history as this suspect commenced his offending in 2013 when the forecast was made, the forecast was therefore made on other predictor variables available in the model. Within 75 days of forecast, the suspect was arrested and 229 days after the forecast the suspect was arrested for a serious offence. The sexual offences were recent offences as opposed to historical sexual cases.

Case Study 4

Forecast High Risk; Actually low risk (False positive, random selection)

What do we know?

Male aged 21

Criminal History

2008: First arrested for an offence of assault (ABH)

2008 – 2012: arrested for a variety of offences; assault (ABH), criminal damage, burglary, affray and minor public order offences.

2013: Drunk and disorderly for which he received a fixed penalty notice – in a neighbouring force area

2013: Affray for which he was convicted – in a neighbouring force area

2013: Incite female under 16 to engage in sexual act (penetration) for which no further action was taken

Presenting Offence

2013: Burglary for which no further action was taken

Follow up

2013: Not arrested but attended the police station as a voluntary attender for an offence of criminal damage for which he received a caution.

2013: Drunk and disorderly for which he received a fixed penalty notice – in a neighbouring force area

How confident was the model and what information did the model have?

High 308 Moderate 165 Low 36

Age of onset (general): 13 years

… Violence: 13 years

… Property: 13 years

… Sexual: 17 years

Prior offences (general): 8

… Violence: 5

… Sexual: 1

… Property: 2

Prior offences (general): 8

… Serious: 1

… Violence: 5

… Sexual: 1

… Property: 2

Prior custody events: 8

Prior intelligence reports: 2

Conclusions

The model was confident in its forecast of high risk. This cases study highlights that the model was blind to some of the suspects offending due to the limitations of the forecasting model, that said the model still forecast high therefore in this case study the other information may not have changed the risk forecast. Data outside of the geographic force area is not used, if PNC were used as part of the model the forecasting model would have been aware of the drunk and disorderly offence in a neighbouring force area. The case study also highlights that the outcome offence, which was dealt with by voluntary attender as opposed to being arrested, was not in the data the model draws from. The voluntary attendee data would have made the suspect actually moderate. If PNC were used in the building of such a model, then information of the voluntary attender offence would have been available. The offender therefore did not turn out to be low risk but was in fact moderate risk. In 2016, outside of the validation period an arrest for affray has occurred, albeit no further action was taken for the offence.

Case Study 5

Forecast high risk; Actually low risk (False positive, random selection)

What do we know?

Male aged 24

Warning Markers; Violent, ADHD, medication taken.

This male is a violent controlling individual with intelligence suggesting a background of serious domestic abuse with previous partners; drugs; and weapons. A number of domestic abuse incidents are also recorded on police records between 2012 and 2014.

Criminal History

2005: First came to police attention at the age of 13 years for an offence of assault (common), and later the same year for offence of criminal damage both of which resulted in warnings/reprimand.

2005 – 2008: Break in offending

2009 – 2012: Arrested for offences that escalate quite rapidly, with offences of handling stolen goods, affray, assault (common), theft, assault (ABH), threats to kill, and assault (GBH Wounding) all of which no further action was taken against him.

2013: Battery for which no further action was taken

2013: Breach of non-molestation order for which no further action was taken

Presenting Offence

2013: Breach of non-molestation order for which no further action was taken

Follow up period

No intelligence or arrests since.

How confident was the model and what information did the model have?

High 264; Moderate 213; Low 32

Age of onset (general): 17 years

… Violence: 17 years

… Property: 17 years

Prior offences (general): 14

… Murder: 1

… Serious: 3

… Violence: 9

… Property: 4

Prior custody events: 9

Prior intelligence reports: 35

Conclusions

The model forecast with somewhat high confidence that the offender was not a low risk offender. He was arrested on suspicion of attempted murder for one of the previous Assault (GBH) charges therefore the model did have a count of 1 recorded against that predictor. A number of intelligence records and incidents suggest domestic abuse may still be ongoing.

Case Study 6

Forecast high risk; Actually low risk (False positive, hand picked)

What do we know?

Male aged 22

Criminal History

2006: First came to police attention for an offence of theft for which no further action was taken

2007 – 2012: arrested for a variety of different offences including, theft, criminal damage, arson, burglary, and racially aggravated criminal damage.

2013 – Public order offence for which he was given a fixed penalty notice

Presenting Offence

2013: Public order offence for which he was charged and convicted at court

Follow up Period

No further arrests since the presenting offence forecast

How confident was the model and what information did the model have?

High 248; Moderate 242; Low 19

Age of onset (general): 15 years

… Property: 15 years

… Violence: 17 years

Prior offences (general): 15

… Violent: 2

…Property: 14

Prior custody events: 9

Prior intelligence reports: 20

Conclusions

The forecast was confident that the offender was not low risk, however only 6 votes separated the high risk forecast from a moderate forecast. In light of no further arrests since the forecast, having been forecast high risk this effectively gives us the high risk false positive outcome.

Case Study 7

Forecast High Risk; Actually High Risk (True Positive, Random selection)

What do we know?

Male aged 27

PND Intelligence across four force areas indicating he is committing theft and burglary offences.

Criminal History

2006: first comes to police attention for theft, for which no further action was taken

2007-2013: every year since then up to 2013 he has been arrested at least once for a variety of offences including, burglary, theft of vehicle, assault, sexual activity with a child under 16, resist/obstruct police constable, no insurance, driving whilst disqualified, and aggravated vehicle taking.

Presenting offence

2013: assisting an offender by impeding his apprehension/prosecution, for which no further action was taken.

Follow up Period

2013: Pursing a course of conduct amounting to harassment for which he received a caution

2013: Going equipped for theft for which no further action was taken

2013: Burglary for which no further action was taken

2013: Causing death by careless driving for which he no further action was taken

2013: Theft for which no further action was taken

2014: Burglary for which no further action was taken

2014: Burglary for which no further action was taken

Actual high risk 2014: Robbery for which no further action was taken

2015: Theft for which following charge the case was dismissed

2015: Handling stolen goods for which he was convicted

How confident was the model and what information did the model have?

High 228; Moderate 217; Low 64

Age of onset (general): 18 years

... Property: 18 years

... Violence: 21 years

... Sexual: 21 years

Prior offences (general): 18

... Serious: 1

... Violence: 2

... Sexual: 1

... Property: 6

Prior custody events: 11

Prior intelligence reports: 125

Conclusions

The forecast was confident that the suspect was not low risk, however between moderate and high the votes were very close. The model did, however, forecast high risk and within 59 days the forecast was proved correct. This individual had a much higher intelligence count.

Case Study 8

Forecast High Risk; Actually High Risk (True Positive Random selection)

What do we know?

Male aged, 35

Warning markers; Violent, Mental Health, Ailment, Suicidal, Self-Harm, Drugs,

Criminal History

1995: First comes to police attention for offence of theft of motor vehicle for which he was convicted

2005 - 2010: There was a five-year break in his offending between 2005 and 2010

2010 - 2012: Arrested for a variety of offences including theft, resist police, assault police, drug offences, handling stolen goods, driving whilst disqualified, public order offences, assault (common), burglary, and criminal damage totalling 42 arrests.

Presenting Offence

2013: aggravated vehicle taking, no insurance and excess alcohol offences for which he was charged however at court the case was dismissed/withdrawn.

Follow up Period

Actual high risk 2014: Robbery for which no further action was taken

2015 arrested for assault (common), resist police, criminal damage for which he was convicted.

How confident was the model and what information did the model have?

High 279; Moderate 217; Low 13

Age of onset (general): 16 years

… Property: 16 years

… Violence: 17 years

Prior offences (general): 54

… Violence: 8

… Property: 31

Prior custody events: 37

Prior intelligence reports: 37

Conclusions

The forecast was very confident that the suspect was not low and there were 62 votes separating the high from the moderate risk group. The offender since the follow up period subject of this research has been arrested for further offences. The model forecast high risk and within 152 days the forecast was proved correct.

Case Study 9

Forecast High risk; Actually high risk (True Positive – Handpicked)

What do we know?

Male aged 24

Warning Markers, [redacted]

Criminal History

2007: First comes to police attention for [redacted]

2008 – 2013: [redacted]

Presenting Offence

2013: [redacted]

Follow up Period

2014: [redacted]

2014: [redacted]

Actual high risk 2014: Murder for which he was convicted.

How confident was the model and what information did the model have?

High 414; Moderate; 87 Low 8

Age of onset (general): [redacted] years

… Violence: [redacted] years

… Property: [redacted] years

Prior offences (general): [redacted]

… Violence: [redacted]

… Property: [redacted]

Prior custody events: [redacted]

Prior intelligence reports: [redacted]

Conclusions

The forecast was very confident in the forecast with 414 votes forecasting high and only 8 votes forecasting low. The confirmation of the forecast which provides the high risk true positive outcome occurred 545 days after the date of the forecast.

16.0 Appendix D: Forecast Group Characteristics

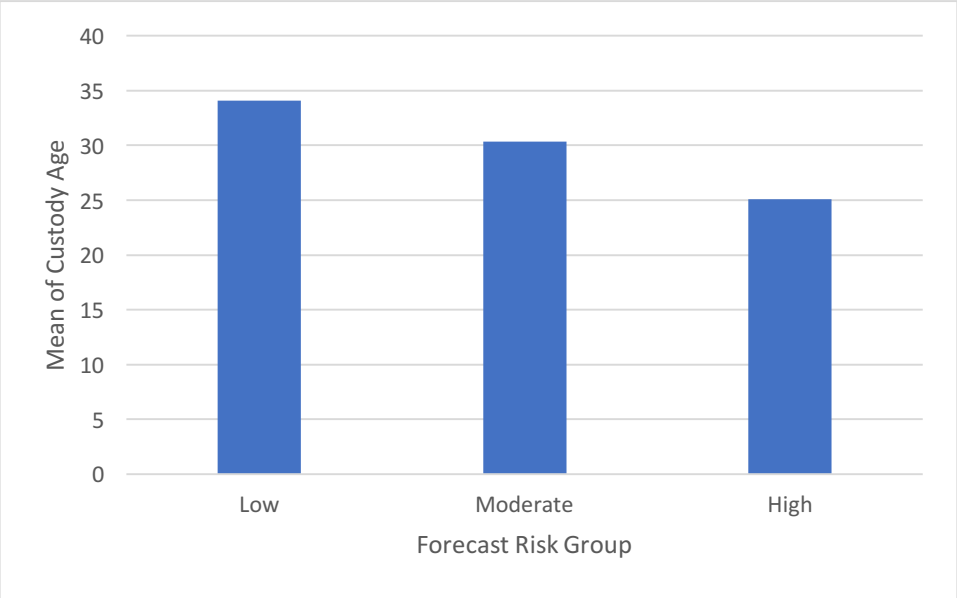

Figure 19:Mean Custody age at the time of the presenting offence in custody

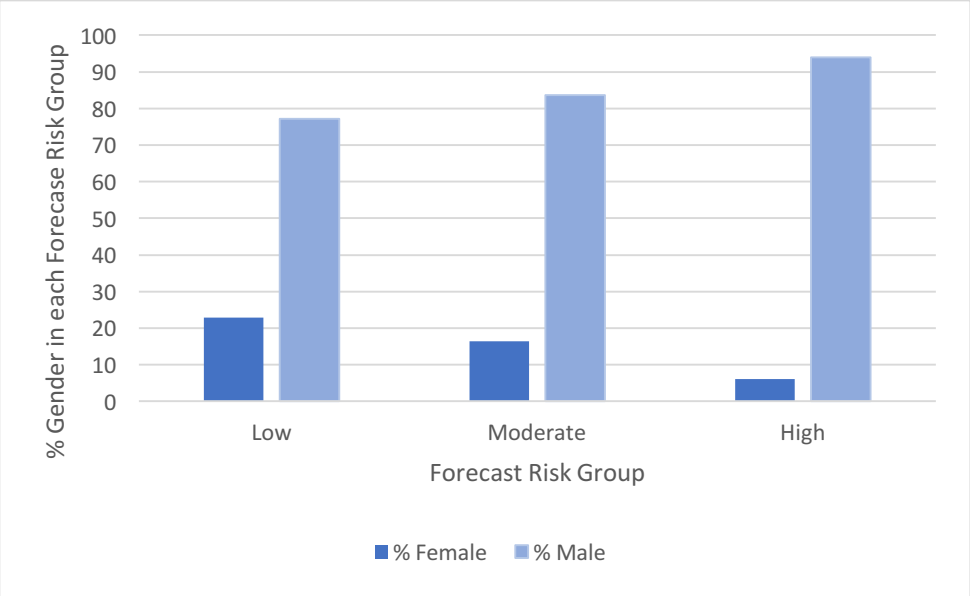

Figure 20: Percentage of gender within forecast risk groups for 2013 validation dataset

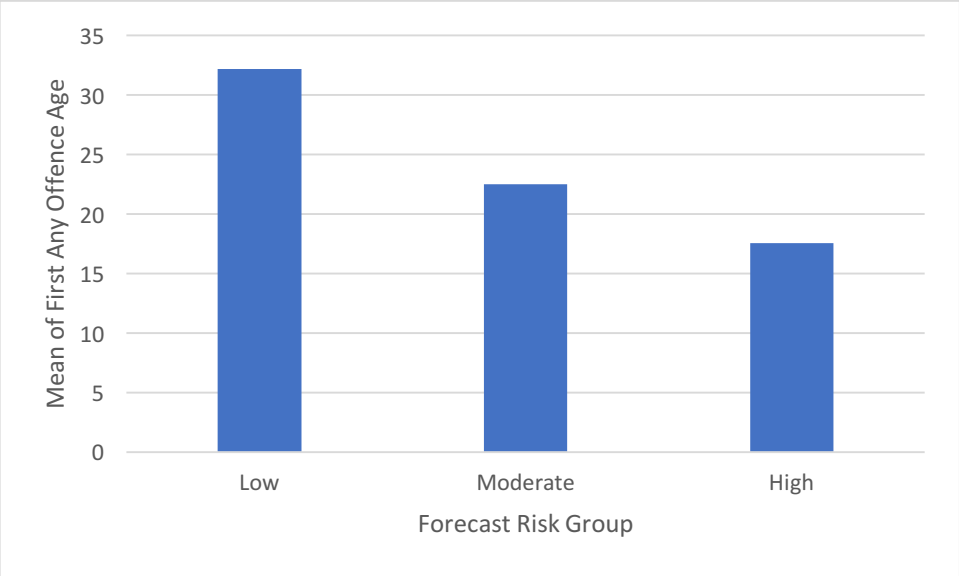

Figure 21: Mean offenders age at first offence

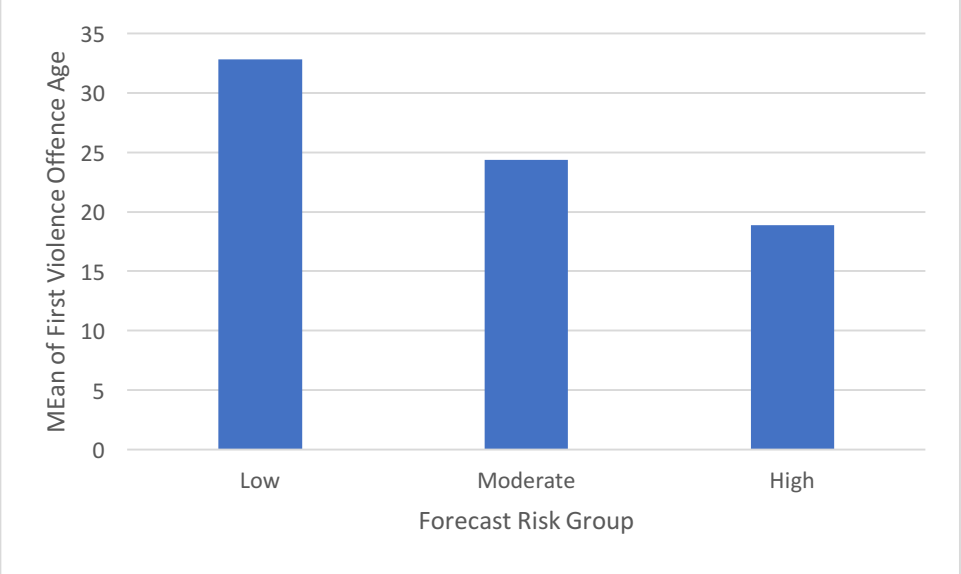

Figure 22: Mean offenders age at the first violent offence

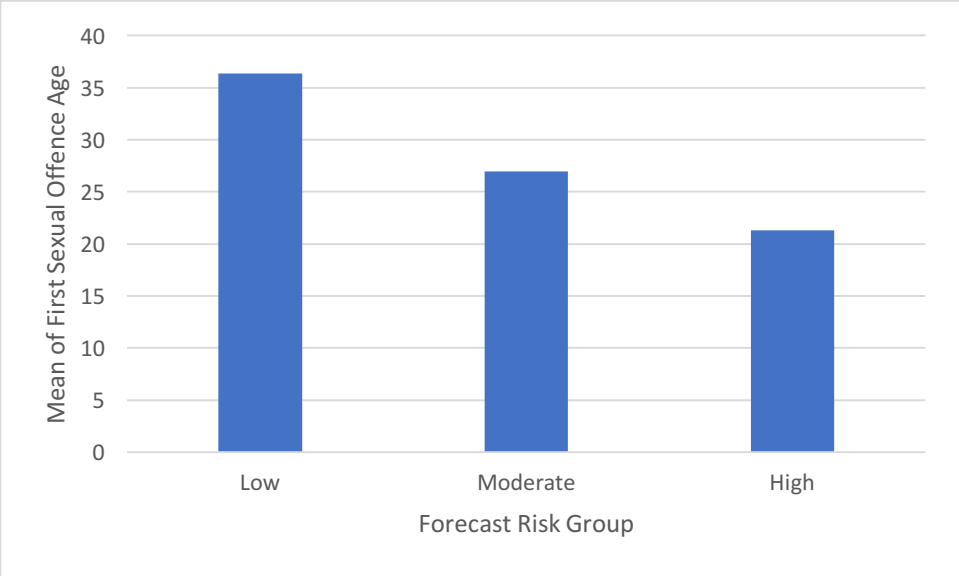

Figure 23: Mean offenders age at first sexual offence

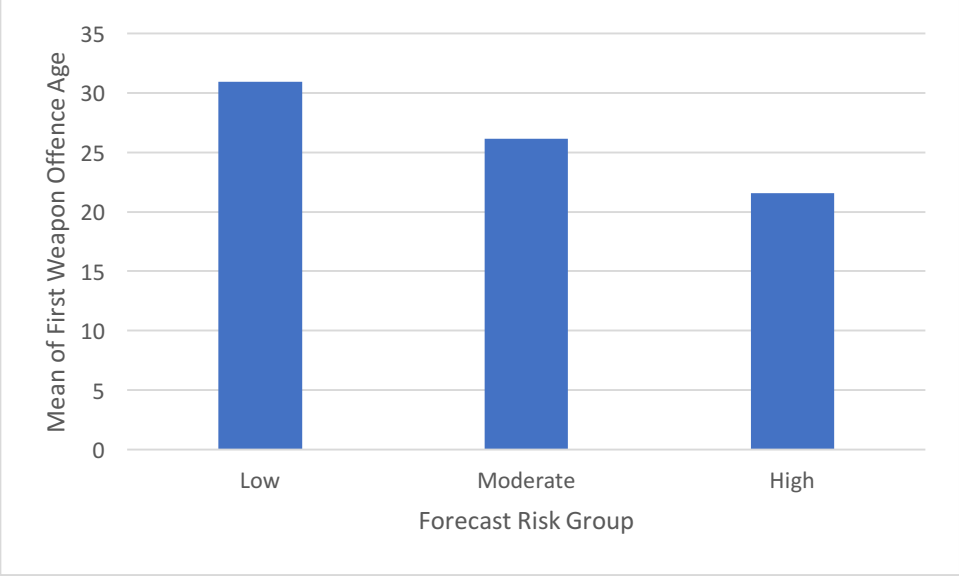

Figure 24: Mean offenders age at first weapon offence

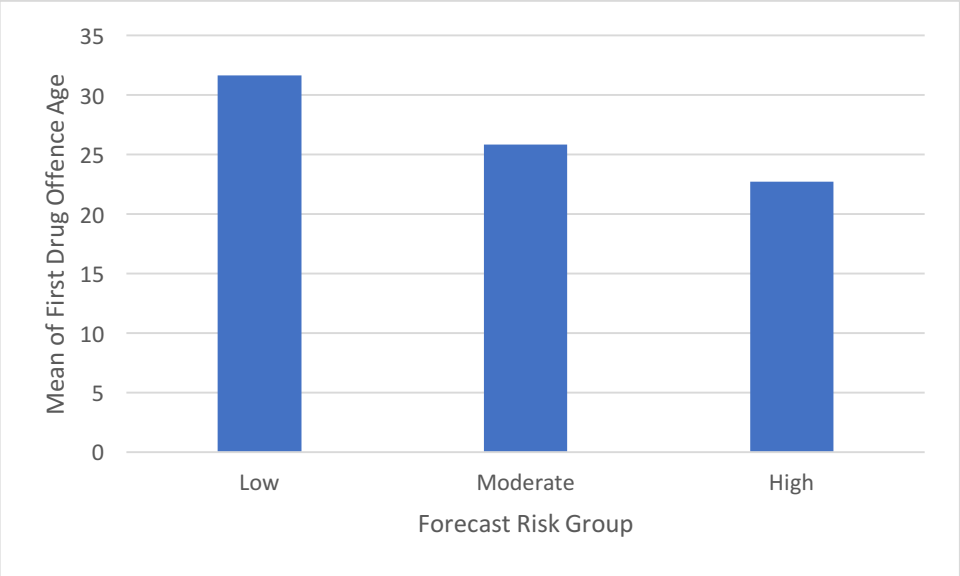

Figure 25: Mean offenders age at first drug offence

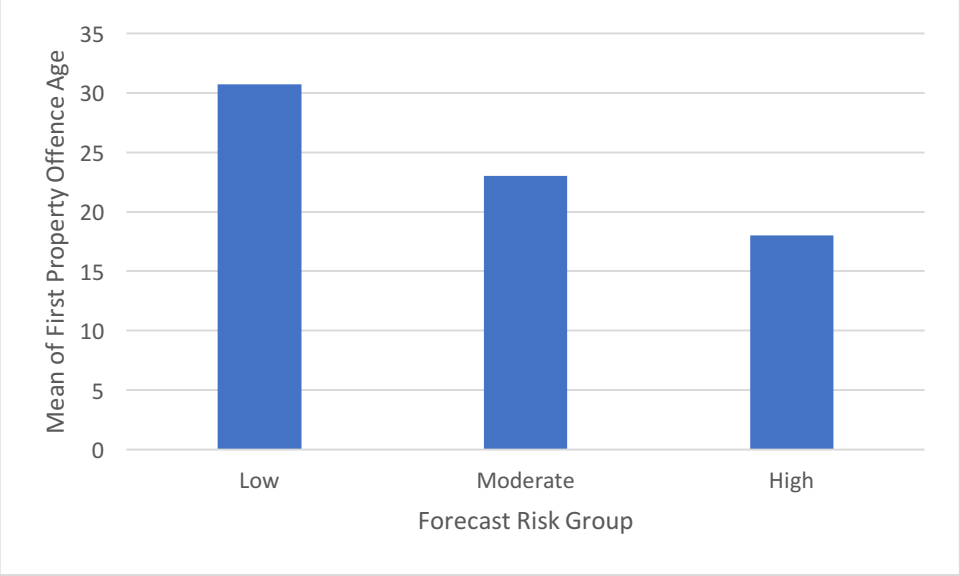

Figure 26: Mean offenders age at first property offence

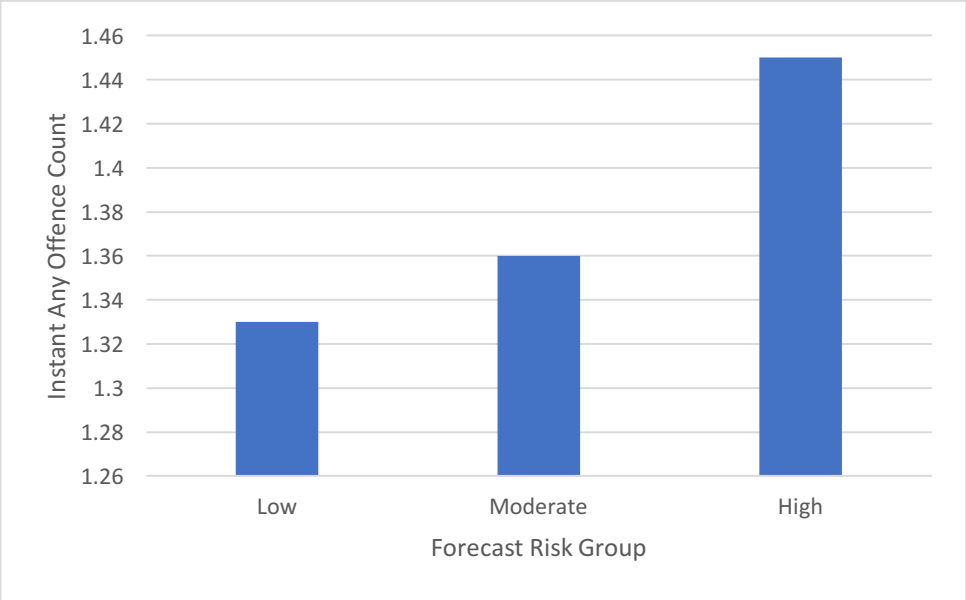

Figure 27: Mean number of presenting offences

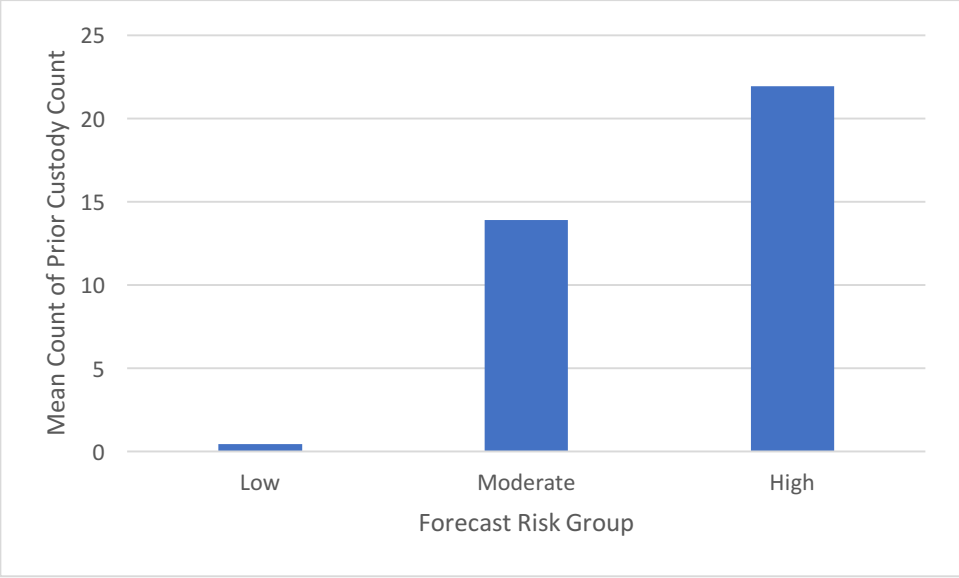

Figure 28: Mean count of custody events prior to the presenting offence

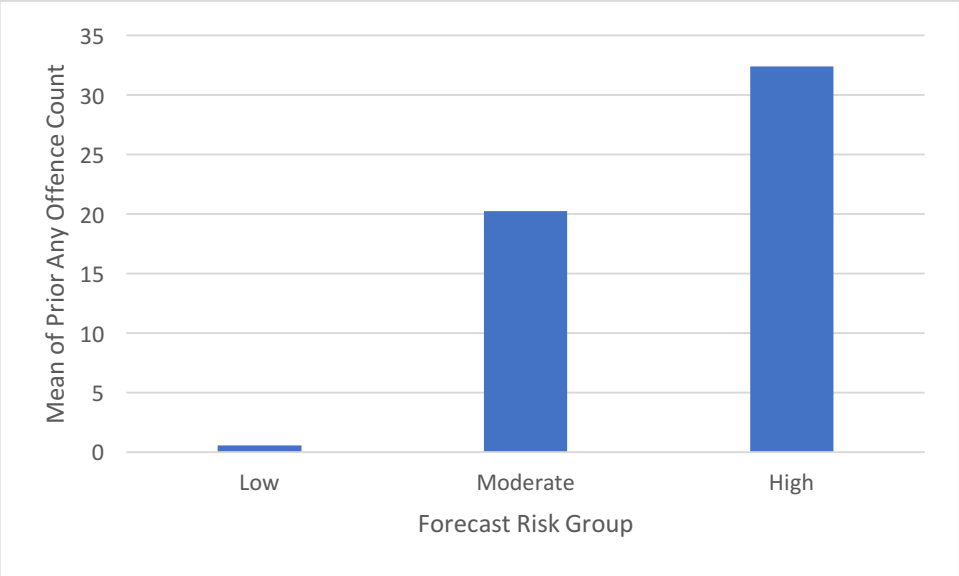

Figure 29: Mean count of offences prior to the presenting offence

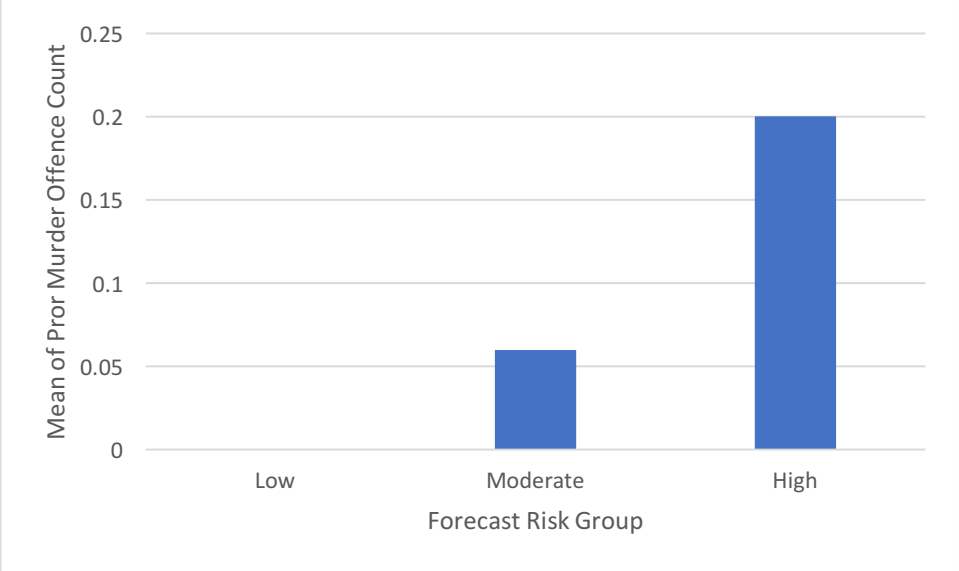

Figure 30: Mean count of murder offences prior to the presenting offence

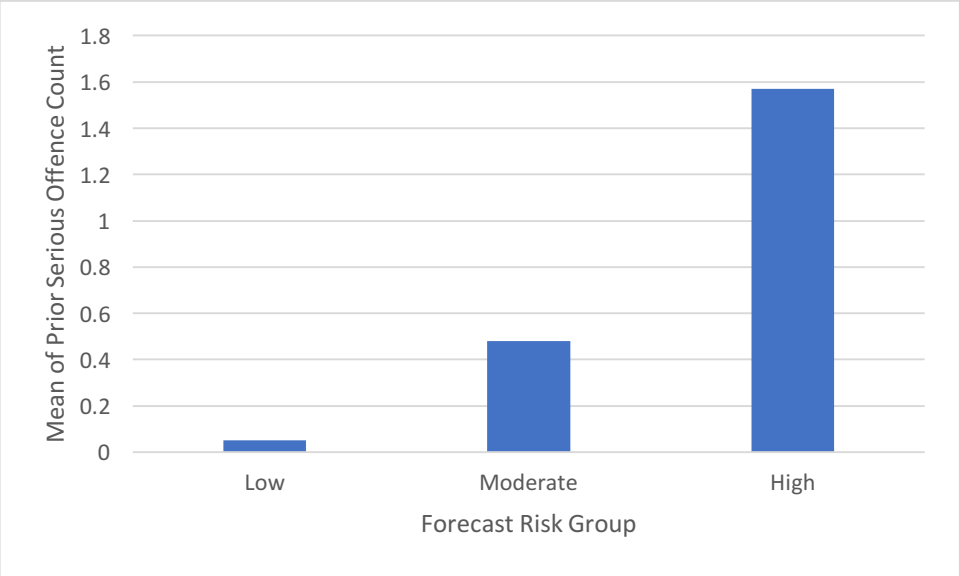

Figure 31: Mean count of serious offences prior to the presenting offence

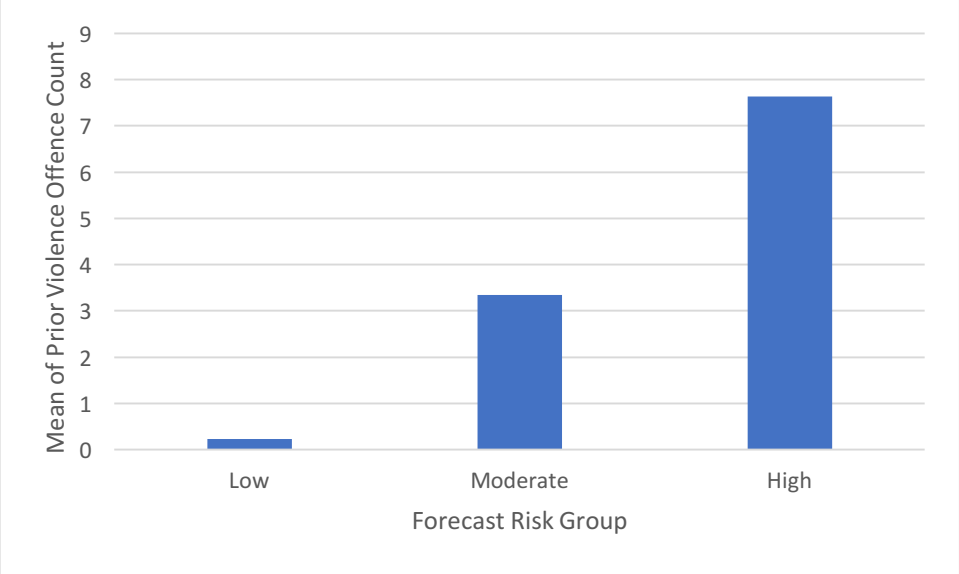

Figure 32: Mean count of violent offences prior to the presenting offence

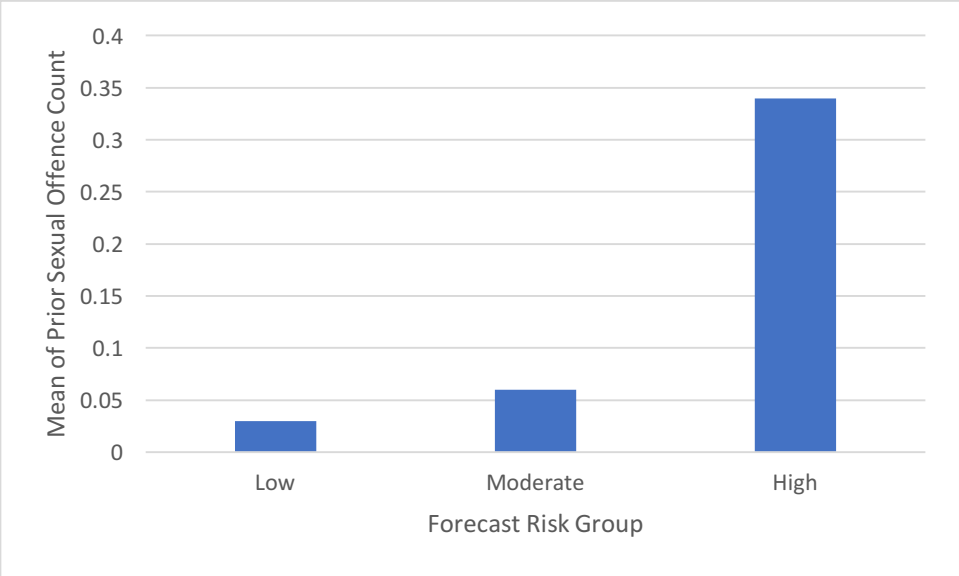

Figure 33: Mean count of sexual offences prior to the presenting offence

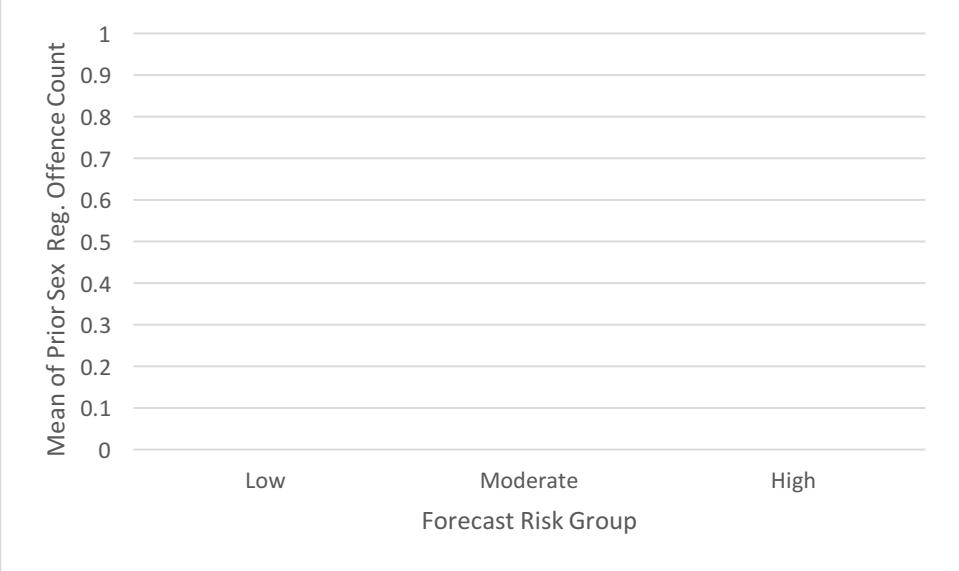

Figure 34: Mean count of sexual registration offences prior to the presenting offence

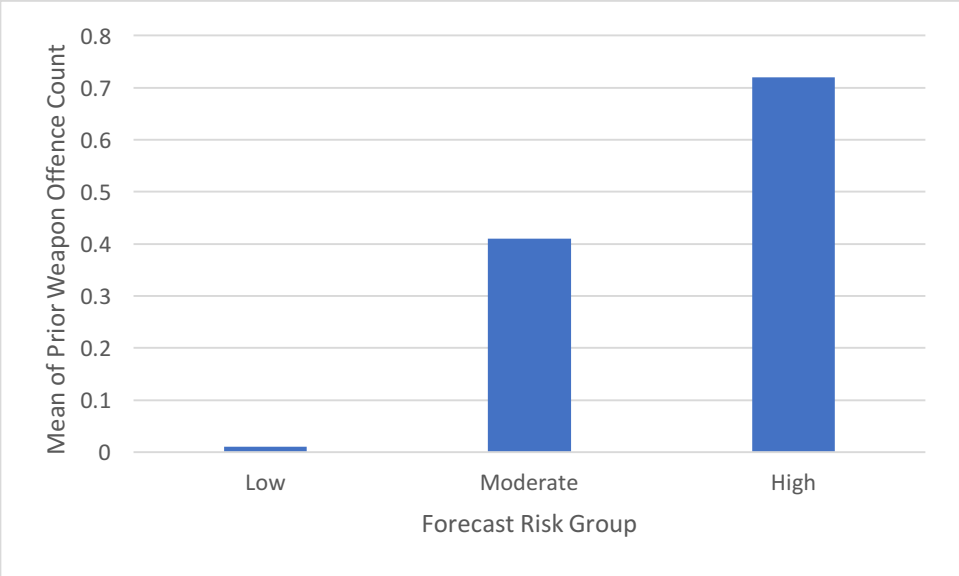

Figure 35: Mean count of weapon offences prior to the presenting offence

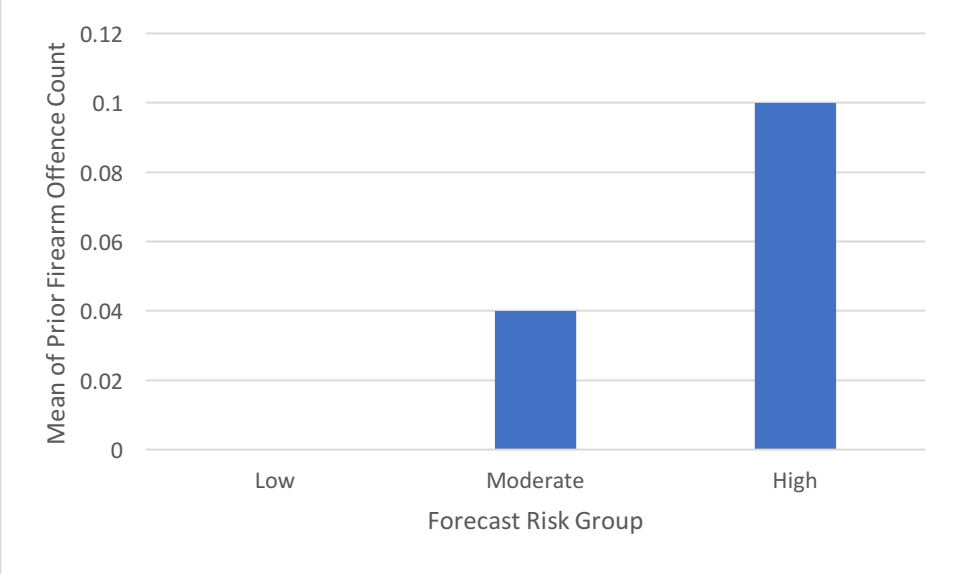

Figure 36: Mean count of firearm offences prior to the presenting offence

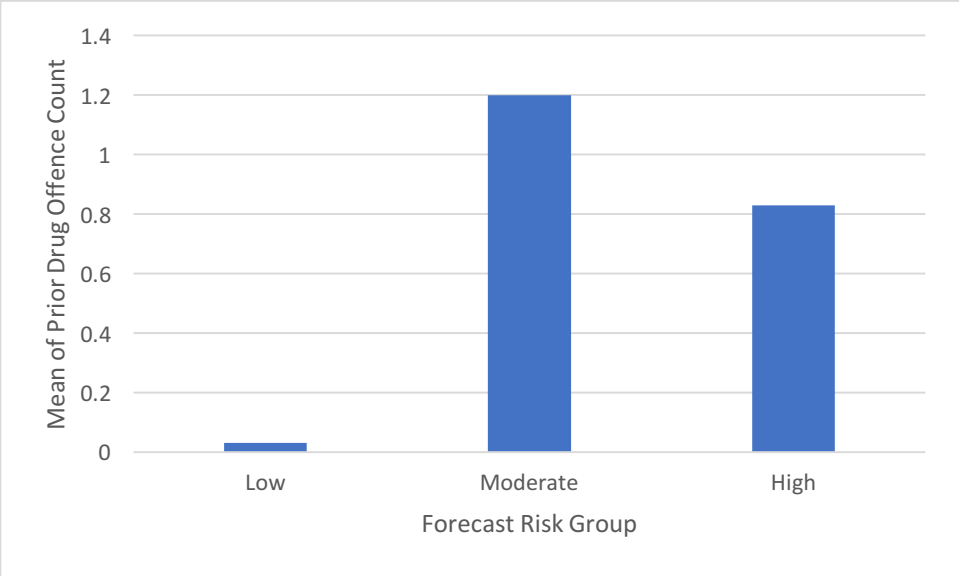

Figure 37: Mean count of drug offences prior to the presenting offence

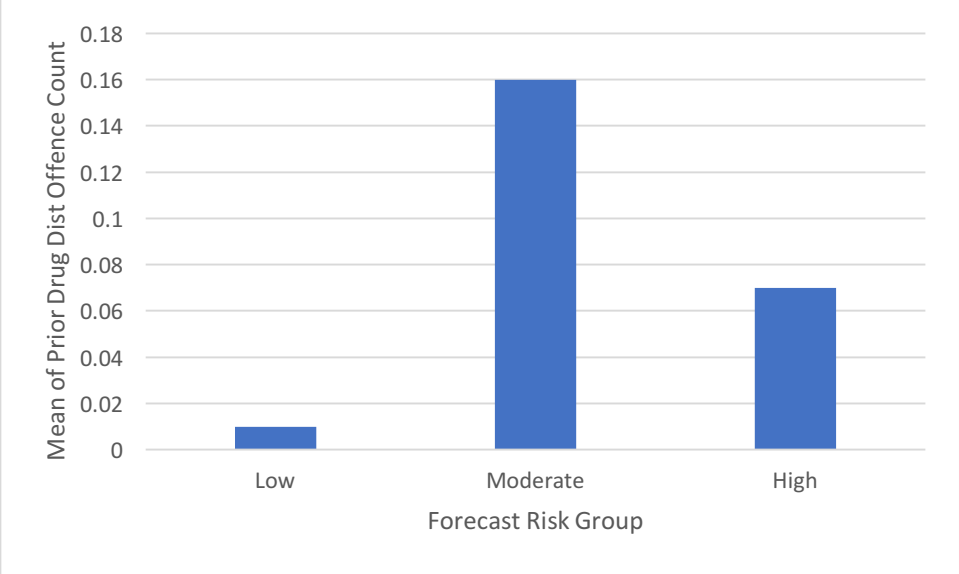

Figure 38: Mean count of drug distribution offences prior to the presenting offence

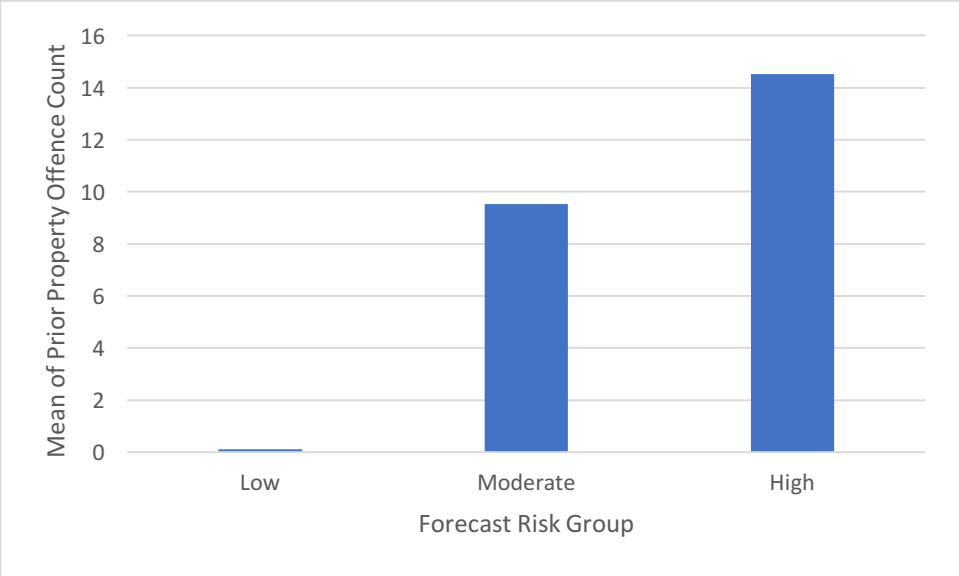

Figure 39: Mean count of property offences prior to the presenting offence

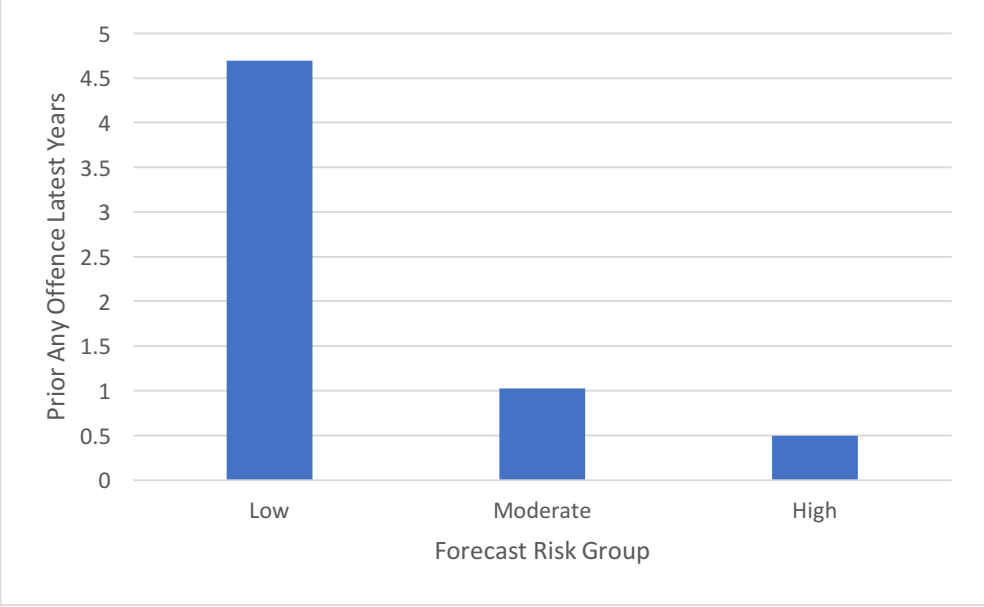

Figure 40: The mean number of years since the most recent custody instance for any offence

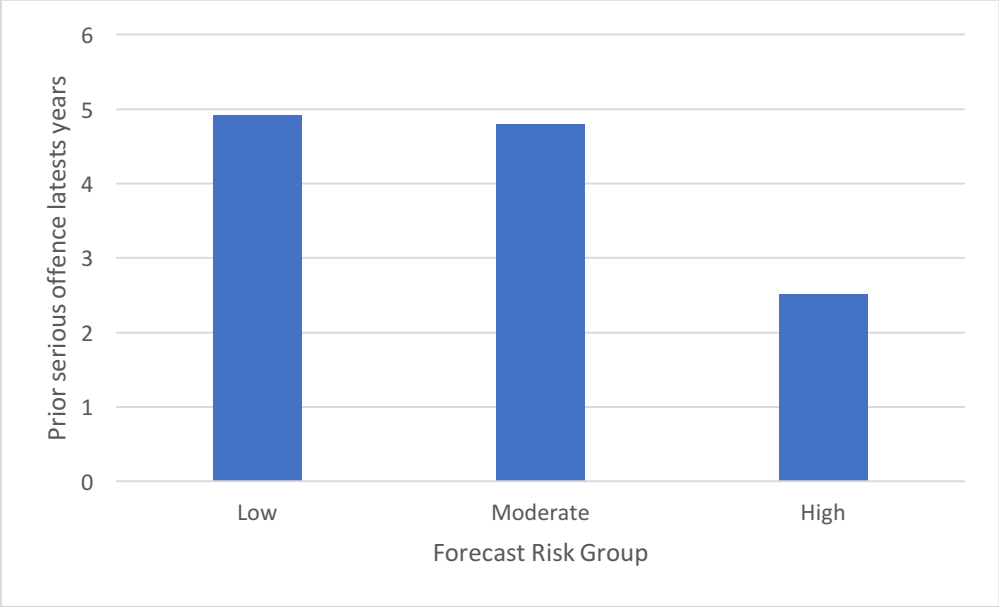

Figure 41: The mean number of years since most recent custody event for serious offences

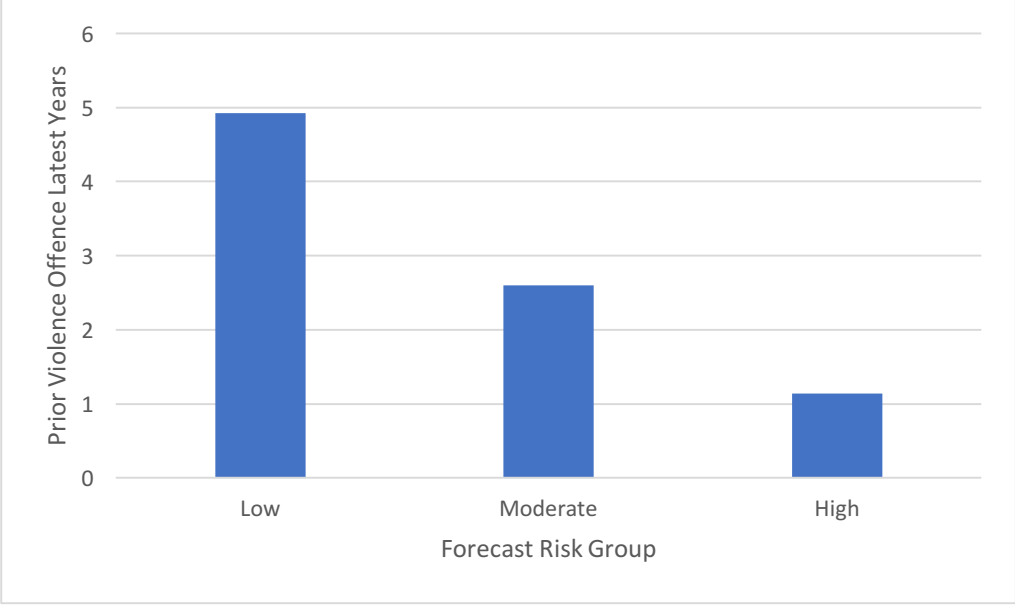

Figure 42: Mean number of years since the most recent custody instance for violent offences

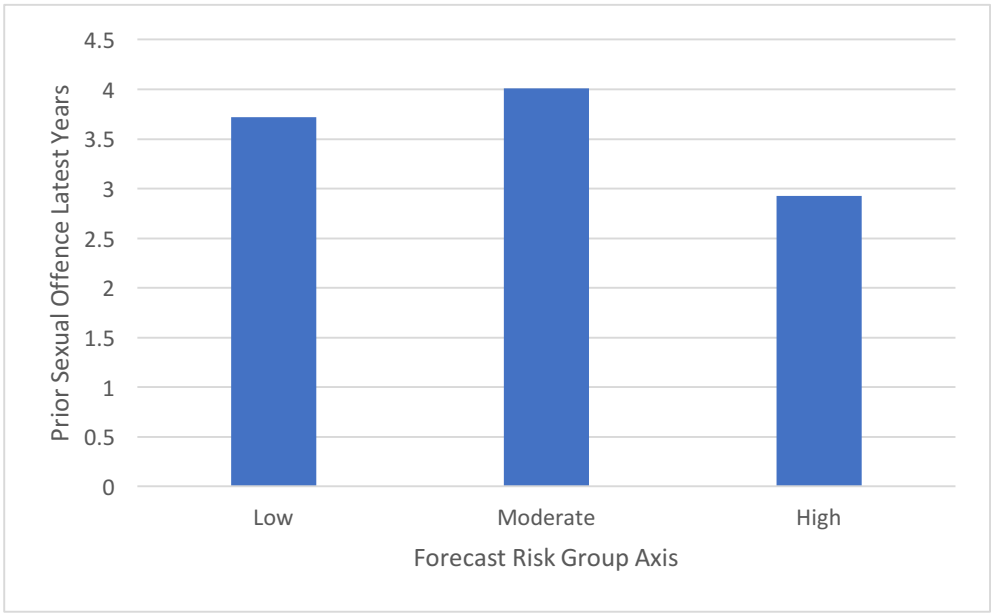

Figure 43: Mean number of years since the most recent custody instance for sexual offences

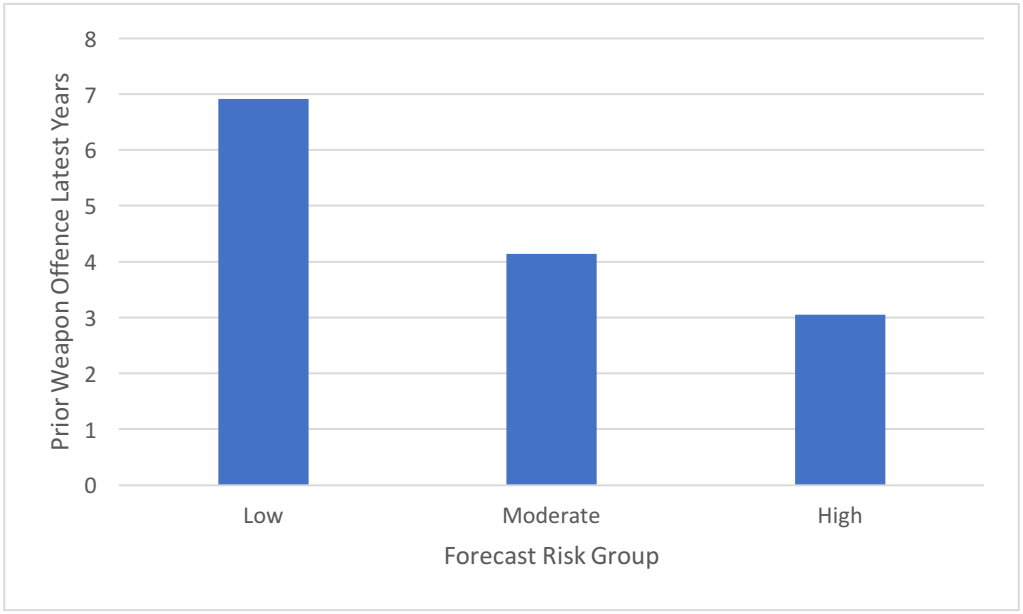

Figure 44: Mean number of years since most recent custody instance for weapon offences

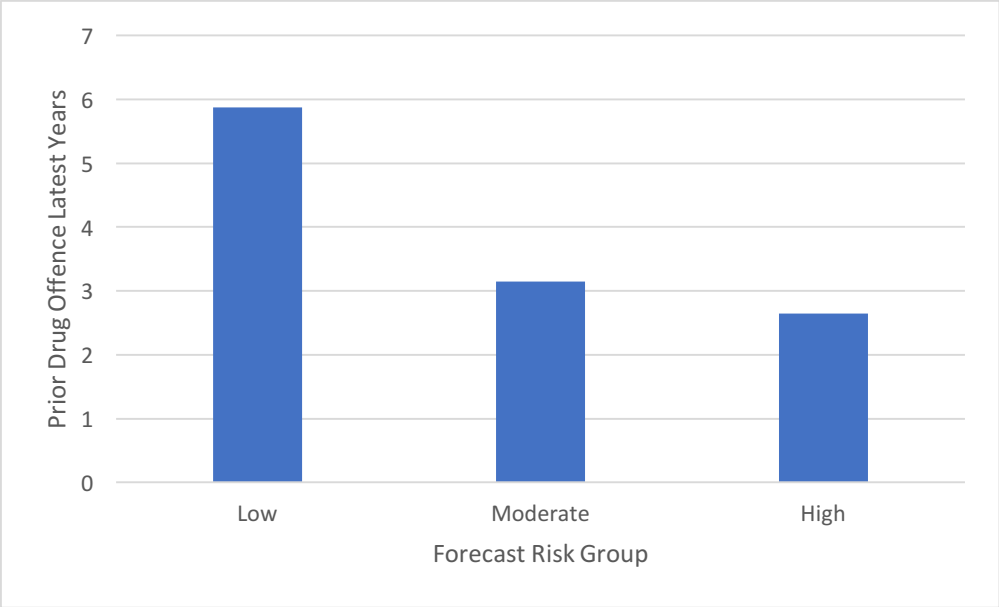

Figure 45: Mean number of years since the most recent custody instance for drug offences

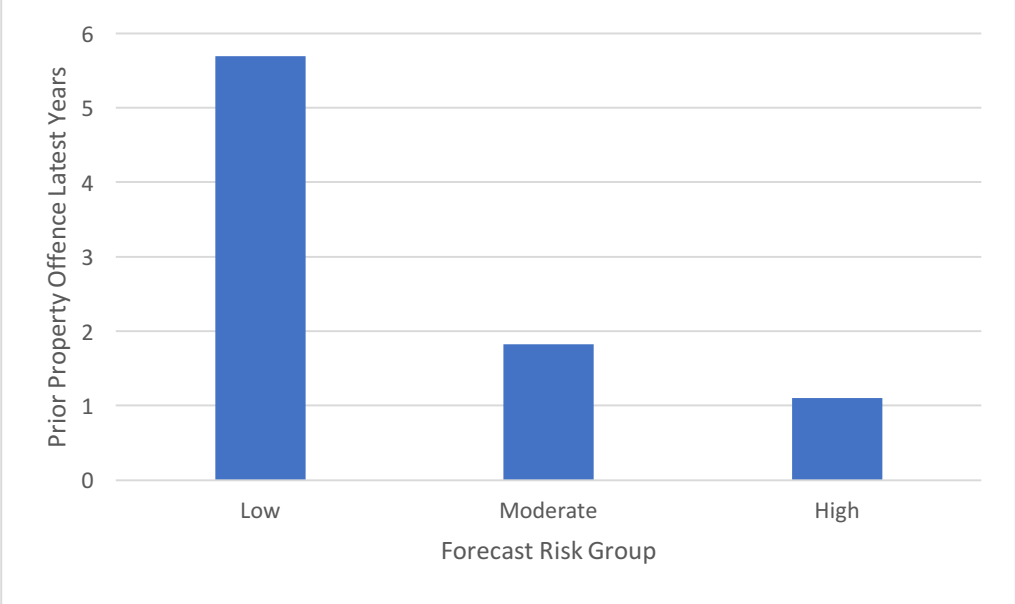

Figure 46: Mean number of years since the most recent custody instance for property offences

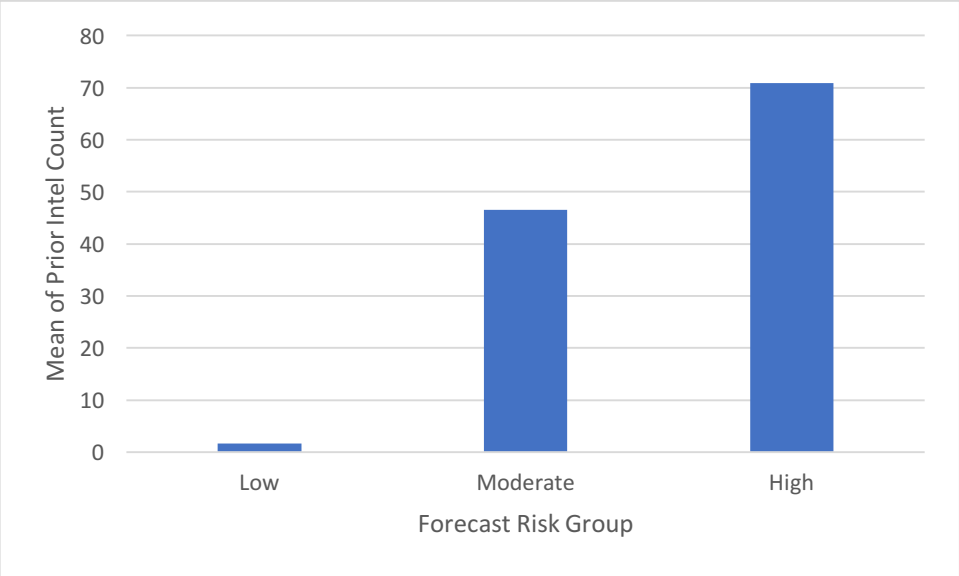

Figure 47: Mean number of intelligence submissions

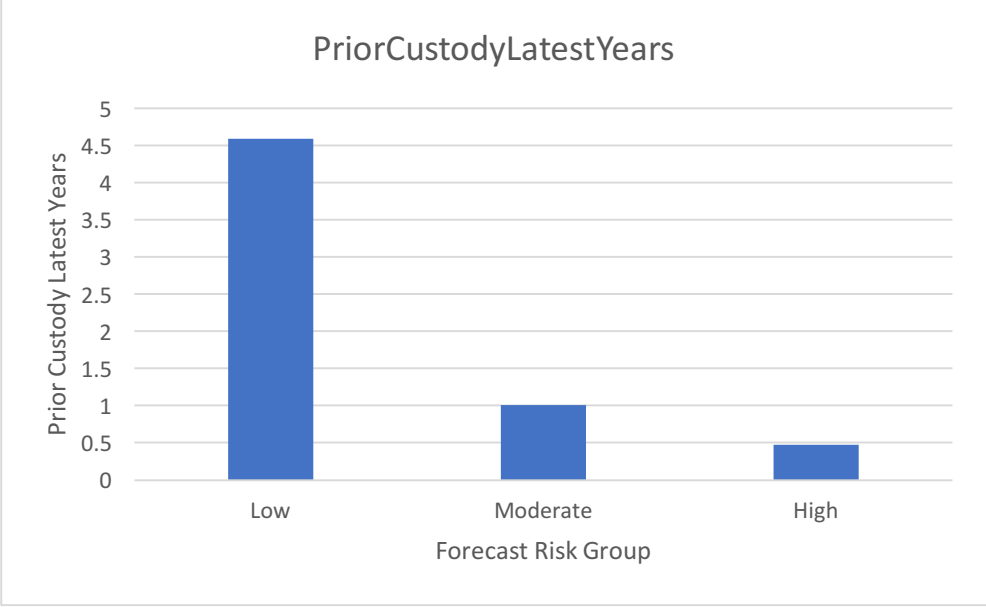

Figure 48: Mean number of years since the most recent custody instance