RISK PREDICTION IN FORENSIC PSYCHIATRY: A PATH FORWARD

RISK PREDICTION IN FORENSIC PSYCHIATRY: A PATH FORWARD

By: Devon Patrick Watts

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

Descriptive Note

MASTER OF SCIENCE (2020) McMaster University

Hamilton, Ontario, CANADA

DEPARTMENT:	Neuroscience
TITLE:	Risk Prediction in Forensic Psychiatry: A Path Forward
AUTHORS:	Devon P. Watts
SUPERVISORS:	Flávio P. Kapczinski, M.D., M.Sc., Ph.D.
COMMITTEE MEMBERS:	Jim Reilly, M. Eng., Ph.D., John Connolly, Ph.D.
NUMBER OF PAGES:	1-101

Lay Abstract

Individuals end up in the forensic mental health system when they commit crimes and are found to be not criminality responsible because of a mental disorder. They are released back into the community when deemed to be low risk. However, it is important to consider the accuracy of the method we use to determine risk at the level of an individual person. Currently, we use group average to assess individual risk, which does not work very well. The range of our predictions become so large, that they are virtually meaningless. In other words, the average of a group is meaningless with respect to you.

Instead, statistical models can be developed that can make predictions accurately, and at an individual level. Therefore, the current work sought to predict the types of criminal offences committed, among 1240 forensic patients. Making accurate predictions of the crimes people may commit in the future is urgently needed to identify better strategies to prevent these crimes from occurring in the first place.

Here, we show that it is possible to predict the type of criminal offense an individual will later commit, using data that is readily available by clinicians. These models perform similarly to the best risk assessment tools available, but unlike these risk assessment tools, can make predictions at an individual level. It is suggested that similar approaches to the ones outlined in this paper could be used to improve risk prediction models, and aid crime prevention strategies.

Abstract

Background: Actuarial risk estimates are considered the gold-standard way to assess whether forensic psychiatry patients are likely to commit prospective criminal offences. However, these risk estimates cannot individually predict the type of criminal offence a patient will subsequently commit, and often simply assess the general likelihood of crime occurring in a group sample. In order to advance the predictive utility of risk assessments, better statistical strategies are required.

Aim: To develop a machine learning model to predict the type of criminal offense committed in forensic psychiatry patients, at an individual level.

Method: Machine learning algorithms (Random Forest, Elastic Net, SVM), were applied to a representative and diverse sample of 1240 patients in the forensic mental health system. Clinical, historical, and sociodemographic variables were considered as potential predictors and assessed in a data-driven way. Separate models were created for each type of criminal offence, and feature selection methods were used to improve the interpretability and generalizability of our findings.

Results: Sexual and violent crimes can be predicted at an individual level with 83.26% sensitivity and 77.42% specificity using only 20 clinical variables. Likewise, nonviolent, and sexual crimes can be individually predicted with 74.60% sensitivity and 80.65% specificity using 30 clinical variables.

Conclusion: The current results suggest that machine learning models have accuracy comparable to existing risk assessment tools (AUCs .70-.80). However, unlike existing risk tools, this approach allows for the prediction of cases at an individual level, which is more clinically useful. The accuracy of prospective models is expected to only improve with further refinement.

Acknowledgements

I would first like to thank my supervisor, Dr. Flávio Kapczinski, who provided the leadership, freedom, scaffolding, and professional insights needed to complete this current work. I would also like to thank Taiane de Azevedo, Bianca Wollenhaupt, and Bianca Pfaffenseller for their guidance and assistance through the completion of my thesis. You have all taught me a great deal about the qualities needed in a good scientist, how to interact positively and maturely in academia, and have been instrumental in developing my problem-solving abilities. I would also like to thank Pedro Ballester, who has helped impart valuable insights into the theoretical components of machine learning, and who has been a conduit of creativity. I would also like to acknowledge several individuals for their support and insight throughout this process, including Drs. Jim Reilly, John Connolly, Benicio Frey, Paul McNicholas, Heather Moulden, and Mini Malak. Additionally, I would like to thank Dr. Gary Chaimowitz, for his leadership in exploring new territories in Forensic Psychiatry, and his support through the current work. I would also like to thank Daniela Russo and her family for their continual support and kindness. Finally, I would like to thank my family for their continued support and encouragement throughout the entirety of my graduate studies.

Table of Contents

3. CHAPTER 3: DISCUSSION

List of Figures/Tables

CHAPTER 2

Declaration of Academic Achievement

Dr. Flavio Kapczinski, Dr. Gary Chaimowitz, and I, were responsible for the development of the research questions associated with the primary research included in Chapter 2 of this thesis. Concerning the primary research presented in Chapter 2, I was the project manager of all associated study tasks, including data preprocessing, machine learning applications, model development, model refinement, model interpretation, and drafting the associated manuscripts for subsequent publication. Drs. Flavio Kapczinski, Gary Chaimowitz, John Connolly, and Jim Reilly provided continued supervision and oversight of the current work.

I would again like to thank everyone who was involved in the completion of this thesis.

CHAPTER 1

General Introduction

Forensic psychiatry is a multidisciplinary field comprising elements of criminology, law, and the diagnosis and treatment of complex and serious mental disorders [1]. Primarily, this involves navigating between justice, correctional and mental health services, given its focus on individuals with mental illness facing criminal convictions [2]. The field of forensic psychiatry largely emerged as a consequence of the interaction between legal, social and medical institutions, and the challenges faced when navigating their convergence [2].

With the integration of psychiatry into a field of medicine in the early 20th century, came the inevitable complication of addressing crimes committed by individuals as a result of their mental illness, and difficulties in determining what constitutes criminal culpability [3]. Additionally, changes in the legal system resulted from the emergence of psychiatry, as medical professionals could be called on to participate in legal decisions and provide their expertise on the mental health and competency of criminal defendants [4]. Forensic psychiatrists are tasked with assessing criminal responsibility, providing expert witness testimony, evaluating patients for prospective criminal risk, determining fitness to stand trial, appraising patient capacity and potential malingering, as well as possessing responsibility in civil actions and legal decision making [2]. Concurrently, primitive tests of cognitive or moral knowledge to determine criminal responsibility became slowly replaced by assessing illness progression and the presence of deranged mental states. This shifted the focus from disease of the intellect to disease of the mind [3].

Given the complexity of balancing both jurisprudence and the appropriate care for an individual patient suffering from mental illness, the field of forensic psychiatry faces several longstanding hurdles [1]. Namely, this involves assessing the fitness of a patient to stand trial, accurately determining whether an individual was competent at the time a crime was committed and judging the likelihood of each patient's prospective risk following their release [5].

Prevalence of criminality among the mentally ill

Considering the importance of identifying the risk that forensic patients pose to themselves and broader society, there is a longstanding and well-established body of literature examining criminality among psychiatric patients. Among the pioneers in the field, Zitrin et al. (1976) examined prospective criminal offenses in 876 inpatients discharged from a psychiatric facility. They found higher rates of subsequent arrest among psychiatric patients than those in the general population both in the same geographic region and among 4,601 cities in the United States [6]. Similarly, Klassen & O'Connor (1988) examined the relationship between arrests, hospitalization, and violence among 304 adult male inpatients at a community mental health centre, with 1-year of follow-up. They found higher rates of arrest and violent crimes among substance abusers, and notably larger violent readmission rates in patients with schizophrenia. Of note, they reported that a significant subset of patients in their sample showed a history of fluctuating between reimprisonment and rehospitalization in psychiatric facilities, highlighting the difficulty of appropriately managing such patients in legal and medical settings [7].

Following early work highlighting notable rates of criminality among those with serious mental illness, more recent efforts have largely focused on characterizing this from an epidemiological framework. For instance, in a systematic review of prevalence studies of serious mental illness among prisoners, comprising 33,588 individuals from 24 different countries, and 109 datasets, high rates of mental illness in prisoners were found in both high- and low-income countries over the timespan of four decades. Specifically, they reported a pooled prevalence of 3.6% (95% CI 3.1-4.2) in male prisoners with psychosis and 3.9% (95% CI 2.7-5.0) among female prisoners. With respect to major depression, the pooled prevalence was 10.2% (95% CI 8.8-11.7) in male prisoners, and 14.1% (95% CI 10.2-18.1) in female prisoners. Of note, they found that although rates of mental illness were high among prisoners, there is little evidence of an acceleration in prevalence over time [8].

Furthermore, in a study by Mullen et al, 2000, 10-years of hospital records and lifetime criminal records were assessed in 6130 patients with schizophrenia, and 6130 controls matched for age, sex, and place of residence. Here, the patient group involved records from two cohorts, with 3719 patients from a 1975 sample, and 2411 patients from a 1985 sample, respectively, to account for potential generational effects. In the 1975 sample, they found that those with schizophrenia showed a 3.5 relative risk of reoffending [95% CI 2.0-5.5], p=0.001, for all categories of crimes, apart from sexual offenses. A similar finding was observed in the 1985 sample, where a 3.0 relative risk of reoffending was reported [95% CI 1.9-4.9, p=0.001] [9].

Likewise, Simpson et al (2004) found that in a sample of 1498 homicides, 8.7% were conducted by those with a serious mental illness. Among them, 29% of those with a serious mental illness showed no history of hospitalization, and of those were admitted, most were only hospitalized on one or two occasions over the last five years. This is suggestive that capital offenses such as murder are not isolated to a subset of the most severe cases who inevitably become repeat offenders. Rather, only 10% of the perpetrators were admitted to the hospital in the month prior to their offense [10]. Additionally, in a sample of 295 inpatients with serious mental illness, 49% of men and 39% of women were found to have committed a form of assault in the past 6 months. Further, rates of crime were found to be higher than the general population. This suggests that aggressive behaviour is a prevalent problem among patients with severe mental illness who require hospitalization [11]. Cumulatively, this work has helped elucidate the societal implications of criminality among a significant minority of patients with serious mental illness, and the importance of developing proactive and accurate ways to assess patient risk of subsequent crime.

Reoffending: prevalence and assessment tools

Although the rates of reoffending in the forensic population remain relatively constant [12–14], available evidence suggests that one in eight men, and one in sixteen women will subsequently commit a grave offense after release from a psychiatric facility [15].

Similarly, a recent study from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) involving 35,306 individuals showed that the presence of mental illness, irrespective of the specific disorder, was associated with a 4 to 5 times greater risk of criminal outcomes [16]. Of note, 28.5% of the participants with mental illness reported a history of criminal behaviour, while a substantial subset, 11.4%, reported a history of incarceration [16]. Additionally, results from a large Swedish registry study comprising 98,082 individuals with a history of hospitalization suggests that those with severe mental illness commit one in every twenty violent crimes [17]. Given the high prevalence of criminal reoffending across cultures in individuals with severe mental illness, there has been a concerted effort to identify predictors of prospective criminal risk following release from psychiatric facilities.

Prior to the development of any standardized tools, clinical judgement was the gold-standard measure to assess prospective patient risk [18]. However, this presented a number of clear limitations, including poor inter-rater reliability between clinicians, confirmation bias, and the propensity for human error [19]. Importantly, clinical judgement alone has not provided a more valid metric by which to identify individuals with mental illness who will prospectively commit serious crime [19].

In response to this, actuarial assessments became increasingly widespread, which concentrated on statistical models, while largely disavowing clinical judgement [20]. This involved using explicit statistical algorithms to identify prospective patient risk, usually at the group level [21]. For instance, the Violence Risk Appraisal Guide (VRAG) is an actuarial assessment of prospective violent risk, developed at the University of Toronto, in Ontario, Canada [22]. This 12-item scale is based solely on historical and relatively static factors for individuals who have committed previous violent offenses [16]. More specifically, the VRAG considers items such as age, marital status, criminal history, psychopathy, previous diagnoses and prior separation from parents.

Validation studies [22–25] have shown that the VRAG identified prospective risk of violent recidivism with an Area Under the Curve (AUC) of 0.73-0.75. However, this metric has been criticized both for being overly simplistic and for its inability to detect prospective violent crime in individuals without prior offenses [22]. Given these clear limitations, the authors of the VRAG recommend supplementing this assessment with clinical judgement [23].

Likewise, other methods available to detect prospective violent recidivism show similar limitations. For instance, the Historical, Clinical and Risk Management Scales (HCR-20) is a structured actuarial assessment that uses pre-discharge information to identify those at risk for committing prospective violent offenses. This comprises 20 items, 10 of which are related to historical factors (e.g. history of mental illness), 5 are related to current clinical presentation (e.g. current symptoms of mental illness), and 5 are related to future risk (e.g. non-compliance with medication). Each item is scored as 0 (not present), 1 (partially/possibly present), or 2 (present), leading to a maximum score of 40, with maximum sub-scores of 20 for the historical scale, and 10 for clinical and risk scales, respectively [26].

In a study comprising 887 male patients discharged from a medium security unit with a 2-year follow-up, the HCR-20 was found to be a good predictor of prospective violent offenses at the group level, with AUCs in the 0.70-0.76 range. Nonetheless, the HCL-20 evaluates clinical risk by stratifying individuals into low, moderate, or high-risk categories, which can lead to an ambiguous prognosis if an individual is identified to be of a moderate risk [19]. Further, while this assessment has been shown to identify prospective violent recidivism with high AUCs, among those identified as high risk, false positive rates between 70-80% have been reported [27].

Moreover, the accuracy of these risk assessment tools varies as a function of the clinical population they are administered to. For example, in a study assessing the predictive validity of the VRAG, the H-10 scale of the HCR-20 and the Psychopathy Checklist Revised (PCL-R) among 169 inpatients with schizophrenia, all risk assessment scales showed poor predictive capabilities in identifying those who would subsequently commit violent crime. Of note, the performance of these instruments in identifying recidivism were similar to simply identifying patients with greater symptom severity and chronicity [28].

Considering the clear limitations of current strategies in detecting which patients will subsequently commit violent crime, there is a major unmet need for an actuarial tool that can be used at an individual level. In the absence of this, given the high false positive and false negative rate of gold-standard actuarial tools, a substantial number of patients will be mischaracterized as either high or low risk for committing violent crime if released from psychiatric care. As such, this precipitates unnecessarily denying civil liberties of patients who will not subsequently reoffend on the one hand and endangering the lives of those in the community when they do, on the other hand.

Actuarial Methods: sexual risk prediction

Apart from screening patients for subsequent violent crime risk, several actuarial risk assessments have focused on the challenging goal of identifying the likelihood of prospective sex offences. While there are a number of actuarial methods, as described elsewhere [29,30], the most replicated assessments include the Static-99, Rapid Risk Assessment for Sex Offense Recidivism (RRASOR), Sex Offender Risk Appraisal Guide (SORAG), the Risk-Matrix 2000 (RM2000/S) and the Static-2002 [31].

Sex Offender Risk Appraisal Guide (SORAG)

The SORAG is a modification of the VRAG, indicated for evaluating the risk of sexual offences. The primary difference between these two assessments involves the addition of questions on the number of previous sexual offences, and phallometric test results indicating sexual deviance [32]. In a validation study assessing the psychometric properties of this assessment in 1104 sexual offenders released from an Australian prison, the SORAG showed an AUC of 0.66 in identifying sexual recidivism, with a 95% Confidence Interval (CI) of 0.61-0.72 [33]. Of note, the SORAG demonstrated better performance in identifying violent and general recidivism with an AUC of 0.74 [34].

Risk Matrix 2000 (RM2000/S)

The Risk Matrix 2000 was developed as an easy to score actuarial assessment in identifying violent and sexual recidivism among adult males who have been convicted of sexual crimes [35,36]. This tool involves subscales that assess the risk of sexual recidivism, non-sexual violent recidivism, and sexual violent recidivism, respectively [37]. Previous research has found that each subscale demonstrates a moderate effect size in identifying prospective recidivism (Cohen’s d= 0.50-0.64), defined as the difference between two group means divided by the pooled standard deviation [37]. However, significant variability in this effect has been observed across studies [37,38]. Intuitively, the sex subscale showed a larger effect size in sexual recidivism, with a Cohen's d of 0.74 [37].

Static-99

The Static-99 was developed using fixed variables that correlate with sexual reconviction among adult males [39]. This involves an amalgamation of two risk assessments, including the 4-item Rapid Risk Assessment of Sex Offender Recidivism (RRASOR), and the 9-item Structured Anchored Clinical Judgement - Minimum (SACJ-Min). Combined, this assessment incorporates age, the sex of the victim, relationship with the victim, prior offences, prior convictions, victim characteristics and marital status [40]. In previous studies, the Static-99 was found to show moderate predictive accuracy (AUC 0.71) in identifying sexual recidivism [41].

Static-2002

The Static-2002 emerged as a revision to the Static-99 to improve both consistency in scoring, and the performance of the metric. It was developed to identify risk of sexual recidivism among adult males who have committed previous sexual offences. This involves 13 items in five separate categories, which includes age at release, persistence of sexual offences, deviant sexual interests, victim characteristics, and questions pertaining to general criminality [29]. In a validation study comprising 10 datasets with a total of 4596 offenders, the Static-2002 showed an AUC of 0.71 for sexual recidivism (N=2142) and 0.71 for any violent recidivism (N=2143) [37,42]. However, in a follow-up study comprising 468 sex offenders followed for an average of 5.9 years, the Static-2002 showed an AUC of 0.69 (95% CI: 0.59-0.78) in identifying serious violent recidivism, and an AUC of 0.67 (95% CI: 0.51-0.81) in identifying sexual recidivism [29]. Altogether, this suggests high variability in the performance accuracy of identifying recidivism.

Similarly, in a systematic review assessing the effectiveness of sexual offender risk assessments, a large degree of variability was observed both within and across instruments. For instance, the Static-99, which is among the most replicated actuarial tools, shows an average AUC of 0.692 with a range between 0.570 to 0.920 depending on the study [31]. Overall, the performance of actuarial tools ranged from an average of 0.692 to 0.737, with less replicated studies showing generally greater accuracy [31]. This suggests that many instruments may present with over-optimistic performance metrics.

Limitations of actuarial assessments

Although actuarial methods have helped shift the focus of risk assessment in forensic psychiatry toward objectivity and reproducibility, there are a number of limitations to their widespread use in making clinical decisions. Namely, there is little evidence that actuarial methods perform any better than clinical judgement in identifying which individual patients will subsequently reoffend [43]. This is largely because most actuarial instruments have been developed psychometrically to assess group-based risk and perform poorly when making individualized predictions [34].

This phenomenon is related to the difference in calculating CIs between a group effect and an individualized prediction, with the latter showing a wider prediction interval [34]. Of note, this higher variability in the prediction interval when making individualized predictions is to a great extent independent of the sample size. This is because while increasing sample size can result in a narrower confidence interval around a regression line in a group-based analysis, this does not translate into a narrower prediction interval in an individual case [44]. Importantly, this is a problem not specific to any individual actuarial assessment, but a result of the high variability in group-based assessments when dealing with individualized predictions [33]. Given the high degree of error using available methods, it is difficult to determine the risk of any individual patient relative to another [21].

Statistical methods used to assess performance

In a similar vein, it is important to appropriately characterize the strengths and limitations of the statistical methods used to evaluate the performance of any given actuarial tool. Most studies thus far have focused on receiver operating characteristics (ROC) curve analyses. A ROC curve is a plot of sensitivity versus 1-specificity (often called the false-positive rate) that offers a summary of sensitivity and specificity across a range of cut points for a continuous predictor [21]. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR). The TPR, also known as sensitivity, is the proportion of actual positive cases that are correctly identified. Conversely, the FPR, or specificity, is the proportion of actual negative cases that are correctly identified [44].

Therefore, the sensitivity and specificity for a given cut-point are the probabilities of correctly identifying a person's group status (i.e. identifying true positives and true negatives) [45]. For instance, if a clinician predicts that a given patient will commit a violent offence within one year after release from a psychiatric facility, and that assessment is found to be correct, this represents a true positive. Conversely, if a clinician predicts that a person will act violently within a year following release and this does not occur, this represents a false positive. In general, an AUC of 0.5 suggests that the model has no discriminative capabilities above chance (approximately 50% correctly classified), while 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered exceptional, respectively [46]. When discrete variables are used, for example, the number of previous violent offences, and when ROC curves are assumed to be based on two underlying Gaussian distributions, a maximum likelihood estimation can be used to fit the data to a smooth curve. This provides a metric to calculate the area under the fitted curve and its associated standard error [47].

The first study in forensic psychiatry to use a ROC curve within an actuarial analysis was Mossman (1994), who attempted to address challenges in conceptualizing accuracy without accounting for base rates or biasing certain outcomes over others. This involved reanalyzing 58 datasets from 45 published studies on violent risk prediction, and this seminal paper established past behaviours as better predictors than clinical judgement in violent risk [48]. Since this point, AUC has become the standard metric to evaluate the performance of risk assessment tools.

What does AUC mean statistically?

Given that AUC has become a ubiquitous measure to assess performance of risk assessment tools, there are several longstanding misconceptions surrounding its use. AUC is not the probability that individuals are classified correctly, or that a person with a high-test score will eventually become a case. Instead, it is a plotting of the positive prediction rate as a function of the negative predictive rate [49]. In other words, it is a mapping of the sensitivity by 1 minus specificity [44,45]. In contrast, the predictive value more closely parallels correctly identified cases, as it is the probability that subjects with a positive screening test truly have the condition [40]. As such, AUC describes how well models can rank order cases and non-cases but is not a measure of the actual predicted probabilities [44].

Why AUC and not another metric?

Effect size relates to the estimated magnitude of an effect. Essentially, this provides a way to evaluate the practical or clinical importance of a statistically significant finding. A common way of assessing effect size, apart from AUC, involves Cohen's *d*, which indicates the standardized difference between two means [41]. A *d* of 1 indicates that two groups differ by one standard deviation, and larger numbers of *d* are proportional to the number of standard deviations of difference between groups. For instance, a *d of 2* reflects that two groups differ by two standard deviations, and so on [41]. In his original work, Cohen suggested that *d*=0.2 could be considered as a 'small' effect size, with 0.5 representing a 'medium' effect and 0.8 representing a large effect [42,43]. Essentially, a lower Cohen's *d* indicates the necessity of generating larger sample sizes. Of note, there are a number of alternative methods available to assess effect sizes, apart from AUC as described elsewhere [44–46].

However, AUC is more commonly used in forensic psychiatry for a number of reasons. Namely, Cohen's *d* was designed for situations where scores were compared between two continuous and normally distributed populations, a circumstance that is seldom the case when dealing with prospective criminal risk [47]. Similarly, with Cohen's *d,* the effect size is dependent upon both the base rate of the variable in question, and the given correlation coefficient. Moreover, although the coefficient of determination, colloquially referred to as R-squared, provides another metric to evaluate effect size, in certain circumstances, it can substantially mischaracterize the importance of a finding, especially in situations where one of the variables is dichotomous [47]. In summary, AUC provides a reasonable metric to assess the relevance of an effect.

Need for confidence intervals

Given that AUC does not inherently assess whether any given patient was correctly classified by a model, and the real-world implications of risk assessment tools in the lives and liberty of individuals, supplementary statistical approaches are warranted. Namely, a number of authors have recommended the use of CIs and corrections for measurement error to be used in conjunction with ROC curves [49–51]. Confidence interval testing relies on proportions and involves two main types. The first involves exact 95% CIs, which use a binomial distribution to reach an exact estimate. The latter assumes a given dataset shows a normal approximation of the sampling distribution. However, when the number of outcomes is small, or the sample size is small, this assumption of normality cannot be met, and CIs are required [48].

Confidence intervals are usually interpreted as a range of values encompassing the population or 'true' value estimated by a certain statistic, with a given probability. In the case of 95% CIs, the range provided would be expected to include the true value of the variable or outcome of interest, with a 5% chance of being incorrect [49]. The standard deviation of effect size is particularly important, as it provides a way to assess the level of uncertainty within a given measurement. In cases where the standard deviation is too large, the measurement is rendered virtually worthless [52].

Therefore, to derive a more realistic framework when evaluating the performance of a given model, CIs should be considered. For example, in a systematic review evaluating the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders, the mean AUC, the number of studies used to derive the mean AUC, and 95% CIs were reported for each actuarial tool. While more replicated assessments showed moderate AUCs ranging from 0.666-0.692, the 95% CIs indicated large disparities, with a range from 0.420-0.920 for the Static-99, RRASOR, SORAG, and Risk Matrix-2000 [31].

Similarly, in a study by Hart et al. (2007), the precision of two commonly used actuarial tools, comprising the Static-99 and VRAG, were evaluated using 95% confidence intervals for group and individual risk assessments. They found large confidence intervals for risk estimates at the group level, whereas at the individual level the margin of error of CIs were so high that the risk estimates imparted little clinical utility [21]. Indeed, in an analysis of the predictive accuracy of the PCL-R in identifying criminal recidivism at both the group and individual level, the range of 95% CIs grew the further the individual patient scores were from the group average. When estimating the CIs for the likelihood of reoffence at an individual level, the CIs showed high variability (0-98%), to the point that they were essentially meaningless. The authors noted that clinicians should derive little confidence in actuarial scales to determine an individual's likelihood of reoffending [34].

Limitations of AUC

Importantly, when assessing the performance of clinical tools using ROC curves, the balance between sensitivity and specificity should be considered of equal importance to the AUC overall. This is because a test can be very sensitive without being specific, or very specific without being sensitive, while still showing a reasonable result [53]. As such, a useful clinical tool will show both a good balance between sensitivity and specificity. Furthermore, the accuracy of an assessment is also determined by how common the outcome or target variable is in the sample. For instance, in situations where the base rate of an outcome is low, such as recidivism only occurring within a small proportion of cases, it is possible to make a better prediction by automatically assuming in each instance that the phenomenon will not occur, rather than trying to ascertain the actual probability of the event [41]. Considering this, the use of a single measure of model fit such as AUC can erroneously eliminate important clinical risk predictors for consideration in scoring algorithms [44]. Given this, ROC curves alone may not be an optimal evaluation metric for models that predict prospective risk or ascertain the risk profile of an individual [44].

Variable selection

Furthermore, concerns have been raised as to the specific variables used to derive actuarial based risk assessments in forensic psychiatry. In general, actuarial methods place a strong emphasis on static risk factors, which are features largely unamenable to change, such as prior offenses, and childhood experiences [54]. However, this largely discounts transient risk factors that may be of importance and potentially modifiable, such as drug abuse, poverty and housing instability. Therefore, by overemphasizing static risk factors without identifying risk factors that are malleable, we run the risk of stigmatizing patients based on their past, without identifying new strategies to improve rehabilitation efforts and decrease prospective recidivism.

Assumption of a linear and additive relationship between risk factors

Apart from the ethical concerns related to the use of actuarial tools in clinical practice, there is the important consideration of the statistical relationship between risk factors used to derive these tools. Most actuarial assessments assume that there is a linear relationship between dependent variables, while others also assume an equal correlation strength between each item and the target outcome. For instance, the HCR-20 determines prospective risk by assessing whether each risk factor is deemed to be absent (score of 0), possibly/partially present (score of 1), or present (score of 2) [26]. As such, each item is inherently assumed to possess the same linear relationship with the outcome, and any interaction effects that exist between variables are ignored.

Additionally, other actuarial strategies, such as the VRAG are posited to reduce such bias by applying a weighting to certain items based on the Pearson Correlation Coefficient, a method that uses covariance to examine the association between variables of interest. Therefore, items with a higher correlation to the outcome, such as a history of bank robbery, have greater weighting relative to other items with a lower correlation such as history of indecent exposure. However, this approach can be susceptible to bias related to the base rate of each dependent variable [55].

Base rates refer to the percentage of a population that demonstrates some characteristics. In this case, certain items may have less weighting simply as a function of being infrequent within the population sample, rather than their actual relative contribution to the outcome [56]. Using the previous example, a history of indecent exposure may be a very important variable in predicting prospective recidivism, but given the rarity of this in clinical populations, it may perhaps be overshadowed by more common factors, such as previous violent behaviour. While it has been argued that such risk assessment tools should be readily interpretable and easy to score by clinicians, it is unlikely that complex phenomena such as predicting criminal recidivism at an individual level can be appropriately modeled using simple linear equations. Altogether this highlights the need for caution in using actuarial assessments to determine prospective clinical risk and the necessity for new tools.

Potential new approaches - Artificial intelligence tools

Given the ethical, psychiatric and legal ramifications of inappropriately mischaracterizing the prospective risk of any given patient, and the resulting consequences to the individual, their families, and broader society, there is a growing interest in the use of artificial intelligence and predictive analytics to facilitate greater accuracy in clinical decision making.

A possible solution to these limitations may lie in the use of machine learning, a field of artificial intelligence that focuses on extracting value from datasets using computational algorithms. These algorithms can detect patterns within a dataset and then apply what they learned to make predictions in unseen data [57]. Unlike traditional statistical approaches that evaluate average differences in outcomes between groups, machine learning methods provide a more straightforward way to predict outcomes at the individual level [56]. This can potentially pave the way for tailor-made tools for the diagnosis, assessment, and treatment of patients [58,59].

Furthermore, machine learning can deal with complex data containing a large volume of information that can be created at a high velocity, and in a wide variety of types, three essential characteristics found in big data[60]. Given the inherent challenges with current actuarial strategies, there is also a growing interest in developing more objective and reliable tools to assess forensic populations[61]. Although these techniques have shown promising results in other fields of science and medicine [62–68], their value for forensic psychiatry has yet to be fully explored.

Differences between actuarial and machine learning approaches

To understand the differences between actuarial and data-driven machine learning approaches to risk management, it is important to briefly discuss where they diverge philosophically and statistically. While both attempt to capture relationships between dependent variables to model an underlying phenomenon and use information from past occurrences to predict future outcomes, there are noteworthy differences in how this is achieved.

Namely, actuarial science is concerned with the probability of certain events occurring, using a group-average aggregate of risk predictors [69]. As such, the primary consideration is risk management, and identifying relevant factors that precipitate higher risk among individuals. This involves a large degree of statistical approaches, including Bayesian inference and generalized linear modelling. While these approaches can also be used in a machine learning context, actuarial science uses domain knowledge to select relevant variables and is oriented toward understanding the underlying phenomenon of interest. As such, actuarial methods statistically analyze patterns of data in stochastic and deterministic scenarios to explain an outcome, placing less of a focus on making precise predictions [70].

Machine learning classification models, on the other hand, are more concerned with predicting a phenomenon of interest with the highest possible accuracy [71]. Indeed, model optimization represents an entire subfield within machine learning [72] . However, in machine learning, interpretability can become a difficult problem [73], especially when using more sophisticated algorithms. Despite this trade-off, machine learning methods provide the benefits of an exploratory approach to selecting relevant variables in classification problems, based on a data-driven, rather than a hypothesis-driven framework [74]. As such, this provides a greater degree of flexibility in feature selection [75], which is an integral component of model development. This is especially important if there are latent or unexamined variables within a given dataset that are useful risk factors but have not yet been identified in previous literature.

Similarly, this approach is often more conducive to novel discoveries, which may be better suited to capturing the idiosyncrasies of a specific population. For instance, a common problem highlighted among actuarial risk assessment tools is the difference in performance accuracy in predicting sexual recidivism between subpopulations of sexual offenders [76]. This suggests that the relationship between risk factors may not be linear across these populations. By disregarding the assumption that each risk factor is related in a linear fashion, machine learning methods can more easily examine the complex interactions between variables to make individualized predictions.

Interpretability in machine learning

Generally speaking, more sophisticated algorithms tend to lead to higher performance in predicting outcomes. This provides a unique avenue to address problems where more traditional statistical analysis techniques have struggled, such as individualized risk prediction. However, as models grow in complexity, this carries the trade-off of greater difficulty in model interpretability and explainability [77]. This is especially the case for non-linear classification models, which use a nonlinear combination of model parameters to predict a specified outcome.

Complex machine learning models have been commonly referred to as ‘black-box’ methods, since we have information about the input and output of the model but lack detailed knowledge about the specific decision-making process [78]. As such, some authors have expressed concerns about applying predictive models in real clinical scenarios if we do not fully understand the functioning of such models[60,61]. This argument follows that opaque models bring forth several concerns, especially when dealing with applications in law and healthcare. Moreover, others have called for an end to using black-box models for high-stakes decisions, and instead advocate for the use of readily interpretable models such as decision trees [79].

Decision trees are non-parametric, since they make no assumptions on the distribution of the data, and are highly structured, and therefore are the most interpretable machine learning model [80]. However, methods such as these are prone to several limitations. For instance, a small change in the data can precipitate a large change in the structure of an optimal decision tree. This presents challenges in performance when applying a model to an independent dataset [81]. Furthermore, decision trees tend to show poor performance against more sophisticated algorithms [82]. There are various ways to substantially improve the performance of decision trees, such as bagging and boosting, but this performance increase occurs at the expense of ready interpretability [83].

However, the dichotomy between interpretable and black-box methods in a strict sense may not be entirely accurate. While the definition and standards of interpretability largely vary between applications, there are a number of standardized interpretability metrics available to assess our models quantitatively. As such, it is possible in some capacity to peer into the black box. Indeed, explainable artificial intelligence (XAI) is a growing field which aims to better understand how black-box methods make key decisions, in order to improve trust and transparency in machine learning applications [84]. In line with this effort, there are a number of ways to improve model interpretability among so-called black box methods [77,85]. For instance, feature relevance and model visualization play a key role as the intermediary between a black box model and the human expert [86]. Likewise, methods such as variable importance plots can be shown, which lists the most significant variables in descending order. The more a model relies on a variable to make predictions, the greater importance of the given variable [87]. By showcasing which features were most important for a given model, and visualizing this with several available methods as described elsewhere [87,88],[89], such approaches can serve as useful tools to assist in the interpretation and comprehension of a model's decision making process [90].

Generalizability in machine learning

Another important consideration in the use of machine learning for classification problems is that of the generalizability of the model. A model that has the highest accuracy in a testing dataset may not necessarily show the highest accuracy in an independent dataset. This is especially a problem when dealing with high-dimensional data, when the number of features exceeds the number of instances [91]. In other words, when there are more predictor variables than there are patients in a sample, there may be redundant variables that harm the model performance [91]. Additionally, redundant variables run the risk of a model learning based on irrelevant features, which can notably decrease its performance in an independent dataset [92].

Feature selection and feature extraction provide two notable ways to address the problem of irrelevant features and dimensionality in a dataset. Briefly, feature selection involves obtaining an important subset of the original features according to a specified selection criterion. Thus, feature selection removes redundant and irrelevant features from a dataset, prior to running the model [92]. There are a number of feature selection methods, with varying degrees of appropriateness depending on the application, as described elsewhere [93,94],[95]. Of note, limiting the number of features tends to improve the generalizability of a model when applied in independent datasets [96].

Conversely, feature engineering refers to creating new input variables from existing ones [97]. This, understandably, requires a degree of domain knowledge about the data itself.

Among the important considerations of feature engineering are imputation methods and feature splitting. Imputation methods are required to deal with missing data, as a majority of algorithms cannot be used on columns containing missing values. Further, feature splitting uses parts of a column to create new features. This helps uncover information that may be useful to the performance of a given model [97].

Assessing the performance of classification models

Confusion Matrix

This leads into the question of how model performance is assessed when dealing with classification problems. Commonly, a confusion matrix is used, which is a table layout that provides a visualization of the performance of the model. This includes the number of correct and incorrect predictions, which are summarized with count values and broken down by each class. For instance, a confusion matrix contains information about overall accuracy, misclassification rate as well as true and false positive and negative predictions. A more detailed explanation of confusion matrices, as well as how they differ between binary and multiclass predictions can be found elsewhere [98].

Cross-validation

Importantly, no single algorithm is necessarily superior across classification problems. Although certain algorithms may be more or less appropriate depending on the nature of the problem, this still necessitates a comparison of multiple algorithms [99]. Moreover, each algorithm has its own set of hyperparameters, which are used to control the learning process, that must be set prior to running the model [100]. In order to evaluate the optimal set of hyperparameters for any given model, cross-validation can be used. Cross-validation is a resampling procedure, where data is divided into *k* subsets or folds [101]. This holdout method is repeated several times, where one of the *k* subsets is used as the test or validation set, while the other *k*-1 subsets are used to form a training set. The error estimation is averaged over all *k* trials to receive a metric on the set of hyperparameters that are optimal for our model [102].

Previous machine learning studies in risk prediction

Considering the benefits of using machine learning algorithms to model complex phenomena and in making individualized predictions, there have been a handful of groups that have pioneered its introduction in a forensic risk prediction context. For instance, Falconer et al., assessed predictors of rearrests within the first 90 days of release from jail, among 2100 adults involved in the criminal justice system, using clinical and demographic variables[48]. The authors reported an AUC of 0.67 using a generalized linear model within a hold-out testing dataset. Similarly, Caulkins et al. assessed predictors of criminal recidivism, among 3508 offenders, within a two-year period following release from federal prison using clinical and administrative data. The authors compared the performance of logistic regression and Artificial Neural Network (ANN) models using either eight, eleven or eighteen predictive variables. The best performance was observed in the model with 18 variables, with an AUC of 0.689 using logistic regression, and 0.699 using ANN [103].

In another study using big data analytics, Palocsay et al. assessed predictors of criminal recidivism among 19,136 individuals released from federal prison using nine clinical and demographic variables. The authors compared the performance of linear regression and artificial neural networks and split the dataset into training (9457 prisoners) and testing (9679) sets. The highest performance was observed using ANNs with an accuracy of 69.23%. However, the sensitivity of all models was poor, with 30.41-41.26% of recidivists correctly identified, and approximately 81.07-88.43% non-recidivists correctly identified. While the performance accuracy overall was similar between models, ANN showed a modest improvement over logistic regression [104].

Moreover, Pflueger et al. used demographic and clinical data to identify predictors of general recidivism in a sample of 365 offenders with a history of mental illness, from which 128 were re-offenders[44]. The authors used a random forest algorithm and performed feature selection according to a CART criterion, by excluding the variables with the least importance to the model. Using six variables, they developed three models with pre-specified sensitivity and specificity weightings. The first model was equally weighted and showed an accuracy of 85%, with 84% sensitivity, and 86% specificity, respectively. The second model used a 95% sensitivity cut-off and achieved a specificity of 0.60 for an overall accuracy of 77%. The third, and final model, used a 95% specificity cut-off and achieved a sensitivity of 0.58, with an overall accuracy of 77% [105].

Among the few available machine learning studies predicting criminal behaviour in forensic patients with mental illness, Linaker et al. investigated the predictors of imminent physical violence requiring restraints in 92 inpatients. Within this sample, 48 incidents of violence occurred in 32 patients. Using factor analysis as a feature selection method, they identified six behavioral variables that were common before violence. Subsequently, these six variables were used to build a logistic regression model, where they obtained an accuracy of 92.1%, with a sensitivity of 81.3% and a specificity of 100% in the testing set[35]. While these results seem quite promising, it is important to note the low prevalence of the outcome event with only 15 patients in the training set, and 17 patients in the testing set displaying violent behavior. Moreover, only a few instances of each behavior were observed among a small subset of patients. Altogether, this suggests that the model accuracy reported is likely to be over-optimistic.

Similarly, Monahan et al. monitored a sample of 939 psychiatric inpatients for violence, assessing 106 risk factors[37]. Using an iterative classification tree (ICT) approach, 72.6% of the sample was classified as either low or high risk. Instead of using a cross-validation procedure, however, the authors used a bootstrapping approach, which is less reliable in estimating model generalizability[37]. In another study using the same sample, the authors obtained an AUC of 0.81 with the ICT and 0.79 with a standard classification tree to classify patients as high or low risk using clinical assessments[76]. Additionally, Soini et al. used a Naive Bayes classifier coupled with clinical variables to predict forensic admission in a sample with 308 psychiatric patients[106]. Of note, the authors used independent data sets from four different countries to train and test the predictive model. They achieved accuracies between 86.1 to 91% in the training sets, and between 82.5 to 87.1% in the testing sets[106].

Aim of the current thesis

Although a limited number of prior studies have assessed the use of machine learning applications to predict criminal recidivism, there are several limitations in available studies. Namely, prior studies have only assessed the presence or absence of recidivism, rather than analyzing the type of crime and whether this varies as a function of the patient population. No studies thus far have used machine learning models to predict the type of crime that an individual would subsequently commit. Such an approach would be more suited to a tailored intervention at an individual level.

Moreover, many previous studies suffer from a substantial class imbalance problem. This occurs when the total number of a class of data is far less than the total number of another class of data. For instance, in a study predicting patient aggressive events in a psychiatric hospital, as little as the data set presented with the outcome of interest (Suchting 2018). Additionally, a majority of studies present a lack of transparency in the exact features used in the model and lack a representation of feature relevance or model visualization when presenting their results. This limits the interpretability and replicability of risk prediction models in the field. Therefore, the current study attempts to address these drawbacks, by analyzing the type of crime committed among recidivists, building separate models on the basis of gender, and mental status, as well as using feature selection, feature relevance and model visualization techniques to facilitate the replicability of its findings.

References

1. Arboleda-Flórez, J. Forensic psychiatry: contemporary scope, challenges and controversies. *World Psychiatry* (2006).

2. Monden, Y. *Principles and Practice of Forensic Psychiatry, 2Ed*. *Principles and Practice of Forensic Psychiatry, 2Ed* (2003). doi:10.1201/b13499.

3. Prosono, M. History of forensic psychiatry. in *Principles and Practice of Forensic Psychiatry, 2Ed* (2003). doi:10.1201/b13499-5.

4. Gutheil, T. G. The history of forensic psychiatry. *Journal of the American Academy of Psychiatry and the Law* (2005).

5. Buchanan, A. & Grounds, A. Forensic psychiatry and public protection. *British Journal of Psychiatry* (2011) doi:10.1192/bjp.bp.111.095471.

6. Zitrin, A., Hardesty, A. S., Burdock, E. I. & Drossman, A. K. Crime and violence among mental patients. *Am. J. Psychiatry* (1976) doi:10.1176/ajp.133.2.142.

7. Klassen, D. & O'Connor, W. A. Crime, inpatient admissions, and violence among male mental patients. *Int. J. Law Psychiatry* (1988) doi:10.1016/0160-2527(88)90017-9.

8. Fazel, S. & Seewald, K. Severe mental illness in 33 588 prisoners worldwide: Systematic review and meta-regression analysis. *British Journal of Psychiatry*

(2012) doi:10.1192/bjp.bp.111.096370.

9. Mullen, P. E., Burgess, P., Wallace, C., Palmer, S. & Ruschena, D. Community care and criminal offending in schizophrenia. *Lancet* (2000) doi:10.1016/S0140-6736(99)05082-5.

10. Simpson, A. I. F. *et al.* Homicide and mental illness in New Zealand, 1970-2000. *Br. J. Psychiatry* (2004) doi:10.1192/bjp.185.5.394.

11. Hodgins, S., Alderton, J., Cree, A., Aboud, A. & Mak, T. Aggressive behaviour, victimisation and crime among severely mentally ill patients requiring hospitalisation. *Br. J. Psychiatry* (2007) doi:10.1192/bjp.bp.106.06.029587.

12. Chang, Z., Larsson, H., Lichtenstein, P. & Fazel, S. Psychiatric disorders and violent reoffending: A national cohort study of convicted prisoners in Sweden. *The Lancet Psychiatry* (2015) doi:10.1016/S2215-0366(15)00234-5.

13. Chang, Z., Lichtenstein, P., Langström, N., Larsson, H. & Fazel, S. Association between prescription of major psychotropic medications and violent reoffending after prison release. *JAMA - J. Am. Med. Assoc.* (2016) doi:10.1001/jama.2016.15380.

14. Coid, J. W. *et al.* Psychiatric diagnosis and differential risks of offending following discharge. *Int. J. Law Psychiatry* (2015) doi:10.1016/j.ijlp.2015.01.009.

15. Coid, J., Mickey, N., Kahtan, N., Zhang, T. & Yang, M. Patients discharged from medium secure forensic psychiatry services: Reconvictions and risk factors. *Br. J. Psychiatry* (2007) doi:10.1192/bjp.bp.105.018788.

16. Moore, K. E. *et al.* Psychiatric disorders and crime in the US population: Results from the national epidemiologic survey on alcohol and related conditions wave III. *J. Clin. Psychiatry* (2019) doi:10.4088/JCP.18m12317.

17. Fazel, S. & Grann, M. The population impact of severe mental illness on violent crime. *Am. J. Psychiatry* (2006) doi:10.1176/ajp.2006.163.8.1397.

18. Clark, T. Review of Forensic psychiatry: Clinical, legal and ethical issues (2nd edn). *The British Journal of Psychiatry* (2015) doi:10.1192/bjp.bp.114.150524.

19. Murray, J. & Thomson, D. M. E. Clinical judgement in violence risk assessment. *Eur. J. Psychol.* (2010) doi:10.5964/ejop.v6i1.175.

20. Sreenivasan, S., Kirkish, P., Garrick, T., Weinberger, L. E. & Phenix, A. Actuarial risk assessment models: A review of critical issues related to violence and sex-offender recidivism assessments. *Journal of the American Academy of Psychiatry and the Law* (2000).

21. Hart, S. D., Michie, C. & Cooke, D. J. Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *Br. J. Psychiatry* (2007) doi:10.1192/bjp.190.5.s60.

22. Grann, M., Belfrage, H. & Tengström, A. Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Crim. Justice Behav.* (2000) doi:10.1177/0093854800027001006.

23. Rice, M. E., Harris, G. T. & Lang, C. Validation of and revision to the VRAG and SORAG: The violence risk appraisal guide-revised (VRAG-R). *Psychol. Assess.*

(2013) doi:10.1037/a0032878.

24. Harris, G. T., Rice, M. E. & Cormier, C. A. Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law and Human Behavior* (2002) doi:10.1023/A:1016347320889.

25. Hastings, M. E., Krishnan, S., Tangney, J. P. & Stuewig, J. Predictive and Incremental Validity of the Violence Risk Appraisal Guide Scores With Male and Female Jail Inmates. *Psychol. Assess.* (2011) doi:10.1037/a0021290.

26. Gray, N. S., Taylor, J. & Snowden, R. J. Predicting violent reconvictions using the HCR-20. *Br. J. Psychiatry* (2008) doi:10.1192/bjp.bp.107.044065.

27. Dahle, K. P. Strengths and limitations of actuarial prediction of criminal reoffence in a German prison sample: A comparative study of LSI-R, HCR-20 and PCL-R. *Int. J. Law Psychiatry* (2006) doi:10.1016/j.ijlp.2006.03.001.

28. Thomson, L., Davidson, M., Brett, C., Steele, J. & Darjee, R. Risk assessment in forensic patients with schizophrenia: The predictive validity of actuarial scales and symptom severity for offending and violence over 8 – 10 years. *Int. J. Forensic Ment. Health* (2008) doi:10.1080/14999013.2008.9914413.

29. Langton, C. M. *et al.* Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the static-2002 and five other instruments. *Crim. Justice Behav.* (2007) doi:10.1177/0093854806291157.

30. Harris, G. T. & Rice, M. E. Characterizing the value of actuarial violence risk assessments. *Crim. Justice Behav.* (2007) doi:10.1177/0093854807307029.

31. Tully, R. J., Chou, S. & Browne, K. D. A systematic review on the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders. *Clinical Psychology Review* (2013) doi:10.1016/j.cpr.2012.12.002.

32. Rossegger, A., Gerth, J., Singh, J. & Endrass, J. Examining the Predictive Validity of the SORAG in Switzerland. *Sex. Offender Treat.* (2013).

33. Rettenberger, M., Rice, M. E., Harris, G. T. & Eher, R. Actuarial risk assessment of sexual offenders: The psychometric properties of the sex offender risk appraisal guide (SORAG). *Psychol. Assess.* (2017) doi:10.1037/pas0000390.

34. Cooke, D. J. & Michie, C. Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law Hum. Behav.* (2010) doi:10.1007/s10979-009-9176-x.

35. D., T. *et al.* Distinguishing and combining risks for sexual and violent recidivism. *Ann. N. Y. Acad. Sci.* (2003).

36. Kingston, D. A., Yates, P. M., Firestone, P., Babchishin, K. & Bradford, J. M. Long-term predictive validity of the risk matrix 2000: A comparison with the static-99 and the sex offender risk appraisal guide. *Sex. Abus. J. Res. Treat.* (2008) doi:10.1177/1079063208325206.

37. Helmus, L. M. & Babchishin, K. M. Primer on Risk Assessment and the Statistics Used to Evaluate Its Accuracy. *Crim. Justice Behav.* (2017) doi:10.1177/0093854816678898.

38. Parent, G., Guay, J. P. & Knight, R. A. An assessment of long-term risk of

recidivism by adult sex offenders: One size doesn't fit all. *Crim. Justice Behav.* (2011) doi:10.1177/0093854810388238.

39. Hanson, R. K. & Thornton, D. *Static 99: Improving actuarial risk assessments for sex offenders*. *The Department of the Solicitor General and Her Majesty's Prison Service, London* (1999).

40. Trevethan, R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front. Public Heal.* (2017) doi:10.3389/fpubh.2017.00307.

41. Mossman, D. Evaluating Risk Assessments Using Receiver Operating Characteristic Analysis: Rationale, Advantages, Insights, and Limitations. *Behav. Sci. Law* (2013) doi:10.1002/bsl.2050.

42. Hanson, R. K., Thornton, D. & Center, S. *Notes on the Development of the STATIC-2002*. *Department of Health and Family Services, WI* (2003).

43. Litwack, T. R. Actuarial versus clinical assessments of dangerousness. *Psychol. Public Policy, Law* (2001) doi:10.1037/1076-8971.7.2.409.

44. Cook, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* (2007) doi:10.1161/CIRCULATIONAHA.106.672402.

45. Perkins, N. J. & Schisterman, E. F. The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* (2006) doi:10.1093/aje/kwj063.

46. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test

assessment. *J. Thorac. Oncol.* (2010) doi:10.1097/JTO.0b013e3181ec173d.

47. McNeil, B. J. & Hanley, J. A. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Med. Decis. Mak.* (1984) doi:10.1177/0272989X8400400203.

48. Mossman, D. Assessing Predictions of Violence: Being Accurate About Accuracy. *J. Consult. Clin. Psychol.* (1994) doi:10.1037/0022-006X.62.4.783.

49. Cortes, C. & Mohri, M. Confidence intervals for the area under the ROC Curve. in *Advances in Neural Information Processing Systems* (2005).

50. Zhu, W., Zeng, N. & Wang, N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *Northeast SAS Users Gr. 2010 Heal. Care Life Sci.* (2010).

51. Kerekes, J. Receiver operating characteristic curve confidence intervals and regions. *IEEE Geosci. Remote Sens. Lett.* (2008) doi:10.1109/LGRS.2008.915928.

52. Osborne, J. & Thompson, B. Computing and Interpreting Effect Sizes, Confidence Intervals, and Confidence Intervals for Effect Sizes. in *Best Practices in Quantitative Methods* (2011). doi:10.4135/9781412995627.d21.

53. Altman, D. G. & Bland, J. M. Statistics Notes: Diagnostic tests 1: Sensitivity and specificity. *BMJ* (1994) doi:10.1136/bmj.308.6943.1552.

54. Rogers, R. The uncritical acceptance of risk assessment in forensic practice. *Law and Human Behavior* (2000) doi:10.1023/A:1005575113507.

55. Saks, M. J. & Koehler, J. J. The coming paradigm shift in forensic identification science. *Science* (2005) doi:10.1126/science.1111565.

56. Tversky, A. & Kahneman, D. Evidential impact of base rates. in *Judgment under Uncertainty* (2013). doi:10.1017/cbo9780511809477.011.

57. Deo, R. C. Machine learning in medicine. *Circulation* (2015) doi:10.1161/CIRCULATIONAHA.115.001593.

58. Passos, I. C., Mwangi, B. & Kapczinski, F. Big data analytics and machine learning: 2015 and beyond. *The Lancet Psychiatry* (2016) doi:10.1016/S2215-0366(15)00549-0.

59. Passos, I. C. & Mwangi, B. Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. *Mol. Psychiatry* (2018) doi:10.1038/s41380-018-0250-y.

60. Khoury, M. J. & Ioannidis, J. P. A. Big data meets public health. *Science* (2014) doi:10.1126/science.aaa2709.

61. Sartori, G., Pellegrini, S. & Mechelli, A. Forensic neurosciences: From basic research to applications and pitfalls. *Curr. Opin. Neurol.* (2011) doi:10.1097/WCO.0b013e3283489754.

62. Haenssle, H. A. *et al.* Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* (2018)

doi:10.1093/annonc/mdy166.

63. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* (2018) doi:10.1038/nature26000.

64. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* (2019) doi:10.1038/s41591-019-0447-x.

65. Havugimana, P. C., Hu, P. & Emili, A. Protein complexes, big data, machine learning and integrative proteomics: lessons learned over a decade of systematic analysis of protein interaction networks. *Expert Rev. Proteomics* (2017) doi:10.1080/14789450.2017.1374179.

66. Frommholz, I. *et al.* On Textual Analysis and Machine Learning for Cyberstalking Detection. *Datenbank-Spektrum* (2016) doi:10.1007/s13222-016-0221-x.

67. Moessner, M., Feldhege, J., Wolf, M. & Bauer, S. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *Int. J. Eat. Disord.* (2018) doi:10.1002/eat.22878.

68. Jaworska, N., De La Salle, S., Ibrahim, M. H., Blier, P. & Knott, V. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. *Front. Psychiatry* (2019) doi:10.3389/fpsyt.2018.00768.

69. McNeil, A. J., Frey, R. & Embrechts, P. *Quantitative risk management: Concepts, techniques, and tools*. *Quantitative Risk Management: Concepts, Techniques,*

and Tools (2005). doi:10.1198/jasa.2006.s156.

70. Jed Frees, E. W., Derrig, R. A. & Meyers, G. Predictive modeling in actuarial science. in *Predictive Modeling Applications in Actuarial Science: Volume I: Predictive Modeling Techniques* (2014). doi:10.1017/CBO9781139342674.001.

71. Gerds, T. A., Cai, T. & Schumacher, M. The performance of risk prediction models. *Biometrical Journal* (2008) doi:10.1002/bimj.200810443.

72. Bennett, K. P. & Parrado-Hernández, E. The interplay of optimization and machine learning research. *J. Mach. Learn. Res.* (2006).

73. Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. 1–13 (2017).

74. Cheng, T. H., Wei, C. P. & Tseng, V. S. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. in *Proceedings - IEEE Symposium on Computer-Based Medical Systems* (2006). doi:10.1109/CBMS.2006.87.

75. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. Feature selection for high-dimensional data. *Prog. Artif. Intell.* (2016) doi:10.1007/s13748-015-0080-y.

76. Eher, R., Rettenberger, M., Matthes, A. & Boer, D. P. Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *Int. J. Offender Ther. Comp. Criminol.* (2010) doi:10.1177/0306624X08328755.

77. Katuwal, G. J. & Chen, R. Machine Learning Model Interpretability for Precision Medicine. (2016).

78. Burrell, J. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data Soc.* (2016) doi:10.1177/2053951715622512.

79. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* (2019) doi:10.1038/s42256-019-0048-x.

80. Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* (1986) doi:10.1023/A:1022643204877.

81. Zorman, M., Štiglic, M. M., Kokol, P. & Malčić, I. The limitations of decision trees and automatic learning in real world medical decision making. *J. Med. Syst.* (1997) doi:10.1023/A:1022876330390.

82. Kokol, P., Zorman, M., Štiglic, M. M. & Malèiæ, I. The limitations of decision trees and automatic learning in real world medical decision making. in *Studies in Health Technology and Informatics* (1998). doi:10.3233/978-1-60750-896-0-529.

83. Dietterich, T. G. Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach. Learn.* (2000) doi:10.1023/A:1007607513941.

84. Barredo Arrieta, A. *et al.* Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).

85. Samek, W. & Müller, K. R. Towards Explainable Artificial Intelligence. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2019). doi:10.1007/978-3-030-28954-6_1.

86. Handelman, G. S. *et al.* Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *American Journal of Roentgenology* (2019) doi:10.2214/AJR.18.20224.

87. Wei, P., Lu, Z. & Song, J. Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety* (2015) doi:10.1016/j.ress.2015.05.018.

88. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* (2008) doi:10.1186/1471-2105-9-307.

89. Williamson, B. D., Gilbert, P. B., Simon, N. & Carone, M. Nonparametric variable importance assessment using machine learning techniques. *UW Biostat. Work. Pap. Ser.* (2017).

90. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* (2019) doi:10.1007/s00521-019-04051-w.

91. Fan, J. & Li, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. in *International Congress of Mathematicians, ICM 2006* (2006). doi:10.4171/022-3/31.

92. Cai, J., Luo, J., Wang, S. & Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* (2018) doi:10.1016/j.neucom.2017.11.077.

93. Van Der Maaten, L. J. P., Postma, E. O. & Van Den Herik, H. J. Dimensionality Reduction: A Comparative Review. *J. Mach. Learn. Res.* (2009) doi:10.1080/13506280444000102.

94. Zhang, D., Zhou, Z. H. & Chen, S. Semi-supervised dimensionality reduction. in *Proceedings of the 7th SIAM International Conference on Data Mining* (2007). doi:10.1137/1.9781611972771.73.

95. Sugiyama, M., Idé, T., Nakajima, S. & Sese, J. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Mach. Learn.* (2010) doi:10.1007/s10994-009-5125-7.

96. Klppel, S. *et al.* A plea for confidence intervals and consideration of generalizability in diagnostic studies. *Brain* (2009) doi:10.1093/brain/awn091.

97. Zheng, A. & Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. *O'Reilly* (2018).

98. Tharwat, A. Classification assessment methods. *Appl. Comput. Informatics* (2018) doi:10.1016/j.aci.2018.08.003.

99. Schaffer, C. Selecting a classification method by cross-validation. *Mach. Learn.* (1993) doi:10.1007/bf00993106.

100. Claesen, M. & De Moor, B. Hyperparameter Search in Machine Learning. 10–14 (2015).

101. Moore, A. W. & Lee, M. S. Efficient Algorithms for Minimizing Cross Validation Error. in *Machine Learning Proceedings 1994* (1994). doi:10.1016/b978-1-55860-335-6.50031-3.

102. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* (2011) doi:10.1007/s11222-009-9153-8.

103. Caulkins, J., Cohen, J., Gorr, W. & Wei, J. Predicting criminal recidivism: A comparison of neural network models. *J. Crim. Justice* (1996).

104. Palocsay, S. W., Wang, P. & Brookshire, R. G. Predicting criminal recidivism using neutral networks. *Socioecon. Plann. Sci.* (2000) doi:10.1016/S0038-0121(00)00003-3.

105. Pflueger, M. O., Franke, I., Graf, M. & Hachtel, H. Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. *BMC Psychiatry* (2015) doi:10.1186/s12888-015-0447-4.

106. Hanson, R. K. Does Static-99 predict recidivism among older sexual offenders? *Sex. Abus. J. Res. Treat.* (2006) doi:10.1007/s11194-006-9027-y.

CHAPTER 2

Risk Prediction in Forensic Psychiatry – A Path Forward

Authors: Devon Watts[1,2], Heather Moulden[1], Mini Mamak[1], Casey Upfold[1], Flávio Kapczinski[1,2,3]; Gary Chaimowitz[1]

1. Department of Psychiatry and Behavioral Neurosciences, McMaster University, Hamilton, Canada.
2. Neuroscience Graduate Program, McMaster University, Hamilton, Canada
3. Instituto Nacional de Ciência e Tecnologia Translacional em Medicina (INCT-TM), Porto Alegre, Brazil.

*Corresponding author:

Gary Chaimowitz, MB, CHB, MBA, FRCP(C)

Professor of Psychiatry

Psychiatry & Behavioural Neurosciences, McMaster University

Head of Service, Forensic Psychiatry Program, St. Joseph's Healthcare

100 West 5th Street, Hamilton, Ontario, L9C 0E3, Canada

Phone: 905-522-1155 ext 35424

Email: chaimow@mcmaster.ca

Email for all authors:

Devon Watts: wattsd@mcmaster.ca

Heather Moulden: hmoulden@stjosham.on.ca

Mini Mamak: mmamak@stjoes.ca

Casey Upfold: cupfold@stjosham.on.ca

Flavio Kapczinski: kapczinf@mcmaster.ca

Gary Chaimowitz: chaimow@mcmaster.ca

Background: Actuarial risk estimates are considered the gold-standard way to assess whether forensic psychiatry patients are likely to commit prospective criminal offences. However, these risk estimates cannot individually predict the type of criminal offence a patient will subsequently commit, and often simply assess the general likelihood of crime occurring in a group sample. In order to advance the predictive utility of risk assessments, better statistical strategies are required.

Aim: To develop a machine learning model to predict the type of criminal offense committed in forensic psychiatry patients, at an individual level.

Method: Machine learning algorithms (Random Forest, Elastic Net, SVM), were applied to a representative and diverse sample of 1240 patients in the forensic mental health system. Clinical, historical, and sociodemographic variables were considered as potential predictors and assessed in a data-driven way. Separate models were created for each type of criminal offence, and feature selection methods were used to improve the interpretability and generalizability of our findings.

Results: Sexual and violent crimes can be predicted at an individual level with 83.26% sensitivity and 77.42% specificity using only 20 clinical variables. Likewise, nonviolent and sexual crimes can be individually predicted with 74.60% sensitivity and 80.65% specificity using 30 clinical variables.

Conclusion: The current results suggest that machine learning models have accuracy comparable to existing risk assessment tools (AUCs .70-.80). However, unlike existing risk tools, this approach allows for the prediction of cases at an individual level, which is more clinically useful. The accuracy of prospective models is expected to only improve with further refinement.

Keywords:

machine learning; big data; computational psychiatry; psychiatric disorders; predictive psychiatry; precision medicine; predictive analysis; pattern recognition; forensic psychiatry; violence; violent behavior; crime; criminality

Manuscript submitted for publication to: British Journal of Psychiatry

Introduction:

Predictors of criminal risk: trials and tribulations

Prior to the development of any standardized tools, clinical judgement was the gold-standard measure to assess a patient's prospective risk of criminal reoffending (1). However, this presented a number of clear limitations, including poor inter-rater reliability between clinicians, confirmation bias, and the propensity for human error (2). Importantly, clinical judgement alone has not provided a valid metric by which to identify individuals with mental illness who will prospectively commit serious criminal offenses (3). In response to this, actuarial risk estimates became increasingly widespread, which concentrated on statistical models, while largely disavowing clinical judgement (4). Broadly speaking, risk estimates attempt to quantify the probability that an event will occur in the future (5). As discussed elsewhere, these risk estimates have demonstrated moderate to high predictive validity in quantifying group-based risk of

of general recidivism(6), as well as violent(7) and sexual recidivism(8).

The ethics of predictors: a question of fairness

Nonetheless, concerns have been raised as to the specific variables used in risk estimates within forensic psychiatry (2). In general, these tools place a strong emphasis on static risk factors, which are patient characteristics largely unamenable to change, such as prior offenses, and childhood experiences (9). This, as a consequence, discounts modifiable risk factors that may be of importance, such as drug abuse, poverty and housing instability(10). Therefore, by overemphasizing static risk factors, we run the risk of stigmatizing patients based on their past, without identifying new strategies to improve rehabilitation efforts and thus decrease prospective recidivism.

Statistical challenges in group-based risk assessment

Although risk estimates have helped shift the focus of risk assessment in forensic psychiatry toward a reproducible and statistical framework, there are important caveats to their use. Namely, there is little evidence that actuary risk estimates perform any better than clinical judgement in determining whether a specific patient will reoffend (11). This is largely because most risk estimates have been developed statistically to assess group-based risk, and perform poorly when making individualized predictions (12).

This phenomenon is partly related to the difference in calculating 95% Confidence Intervals (CIs) between a group effect and an individualized prediction, with the latter showing higher variability (12). Of note, this phenomenon of high variability in prediction intervals occurs largely independent of sample size consideration (12). This is because while increasing sample size can result in a more narrow confidence interval around a regression line in a group-based analysis, this does not translate into a more narrow prediction interval in an individual case (12). While it has been argued that such risk estimates should be readily interpretable and easy to score by clinicians, it is unlikely that complex phenomena such as predicting criminal recidivism at an individual level can be appropriately modeled using simple statistical approaches(5). Altogether this highlights the need for caution in using actuarial risk estimates to determine prospective clinical risk and the necessity for new approaches.

The age of artificial intelligence. A path forward?

Given the ethical, psychiatric and legal ramifications of mischaracterizing the prospective risk of any given patient, and the resulting consequences to the individual, their families, and broader society, there is a growing interest in the use of artificial intelligence and predictive analytics to improve accuracy in clinical decision making. A possible solution to these challenges may lie in the use of machine learning, which broadly speaking, focuses on extracting value from datasets using computational algorithms (13). These algorithms can detect patterns within a dataset and then apply what they learned to make predictions in new, unseen data (14). Unlike traditional statistical approaches that evaluate average differences in outcomes between groups, machine learning methods provide a more straightforward way to predict outcomes at an individual level (15). This can potentially pave the way for tailor-made tools for the diagnosis, assessment, and treatment of patients(16,17). Although these techniques have shown promise in other fields of science and medicine, their value in forensic psychiatry has yet to be fully explored.

The urgent need for individualized risk assessment

While few studies have assessed the risk of subsequent violence(18,19) and criminal recidivism(20,21) among psychiatric patients, no studies thus far have focused on predicting the type of crime committed at an individual level. Knowing the type of crime an individual is likely to commit, before the offence occurs, is urgently needed in order to guide more targeted and precise risk assessment strategies and frontline therapeutic interventions(1). Furthermore, the vast majority of work thus far has focused on predicting recidivism in non-psychiatric prison populations (22–25). Importantly, it is largely unclear whether such models can be appropriately extrapolated to offences committed by those with severe mental illness.

In the present study, we used machine learning models to predict the type of criminal offense committed at an individual level. That was carried out in a representative sample of 1240 forensic inpatients from 10 psychiatric institutions, found unfit to stand trial (UST), or not criminally responsible (NCR).

2. Methods

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects/patients were approved by the Hamilton Integrated Research Ethics

Board #0538 and # 10529, in affiliation with McMaster University. Every participant in the sample was found not criminally responsible or unfit to stand trial for criminal offenses, between 2014-2015. Explicit written consent was not required from participants for ethical approval of the study, given the inherent challenges of obtaining consent from a large sample of patients from multiple psychiatric facilities. However, patient data was anonymized with digital identifiers removed, in line with ethical standards.

2.1. Study population

The present study consisted of 1240 individuals charged with a criminal offense, and subsequently deemed either Unfit to Stand Trial (UST) or Not Criminally Responsible (NCR) as a result of serious mental illness. That comprised a diverse sample of patients from 10 forensic psychiatry facilities, representing patients who were subject to oversight by the Ontario Review Board (ORB) between 2014-2015. The ORB is an independent tribunal established under the Criminal Code of Canada that reviews the status of every person who has been found to be NCR or UST for criminal offences on account of a mental disorder.

2.2. Variables

The ORB database comprised 1100 clinical, demographic, administrative and behavioral variables. Among them, 246 variables were extracted from ORB files. This involved a diverse set of factors, such as adverse events in childhood, income, housing, comorbidities, family history, prescribed medications, substance use, and presumed indicators of risk.

2.3. Feature selection

Within machine learning applications, feature selection provides an important way to reduce the dimensionality of a dataset by removing irrelevant features (26). There are several available feature selection methods, with varying degrees of appropriateness depending on the application, as described elsewhere (27). Of note, limiting the number of features tends to improve the generalizability of a model when applied in independent datasets (28). In the present study, a data-driven approach to feature selection was used. This encompasses a series of feature selection methods that do not rely on preconceived notions as to which variables will be the most important in the model (29). Specifically, three methods were compared. This included Recursive Feature Elimination (RFE), Ensemble Feature Selection (EFS), and selecting the top 20 weighting factors using variable importance plots.

Briefly, RFE is a method that removes features that have the least impact on training error. This tends to remove redundant features, while retaining independent features(30). However, each feature selection method is prone to a set of biases. Considering this, EFS was used, which comprises eight feature selection methods, and linearly combines their normalized outputs to derive a quantitative feature importance metric(31). Furthermore, a straightforward and visually interpretable feature selection method was used as a point of comparison, where a subset of the data with all predictors were used to build a model, and the top 20 variables were identified using a variable importance plot. These 20 variables were then used as the sole predictors within the respective model.

Importantly, only variables occurring prior to the index offense were considered as potential predictors. Variables with 15% or more missing data were excluded, considering both the impact of missing values in model performance, and the limitations of available imputation strategies (32,33). Imputation was performed using the mean and mode, for numerical and categorical variables, respectively.

2.4. Machine Learning Algorithms

Three machine learning algorithms (Elastic Net, Support Vector Machines (SVM) and Random Forest) were implemented in R using various packages (34–36). Predictor variables were centered and scaled using the preProcess function available in Caret (37). Zero and near-zero variance predictors were removed using the nearZeroVar function available in Caret (37). All categorical variables were transformed into dichotomous, quantitative variables, colloquially known as dummy variables. A thorough explanation as to the strengths and technicalities of these machine learning algorithms can be found elsewhere (38–40).

Fundamentally, elastic net is a regularized method of logistic regression that linearly combines the L1 and L2 penalties of lasso and ridge methods (38). Elastic net is both computationally efficient, and well suited to cases of highly correlated predictors (38). Random Forest is an ensemble method that builds a number of decision trees, with each node split using the best of a subset of randomly chosen predictors (36). By averaging a set of observations using random sampling when building trees and nodes, random forest notably reduces variance, and tends to perform very well in classification problems (40). SVM is an extension of the maximal margin classifier that can accommodate non-linear class boundaries (40). SVM attempts to find a maximal margin hyperplane that directly depends on a series of support vectors, rather than relying on all observations in the dataset (39). SVM does not require a hyperplane that perfectly separates hyperplanes, and allows some observations to be incorrect, in the interest of better classifying observations overall (39). This algorithm is computationally efficient and allows for a number of potential non-linear boundaries between classes, by incorporating various possible kernel functions (41).

2.4. Addressing the class imbalance problem

A common challenge in machine learning classifiers is that of class imbalance (42,43). This occurs when the number of one class (e.g. nonviolent crimes) is far less than the total number of another class of data (e.g. violent crimes). In the current study, as detailed in Table 1, 863 patients were charged for violent crimes, 253 patients for non-violent crime, and 124 for sexual crimes, respectively. Three separate approaches were compared to address the imbalance between these classes. Namely, under-sampling of the majority class was used, which involves randomly eliminating elements of the majority class until it matches the size of the minority class (43). Conversely, the minority class was oversampled at random until it contained as many examples as the majority class (43). This was achieved in both instances by setting their respective arguments in the trainControl function in Caret (37). Furthermore, class weighting was used, which modifies the relative cost of misclassifying the majority and minority classes to compensate for their unbalanced ratio (43).

2.5. Model testing and validation

The ORB dataset was divided into training and testing sets, comprising 70% and 30% of the data, respectively. In order to estimate prediction error, 10-fold cross-validation was used, as described elsewhere (44). This involves a form of out-of-sample testing where data is partitioned into 10-folds, where a single subsample is retained as validation data to test the model, and the remaining *k*-1 folds are used as training data (44). This process was repeated 10 times, where the results were combined to produce a single estimation.

Another important consideration is hyperparameter optimization (45). Various parameters within machine learning models can be tuned to minimize a given loss function on independent data. This largely involves adjusting the learning rate of the model in order to improve model performance (45). In the present study, both grid and random search strategies were employed. In the present study, model performance was assessed using the confusionMatrix function in R (37). A confusion matrix is a table layout that provides an overview of model accuracy, misclassification rate, sensitivity, specificity, as well as true and false predictive values. This includes the number of correct and incorrect predictions, which are summarized with count values and broken down by each class. A more detailed explanation of confusion matrices, as well as how they differ in binary and multiclass prediction problems can be found elsewhere (46).

Results

The present study included a total of 1240 patients with mental illness who were found either Not Criminally Responsible, or Unfit to Stand Trial by the ORB between 2014-2015 for a criminal offense. This comprised 863 violent, 253 nonviolent, and 124 sexual offenses, respectively. All algorithms (Random Forest, SVM, and Elastic Net) were used to train binary classifiers. A summary of patient demographics is presented in Table 1.

Random Forest and Elastic Net identified patients who would subsequently commit sexual crime, from nonviolent and violent crimes, with predictive accuracy ranging from 61.5-80.3% in the total model. In particular, the Elastic Net algorithm correctly identified patients who would subsequently commit nonviolent or sexual crimes at an individual level, with a sensitivity of 71.4% and specificity of 70.9%. Furthermore, an Elastic Net algorithm correctly identified which patients would prospectively commit sexual or violent crimes with a sensitivity of 85.1% and specificity of 74.1% in the total model. A receiver operating characteristic (ROC) curve and ‘confusion matrix’ were used to calculate the sensitivity, specificity, balanced accuracy, AUC, and 95% confidence intervals of each of these models, as detailed in Tables 2 and 3.

Table 1. Summary of Demographic Variables

	Violent (n= 863)	**Non-violent (n=253)**	**Sexual (n=124)**	***p*-Value**
Age (years)	34.39±12.30	36.30 ±11.96	34.90 ±14.16	0.0995
Education				0.1715
Primary	91 (10.1%)	25 (9.8%)	19 (15.3%)	
Secondary	639 (70.6%)	180 (70.3%)	90 (72.6%)	
University	175 (19.3%)	51 (19.9%)	15 (12.1%)	
Gender				0.0035
Male	798 (86.5%)	209 (74.1%)	118 (95.9%)	
Female	124 (13.5%)	41 (14.5%)	5 (4.1%)	
Currently Employed				0.0862
Unemployed	534 (84.8%)	170 (89.4%)	71 (79.7%)	
Employed	95 (15.2%)	20 (10.6%)	18 (18.5%)	
Race				0.0442
Caucasian	156 (35.7%)	48 (39.1%)	23 (33.8%)	
Aboriginal	46 (10.2%)	15 (13.5%)	9 (11.7%)	
Black	116 (25.3%)	20 (16.5%)	17 (23.4%)	
Asian	60 (13.8%)	25 (18.8%)	18 (26%)	
Hispanic	17 (3.6%)	3 (2.3%)	2 (2.6%)	
Other	48 (11.3%)	13 (9.8%)	2 (2.6%)	
Marital Status				0.4586
Single	576 (70.6%)	166 (67.7%)	82 (69.4%)	
Married/Common Law	35 (4.2%)	7 (2.8%)	7 (5.9%)	
Other	204 (25.0%)	72 (29.3%)	29 (24.5%)	
History of Substance Abuse				0.5297
No history	234 (27.1%)	74 (29.6%)	37 (30.0%)	
Yes – alcohol and drugs	374 (43.3%)	104 (41.6%)	49 (39.8%)	
Yes – alcohol only	68 (7.8%)	26 (10.4%)	14 (50.9%)	
Diagnosis				<0.001
Schizophrenia	516 (59.8%)	135 (54.0%)	63 (51.2%)	
Schizoaffective	148 (17.1%)	62 (24.8%)	9 (7.3%)	
Delusional Disorder	25 (2.9%)	15 (6.0%)	4 (3.2%)	
Psychosis NOS	48 (5.5%)	11 (4.4%)	5 (4.1%)	
Bipolar Disorder	57 (6.6%)	20 (8.0%)	7 (5.6%)	
Paraphilia	19 (2.2%)	7 (2.8%)	34 (27.6%)	
Dementia/Cognitive Impairment	28 (3.2%)	8 (3.2%)	19 (15.4%)	

Chi-Square test with Yates correction used for categorical variables. One-way ANOVA used for numeric variables

Table 2: Model Performance: Nonviolent vs. Sexual Offences

	Elastic Net	Random Forest	Support Vector Machine (Radial Kernel)
Full Model (156 variables)	95% CI (61.02, 80.14) Specificity: 70.90 Sensitivity: 71.40 PPV: 83.30 NPV 55.00 Balanced Accuracy: 71.20	95% CI (61.02, 80.14) Sensitivity: 71.43 Specificity: 70.97 PPV: 83.33 NPV: 55.00 Balanced Accuracy: **71.20**	95% CI (47.88, 68.59) Sensitivity: 74.60 Specificity: 48.39 PPV: 74.60 NPV: 48.39 Balanced Accuracy: 61.50
Recursive Feature Elimination (29 variables)	95% CI (56.56, 76.38) Sensitivity: 69.84 Specificity: 61.29 PPV: 78.57 NPV: 50.00 Balanced Accuracy: 65.57	95% CI (65.58, 83.81) Sensitivity: 85.71 Specificity: 54.84 PPV: 79.41 NPV: 65.38 Balanced Accuracy: **70.28**	95% CI (56.56, 76.38) Sensitivity: 77.78 Specificity: 45.16 PPV: 74.24 NPV: 50.00 Balanced Accuracy: 61.47
Ensemble Feature Selection (30 variables)	95% CI (65.58, 83.81) Sensitivity: 79.37 Specificity: 67.74 PPV: 83.33 NPV: 61.76 Balanced Accuracy: 73.55	95% CI (66.74, 84.71) Sensitivity: 74.60 Specificity: 80.65 PPV: 80.65 NPV: 60.98 Balanced Accuracy: **77.62**	95% CI (40.54, 61.52) Sensitivity: 53.97 Specificity: 45.16 PPV: 66.67 NPV: 32.56 Balanced Accuracy: 49.56

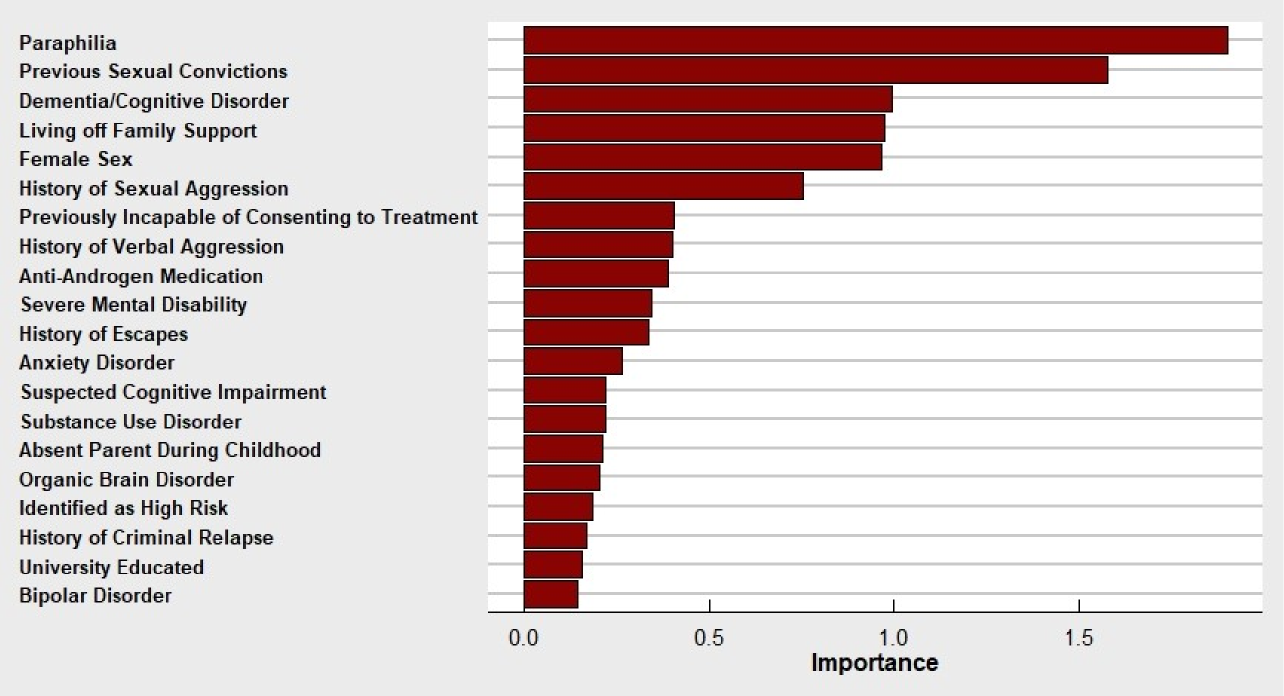
Table 3: Model Performance - Sexual vs. Violent Offences

	Elastic Net (Logistic Regression with L1 and L2 regularization)	Random Forest (Ensembles of decision trees)	Support Vector Machine (Radial Kernel)
Full Model (156 variables)	95% CI (78.52, 88.12) Sensitivity: 85.12 Specificity: 74.19 Pos Pred Value: 95.81 Neg Pred Value: 41.82 Balanced Accuracy: 79.65	95% CI (71.91, 82.69) Sensitivity:76.74 Specificity: 83.87 Pos Pred Value: 97.06 Neg Pred Value 34.21 **Balanced Accuracy: 80.31**	95% CI (58.31, 70.06) Sensitivity: 62.79 Specificity: 67.74 Pos Pred Value: 93.10 Neg Pred Value: 20.79 Balanced Accuracy: 65.27
Recursive Feature Elimination (45 variables)	95% CI (72.35, 83.06) Sensitivity: 80.47 Specificity: 61.29 Pos Pred Value: 93.51 Neg Pred Value: 31.15 **Balanced Accuracy: 70.88**	95% CI (68.03, 79.35) Sensitivity: 75.35 Specificity: 64.52 Pos Pred Value: 93.64 Neg Pred Value: 27.40 Balanced Accuracy: 69.93	95% CI (70.18, 81.21) Sensitivity: 78.14 Specificity: 61.29 Pos Pred Value: 93.33 Neg Pred Value: 28.79 Balanced Accuracy: 69.71
Ensemble Feature Selection (30 variables)	95% CI (76.74, 86.69) Sensitivity: 83.72 Specificity: 70.97 Pos Pred Value: 95.24 Neg Pred Value: 38.60 **Balanced Accuracy: 77.34**	95% CI (71.91, 82.69) Sensitivity: 79.53 Specificity: 64.52 Pos Pred Value: 93.96 Neg Pred Value: 31.25 Balanced Accuracy: 72.03	95% CI (76.3, 86.33) Sensitivity: 83.26 Specificity: 70.97 Pos Pred Value: 95.21 Neg Pred Value: 37.93 Balanced Accuracy: 77.11

The most relevant predictor variables to distinguish between sexual and violent criminal offences were identified with a variable importance plot, using the ‘caret’ package available in R, as detailed in Figure 1. Of note, relevant variables in the model included paraphilia, previous charges for indecent acts, current impulse control disorders, currently medicated for substance abuse, and previously deemed incapable of consenting to medication.

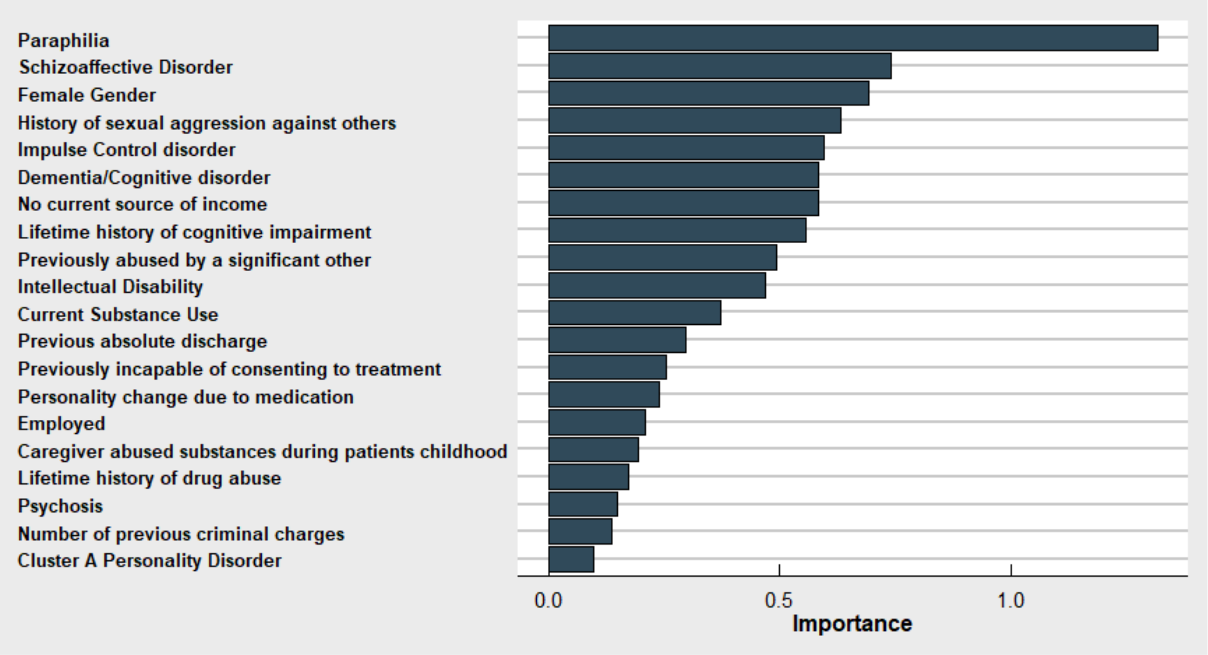
Similarly, the most relevant predictor variables to distinguish between sexual and nonviolent criminal offences were identified using a variable importance plot with the ‘rminer’ package in R, as detailed in Figure 2. The most important variables to differentiate between sexual and nonviolent crimes at an individual level included schizoaffective disorder, personality disorders characterized by odd, eccentric thinking, or behaviour, as well as a caregiver with mental illness during childhood, and the presence of dementia or cognitive disorders. Importantly, in predicting sexual, violent and nonviolent crime, model accuracy was largely preserved following various feature selection methods, as detailed in Tables 2 and 3.

Figure 1: Variable Importance - Sexual vs. Violent Offences



A visual depiction of important features for the model. A variable importance plot was generated using the VarImp() function available in the Caret Package in R Studio.

Figure 2: Variable Importance - Sexual vs. Nonviolent Offences



A visual depiction of important features for the model. A variable importance plot was generated using the VarImp() function available in the Caret Package in R Studio.

DISCUSSION

Available risk assessment tools in forensic psychiatry perform poorly when making individualized predictions. The current study is the first to demonstrate that violent and sexual crimes can be predicted at an individual level based on clinical and demographic variables occurring prior to the event. Moreover, we were able to predict whether an individual will commit a sexual or nonviolent offense. This was found in a representative sample of forensic patients, both with and without psychosis. Model performance was also largely preserved after drastically reducing the number of predictor variables, which aid in the generalizability and replicability of these findings. However, our models were not able to distinguish violent from nonviolent crimes using the same data and preprocessing pipeline. While the clinical implications of these findings need to be refined with prospective studies, the possibility of using machine learning to predict crime related behaviours is clearly demonstrated in this paper.

Here, static and malleable risk factors, alongside other clinical and demographic variables, were assessed in a data-driven way to determine their relative importance in predicting the type of crime committed. This represents a significant departure from prior methodologies(22–25), which largely involved selecting variables using domain knowledge. This data-driven approach to feature selection, however, can potentially identify novel variables that are salient, but unexpected. For instance, in predicting nonviolent and sexual crimes, impulse control disorders, and the absence of financial income were much more important than commonly cited risk factors(47) such as the number of prior criminal charges, or the presence of childhood abuse. Although static risk factors were relevant in all machine learning models, our results suggest that there are important malleable risk factors that may be useful targets to improve patient rehabilitation and decrease criminal recidivism.

As mentioned previously, a major thrust of machine learning models in recidivism prediction have focused on non-psychiatric prison populations, with no studies thus far predicting the type of crime at an individual level. Therefore, the present study represents an initial effort to predict likely offenses before they occur and move the field toward individualized risk assessment. Based on our results, it has been demonstrated that it is possible to make such individualized predictions with a reasonable degree of accuracy. However, it is interesting to note that our model was unable to differentiate non-violent from violent crime. This was found even after running separate models based on biological sex, for men (n=1065), and women (n=170), respectively. While this could occur for a variety of reasons, it suggests that those who commit sexual crimes may represent a distinct clinical profile that was adequately modeled using our algorithms. Conversely, individuals who commit violent and nonviolent crimes may represent a similar profile, that may not be adequately differentiated using clinical and demographic information alone.

With respect to algorithm performance, Elastic Net and Random Forest tended to show similar AUC and accuracy. Although both algorithms performed very similar in the full model, Random Forest showed a notable improvement over Elastic Net following RFE and EFS feature selection methods. As such, Random Forest presented with a preferable balance between sensitivity and specificity. This is an important consideration when dealing with prognostications that have real life consequences, such as crime prediction. Moreover, in all models, SVM showed a substantially poorer performance relative to random forest or elastic net. Since a favourable linear decision boundary was observed with Elastic Net, it is possible that this decrease in performance with SVM was due to the radial kernel used, which creates a nonlinear decision boundary. Moreover, altering cost and gamma hyperparameters using a grid search did not significantly improve performance.

In summary, the current results suggest that machine learning models have accuracy comparable to existing risk assessment tools (AUCs .70-.80). However, unlike existing risk tools, this approach allows for the prediction of cases at an individual level, which is more clinically useful. Moreover, this represents a first attempt to predict the type of crime an individual will commit, using variables occurring prior to the offense. The accuracy of prospective models is expected to only improve with further refinement.

Limitations

The current study has some potential limitations. While a representative sample of 1240 forensic inpatients were used from 10 psychiatric institutions across Ontario, Canada, this may not be representative of patients in other countries, or jurisdictions. Furthermore, while reasonable performance accuracy was observed in the present study, further refinement of risk prediction models is needed. Similarly, a much smaller error rate is required to implement such predictive models as clinical tools. Additionally, variables with missing data were excluded from the analysis. While some of these excluded variables may have proven to be useful, increasing the imputation threshold from 15% to 30% did not result in a significant change in accuracy in any of the models. Likewise, other imputation strategies, such as *k*-nearest neighbours(48), may be a useful alternative in the case of missing data. However, it is important to note that each imputation strategy has its own set of limitations(32,33). Apart from this, the current study used binary classifiers to distinguish between types of crime at an individual level. Other studies may benefit from using one-vs-one and one-vs-rest classifiers (49). Also, other algorithms, and preprocessing pipelines may lead to different performance metrics.

Perspectives

Moving forward, a further refinement of predictive models in forensic risk prediction is required. Potentially, this may be facilitated by using a wider framework when selecting the input data in our models. Considering that our model performance is directly dependent on the available input data, an exploratory data-driven approach may be warranted in risk prediction studies. Moreover, unstructured data such as neuroimaging and neurophysiology, may prove useful to facilitate model performance, when used in combination with structured clinical data.

References

1. Arboleda-Flórez J. Forensic psychiatry: contemporary scope, challenges and controversies. World Psychiatry. 2006;

2. Clark T. Review of Forensic psychiatry: Clinical, legal and ethical issues (2nd edn). The British Journal of Psychiatry. 2015.

3. Murray J, Thomson DME. Clinical judgement in violence risk assessment. Eur J Psychol. 2010;

4. Sreenivasan S, Kirkish P, Garrick T, Weinberger LE, Phenix A. Actuarial risk assessment models: A review of critical issues related to violence and sex-offender recidivism assessments. Journal of the American Academy of Psychiatry and the Law. 2000.

5. Hart SD, Michie C, Cooke DJ. Precision of actuarial risk assessment instruments: Evaluating the "margins of error" of group v. individual predictions of violence. Br J Psychiatry. 2007;

6. Dahle KP. Strengths and limitations of actuarial prediction of criminal reoffence in a German prison sample: A comparative study of LSI-R, HCR-20 and PCL-R. Int J Law Psychiatry. 2006;

7. Gray NS, Taylor J, Snowden RJ. Predicting violent reconvictions using the HCR-20. Br J Psychiatry. 2008;

8. Tully RJ, Chou S, Browne KD. A systematic review on the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders. Clinical Psychology Review. 2013.

9. Rogers R. The uncritical acceptance of risk assessment in forensic practice. Law and Human Behavior. 2000.

10. Lindqvist P, Skipworth J. Evidence-based rehabilitation in forensic psychiatry. Br J Psychiatry. 2000;

11. Litwack TR. Actuarial versus clinical assessments of dangerousness. Psychol Public Policy, Law. 2001;

12. Cooke DJ, Michie C. Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. Law Hum Behav. 2010;

13. Alpaydin E. Introduction to Machine Learning Ethem Alpaydin. Introd to Mach Learn Third Ed. 2014;

14. Deo RC. Machine learning in medicine. Circulation. 2015;

15. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA - Journal of the American Medical Association. 2018.

16. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. The Lancet Psychiatry. 2016.

17. Passos IC, Mwangi B. Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. Mol Psychiatry [Internet]. 2018; Available from: http://www.nature.com/articles/s41380-018-0250-y

18. Linaker OM, Busch-Iversen H. Predictors of imminent violence in psychiatric inpatients. Acta Psychiatr Scand. 1995;

19. Liu YY, Yang M, Ramsay M, Li XS, Coid JW. A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending. J Quant Criminol. 2011;

20. Caulkins J, Cohen J, Gorr W, Wei J. Predicting criminal recidivism: A comparison of neural network models. J Crim Justice. 1996;

21. Pflueger MO, Franke I, Graf M, Hachtel H. Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. BMC Psychiatry. 2015;

22. Palocsay SW, Wang P, Brookshire RG. Predicting criminal recidivism using neutral networks. Socioecon Plann Sci. 2000;

23. Veronezi BP, Moffa AH, Carvalho AF, Galhardoni R, Simis M, Benseñor IM, et al. Evidence for increased motor cortical facilitation and decreased inhibition in atypical depression. Acta Psychiatr Scand. 2016;

24. Cohen MI, Spodak MK, Silver SB, Williams K. Predicting outcome of insanity

acquittees released to the community. Behav Sci Law. 1988;

25. Silver E, Smith WR, Banks S. Constructing actuarial devices for predicting recidivism a comparison of methods. Crim Justice Behav. 2000;

26. Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data. In 2010.

27. Van Der Maaten LJP, Postma EO, Van Den Herik HJ. Dimensionality Reduction: A Comparative Review. J Mach Learn Res. 2009;

28. Klppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. A plea for confidence intervals and consideration of generalizability in diagnostic studies. Brain. 2009.

29. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng. 2014;

30. Chen XW, Jeong JC. Enhanced recursive feature elimination. In: Proceedings - 6th International Conference on Machine Learning and Applications, ICMLA 2007. 2007.

31. Neumann U, Genze N, Heider D. EFS: An ensemble feature selection tool implemented as R-package and web-application. BioData Min. 2017;

32. Somasundaram RS, Nedunchezhian R. Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values. Int J Comput Appl. 2011;

33. Zhang S, Wu X, Zhu M. Efficient missing data imputation for supervised learning. In: Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010. 2010.

34. Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalized linear models. R Packag version. 2009;

35. Karatzoglou A, Meyer D, Hornik K. Support vector machines in R. J Stat Softw. 2006;

36. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;

37. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;

38. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;

39. Fletcher T. Support Vector Machines Explained. Online] http//sutikno blog undip ac id/files/2011/11/SVM-Explained pdf[Accessed 06 06 2013]. 2009;

40. Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012.

41. Noble WS. What is a support vector machine? Nature Biotechnology. 2006.

42. Japkowicz N. The Class Imbalance Problem: Significance and Strategies. Proc 2000 Int Conf Artif Intell. 2000;

43. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal. 2002;

44. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput. 2011;

45. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;

46. Tharwat A. Classification assessment methods. Appl Comput Informatics. 2018;

47. Coid J, Mickey N, Kahtan N, Zhang T, Yang M. Patients discharged from medium secure forensic psychiatry services: Reconvictions and risk factors. Br J Psychiatry. 2007;

48. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: A critical evaluation. BMC Med Inform Decis Mak. 2016;

49. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognit. 2011;

CHAPTER 3:

DISCUSSION

Available risk assessment tools in forensic psychiatry perform poorly when making individualized predictions[1]. The current study is the first to demonstrate that violent and sexual crimes can be predicted at an individual level based on clinical and demographic variables occurring prior to the event. Moreover, we were able to predict whether an individual will commit a sexual or nonviolent offense. This was found in a representative sample of forensic patients, both with and without psychosis. Model performance was also largely preserved after drastically reducing the number of predictor variables, which aid in the generalizability and replicability of these findings. However, our models were not able to distinguish violent from nonviolent crimes using the same data and preprocessing pipeline. While the clinical implications of these findings need to be refined with prospective studies, the possibility of using machine learning to predict crime related behaviours is clearly demonstrated in this paper.

Data-driven approach to feature selection

Here, static and malleable risk factors, alongside other clinical and demographic variables, were assessed in a data-driven way to determine their relative importance in predicting the type of crime committed. This represents a significant departure from prior methodologies[2–5], which largely involved selecting variables using domain knowledge. A data-driven approach to feature selection, however, can potentially identify novel variables that are salient, but unexpected. For instance, in predicting nonviolent and sexual crimes, impulse control disorders, and the absence of financial income were much more important than commonly cited risk factors[6] such as the number of prior criminal charges, or the presence of childhood abuse. Although static risk factors were relevant in all machine learning models, our results suggest that there are important malleable risk factors that may be useful targets to improve patient rehabilitation and decrease criminal recidivism.

Feature selection strategies

An important consideration in classification tasks in machine learning is feature selection, since irrelevant input features can impair model accuracy, and unnecessarily increase model complexity. As such, feature selection is a useful strategy to improve model generalizability [7]. Generally speaking, feature selection methods in classification tasks seek to identify a minimal number of features that does not result in a significant decrease in classification accuracy, while retaining the class distribution to be as close as possible to that observed in the full feature set [8]. Though several approaches exist, broadly speaking, they can be conceptualized as unsupervised, semi-supervised, and supervised approaches [9].

While unsupervised feature selection methods can work well with unlabeled data, it is inherently difficult to evaluate the relevance of features using these approaches [10]. Moreover, depending on how clustering performance is assessed, several equally valid feature subsets can be identified. Semi-supervised feature selection methods learn from a small number of labeled data, and a large number of unlabeled data [11]. Although they can be worthwhile in several applications, semi-and unsupervised feature selection methods were not used in the current work.

In the overarching category of supervised feature selection methods, exists filter, wrapper and embedded models [12]. Filter based models separate feature selection from classifier learning, so the bias of a learning algorithm does not interact with the bias of a feature selection algorithm. This relies on measures of the general characteristics of the training data, such as dependency, information, distance, consistency, and correlation [13]. Wrapper models use the predictive accuracy of a prespecified learning algorithm to determine the quality of selected features [14] . Embedded models, on the other hand, serve as a bridge between filter and wrapper models, by fitting a model and performing feature selection simultaneously [7]. This approach usually achieves comparable accuracy to the wrapper method, and comparable efficiency to the filter method [7].

In feature selection applications, the combination of individually important features does not necessarily translate into optimal classification performance [15]. Moreover, including a certain number of redundant features can help improve the robustness of a given predictor, which may involve a number of highly correlated variables [16]. One such strategy to address this is to simultaneously minimize the redundancy of irrelevant features, while maximizing the relevance of important ones. This leads to a compact subset of superior features at a low computational cost. Collectively, such approaches are referred to as mutual-information-based feature selection algorithms [17]. Among them, minimum-redundancy maximal-relevance (mRMR) is a popular feature selection method that adopts a greedy search to incrementally select a candidate feature set. As such, mRMR finds an optimal solution using a local minimum. Moreover, cross-validation is used to identify the classification error for a large number of features, and to find a relatively stable range of features with a small error [18]. Although mRMR was not used in the current work, it presents as a promising feature selection method in similar applications.

Additionally, Localized Feature Selection (LFS) offers an alternative to conventional feature selection methods that identify a shared feature set to characterize all cases in the sample space[19]. While other algorithms consider local sample behavior during feature selection, they require the entire sample space to be modeled by a common feature set. Instead, LFS considers each training sample as a point relative to its neighboring region and selects an optimal feature set for that region[19]. Essentially, this involves minimizing the local within class distances, and maximizing the between class distances [19]. Future work may serve to benefit from the use of more sophisticated methods of feature selection, such as mRMR and LFS. However, this is by no means an exhaustive list, and the appropriate feature selection method may vary as a function of the specific dataset and task.

Predictive models in forensic psychiatry

As mentioned previously, a major thrust of machine learning models in recidivism prediction have focused on non-psychiatric prison populations, with no studies thus far predicting the type of crime at an individual level. Therefore, the present study represents an initial effort to predict likely offenses before they occur and move the field toward individualized risk assessment. Based on our results, it has been demonstrated that it is possible to make such individualized predictions with a reasonable degree of accuracy. However, it is interesting to note that our model was unable to differentiate non-violent from violent crime. This was found even after running separate models based on biological sex, for men (n=1065), and women (n=170), respectively. While this could occur for a variety of reasons, it suggests that those who commit sexual crimes may represent a distinct clinical profile that was adequately modeled using our algorithms. Conversely, individuals who commit violent and nonviolent crimes may represent a similar profile, that may not be adequately differentiated using clinical and demographic information alone. It is possible that applying feature selection such as mRMR or LFS prior to training the violent and nonviolent binary classifier may improve performance.

Model performance and data modalities

With respect to algorithm performance, Elastic Net and Random Forest tended to show similar AUC and balanced accuracy. Although both algorithms performed very similar in the full model, Random Forest showed a notable improvement over Elastic Net following RFE and EFS feature selection methods. As such, Random Forest presented with a preferable balance between sensitivity and specificity. This is an important consideration when dealing with prognostications that have real life consequences, such as crime prediction. Moreover, in all models, SVM showed a substantially poorer performance relative to random forest or elastic net. Since a favourable linear decision boundary was observed with Elastic Net, it is possible that this decrease in performance with SVM was due to the radial kernel used, which creates a nonlinear decision boundary [20]. Moreover, altering cost and gamma hyperparameters using a grid search did not significantly improve performance.

Algorithms for classification tasks

As mentioned previously, the current study used Elastic Net, Random Forest, and SVM algorithms to fit classifiers. However, there are several other algorithms useful in classification problems, such as various boosting algorithms (AdaBoost[21] and XGBoost[22]), as well as deep learning [23]. In brevity, AdaBoost uses a weak learning condition to derive a boosting algorithm that produces a final classifier with an arbitrarily small generalization error. It tends to perform well in classification problems [21]. Furthermore, extreme gradient boosting (XGBoost), also involves an ensemble of weak prediction models, and can be used with tree or linear algorithms [24]. The primary difference between these two forms of boosting algorithms is that XGBoost introduces a more regularized model to control overfitting, which has been demonstrated to enhance performance in many applications [22]. This also lends itself to the question as to the main difference between Random Forest and various gradient boosted trees. Primarily, this relates to how trees are built, with random forest building each tree independently, while gradient boosting builds one tree at a time [25]. In addition, various forms of deep learning have been shown to achieve state-of-the-art performance in applications such as image and speech recognition tasks [26]. However, tree-based models have been found to generally outperform deep learning on tabular-style datasets, that lack strong multiscale temporal or spatial structures, such as the dataset in the current project [27].

Explainable AI

Although more sophisticated classifiers generally result in better performance, interpretability can become a challenge [28]. For instance, decision trees are a highly interpretable method, but are generally inaccurate and prone to large changes in tree structure as a result of small changes in the training set [29]. However, random forest, which involves an ensemble of decision trees, generally shows high performance, but is difficult to interpret [30]. Although methods such as variable importance plots allow for a global interpretation of the relative weight of features within the model, this largely neglects the impact of input features in predicting single samples [31]. Recently, new local explanation methods have been developed, including Tree Explainer [27], which uses a similar strategy as the concept of the explanation space [32]. Tree Explainer uses the internal structure of tree-based models to efficiently compute local explanations using Shapley values, a concept from cooperative game theory that distributes gains and costs to players working together to obtain a payoff [27]. Moreover, this method captures interaction effects between features, and allows for directly monitoring the impact of individual features on model loss [27].

Both methods can be conceptualized as a form of supervised clustering, where samples are grouped based on their explanations [27,32], which may prove useful in facilitating further transparency and interpretability in models with high-stakes decisions. Furthermore, these methods have the potential of uncovering clusters, or phenotypes, in binary classifiers [27,32]. While this falls outside of the scope of the current work, methods such as Tree Explainer may be useful to identify phenotypes of criminal recidivism and forensic outcomes in prospective studies.

Binary vs Multiclass Machine Learning

Of additional note, it is possible to train a classifier using either supervised or semi-supervised algorithms in multiclass problems [33,34]. In other words, models can be designed to distinguish between more than two classes. Depending on the specific task, binary or multiclass classification may be more suitable. Briefly, one-vs-one and one-vs-all classifiers are among the available options to achieve this [35, 36]. One-vs-one classifiers selects two classes at a time, where a binary classifier is trained for each. This is performed for each pair of classes, with a maximum of n(n-1)/2 classes in total. During the classification task, all binary classifiers are trained. However, in one-vs-one classifiers, interpretability of performance metrics become more of a challenge [36]. Similarly, one-vs-all classifiers involve training a classifier, where each class is fitted against other classes. In this case, each class is represented by only one classifier, which improves the interpretability of this approach, relative to one-vs-one classifiers. As such, it is a commonly used strategy in multiclass scenarios [37].

The current work involved binary classification only. This was done in the interest of improving model interpretability, as important features between two classes may be obfuscated in multiclass scenarios. Despite this rationale, one-vs-all classifiers can still provide valuable insight in identifying important features specific to the individual class [37]. Moreover, it is argued that multiclass classifiers may provide better utility in several healthcare applications, and future work may benefit from their use [38]. Experimental designs that encompass both multiclass classifiers, and binary classifiers with improved interpretability, such as Tree Explainer, may be useful as a compromise between the interests of interpretability and clinical utility.

Limitations

The current work has some potential limitations. While a representative sample of 1240 forensic inpatients were used from 10 psychiatric institutions across Ontario, Canada, this may not be representative of patients in other countries, or jurisdictions. Furthermore, while reasonable performance accuracy was observed in the present study, further refinement of risk prediction models is needed. Similarly, a much smaller error rate is required to implement such predictive models as clinical tools. Additionally, variables with missing data were excluded from the analysis. While some of these excluded variables may have proven to be useful, increasing the imputation threshold from 15% to 30% did not result in a significant change in accuracy in any of the models. Likewise, other imputation strategies, such as *k*-nearest neighbours[39], may be a useful alternative in the case of missing data. However, it is important to note that each imputation strategy has its own set of limitations[40,41]. Apart from this, the current study used binary classifiers to distinguish between types of crime at an individual level. Other studies may benefit from using one-vs-one and one-vs-rest classifiers [36]. Also, other algorithms, and preprocessing pipelines may lead to different performance metrics.

Summary

In summary, the present results suggest that machine learning models have accuracy comparable to existing risk assessment tools (AUCs .70-.80). However, unlike existing risk tools, this approach allows for the prediction of cases at an individual level, which is more clinically useful. Moreover, this represents a first attempt to predict the type of crime an individual will commit, using variables occurring prior to the offense. The accuracy of prospective models is expected to only improve with further refinement.

Perspectives

Moving forward, a further refinement of predictive models in forensic risk prediction is required. Potentially, this may be facilitated by using a wider framework when selecting the input data in our models. Considering that our model performance is directly dependent on the available input data, an exploratory data-driven approach may be warranted in risk prediction studies.

The vast majority of ML studies in forensic psychiatry thus far focus purely on clinical and administrative data, given the widespread availability of such data. However, other modalities, such as neuroimaging (MRI, fMRI, DTI), electrophysiology (EEG, MEG, ERG) and various sensors (actigraphy, heart rate variability), may prove to facilitate model performance, when used in conjunction with clinical data. Moreover, longitudinal studies with larger multicentric samples and adequate external validation are needed to translate proof-of-concept predictive models into applications to be used in clinical and legal settings. We hypothesize that such models may facilitate a more personalized approach to patient evaluation and risk management, provide greater precision in deriving a tailored treatment plan, and aid clinicians and the legal system in the decision-making process as it pertains to mentally disordered offenders. Ultimately, they may become critical tools to assist in prison sentencing, to determine fitness to stand trial, and to optimize the progress of individuals in the forensic system towards rehabilitation.

References

1. Hart, S. D., Michie, C. & Cooke, D. J. Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *Br. J. Psychiatry* (2007). doi:10.1192/bjp.190.5.s60

2. Palocsay, S. W., Wang, P. & Brookshire, R. G. Predicting criminal recidivism using neutral networks. *Socioecon. Plann. Sci.* (2000). doi:10.1016/S0038-0121(00)00003-3

3. Veronezi, B. P. *et al.* Evidence for increased motor cortical facilitation and decreased inhibition in atypical depression. *Acta Psychiatr. Scand.* (2016). doi:10.1111/acps.12565

4. Cohen, M. I., Spodak, M. K., Silver, S. B. & Williams, K. Predicting outcome of insanity acquittees released to the community. *Behav. Sci. Law* (1988). doi:10.1002/bsl.2370060408

5. Silver, E., Smith, W. R. & Banks, S. Constructing actuarial devices for predicting recidivism a comparison of methods. *Crim. Justice Behav.* (2000). doi:10.1177/0093854800027006004

6. Coid, J., Mickey, N., Kahtan, N., Zhang, T. & Yang, M. Patients discharged from medium secure forensic psychiatry services: Reconvictions and risk factors. *Br. J. Psychiatry* (2007). doi:10.1192/bjp.bp.105.018788

7. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. in *Data Classification: Algorithms and Applications* (2014). doi:10.1201/b17320

8. Dash, M. & Liu, H. Feature selection for classification. *Intell. Data Anal.* (1997). doi:10.3233/IDA-1997-1302

9. Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning (Adaptive Computation and Machine Learning series)*. *The MIT Press* (2012). doi:10.1007/978-3-642-34106-9_15

10. Solorio-Fernández, S., Carrasco-Ochoa, J. A. & Martínez-Trinidad, J. F. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* (2020). doi:10.1007/s10462-019-09682-y

11. Ang, J. C., Mirzal, A., Haron, H. & Hamed, H. N. A. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* (2016). doi:10.1109/TCBB.2015.2478454

12. Smialowski, P., Frishman, D. & Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp621

13. Porkodi, R. Comparison of Filter Based Feature Selection Algorithms : an Overview. *Int. J. Innov. Res. Technol. Sci.* (2014).

14. Leng, J., Valli, C. & Armstrong, L. A Wrapper-Based Feature Selection for Analysis

of Large Data Sets. in *Proceedings of 2010 3rd International Conference on Computer and Electrical Engineerings (ICCEE 2010)* (2010).

15. Kwak, N. & Choi, C. H. Input feature selection for classification problems. *IEEE Trans. Neural Networks* (2002). doi:10.1109/72.977291

16. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* (2007). doi:10.1093/bioinformatics/btm344

17. Huang, D. & Chow, T. W. S. Effective feature selection scheme using mutual information. *Neurocomputing* (2005). doi:10.1016/j.neucom.2004.01.194

18. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. in *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003* (2003). doi:10.1109/CSB.2003.1227396

19. Armanfard, N., Reilly, J. P. & Komeili, M. Local Feature Selection for Data Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016). doi:10.1109/TPAMI.2015.2478471

20. Amari, S. & Wu, S. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks* (1999). doi:10.1016/S0893-6080(99)00032-5

21. Schapire, R. E. Explaining adaboost. in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (2013). doi:10.1007/978-3-642-41136-6_5

22. Melville, S. xgboost: Extreme Gradient Boosting. *R Lect.* 1–84 (2014). doi:10.1145/2939672.2939785>.This

23. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* (2015).

doi:10.1038/nature14539

24. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). doi:10.1145/2939672.2939785

25. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statiscal Learning with Applications in R*. *Springer* (2013). doi:10.1016/j.peva.2007.06.006

26. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx044

27. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* (2020). doi:10.1038/s42256-019-0138-9

28. Gilpin, L. H. *et al.* Explaining explanations: An overview of interpretability of machine learning. in *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018* (2019). doi:10.1109/DSAA.2018.00018

29. Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* (1986). doi:10.1023/A:1022643204877

30. Qi, Y. Random forest for bioinformatics. in *Ensemble Machine Learning: Methods and ApplicatiOns* (2012). doi:10.1007/9781441993267_10

31. Molnar, C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book* (2019).

32. Schulz, M.-A., Chapman-Rounds, M., Verma, M., Bzdok, D. & Georgatzis, K. Clusters in Explanation Space: Inferring disease subtypes from model explanations. 1–8 (2019).

33. Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. Machine learning: A review of classification and combining techniques. *Artif. Intell. Rev.* (2006). doi:10.1007/s10462-007-9052-3

34. Sinha, K. Semi-supervised learning. in *Data Classification: Algorithms and Applications* (2014). doi:10.1201/b17320

35. Aly, M. & Edu>, <malaa@caltech. Survey on multiclass classification methods. *Neural Netw* (2005).

36. Galar, M., Fernández, A., Barrenechea, E., Bustince, H. & Herrera, F. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit.* (2011). doi:10.1016/j.patcog.2011.01.017

37. Rifkin, R. & Klautau, A. In defense of one-vs-all classification. *J. Mach. Learn. Res.* (2004).

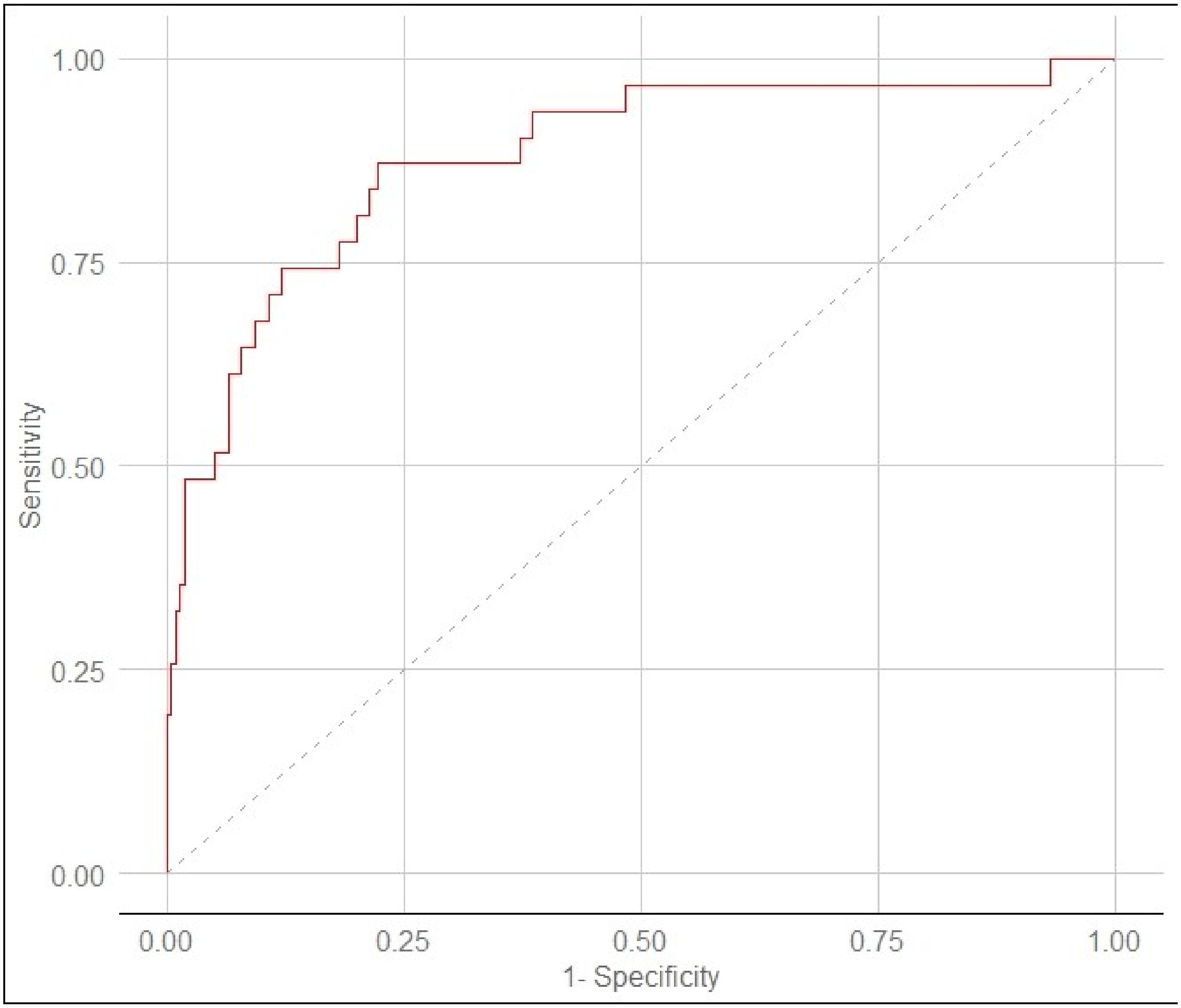
38. Oza, N. C. & Tumer, K. Classifier ensembles: Select real-world applications. *Inf. Fusion* (2008). doi:10.1016/j.inffus.2007.07.002

39. Beretta, L. & Santaniello, A. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med. Inform. Decis. Mak.* (2016). doi:10.1186/s12911-016-0318-z

40. Somasundaram, R. S. & Nedunchezhian, R. Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values. *Int. J. Comput. Appl.* (2011). doi:10.5120/2619-3544

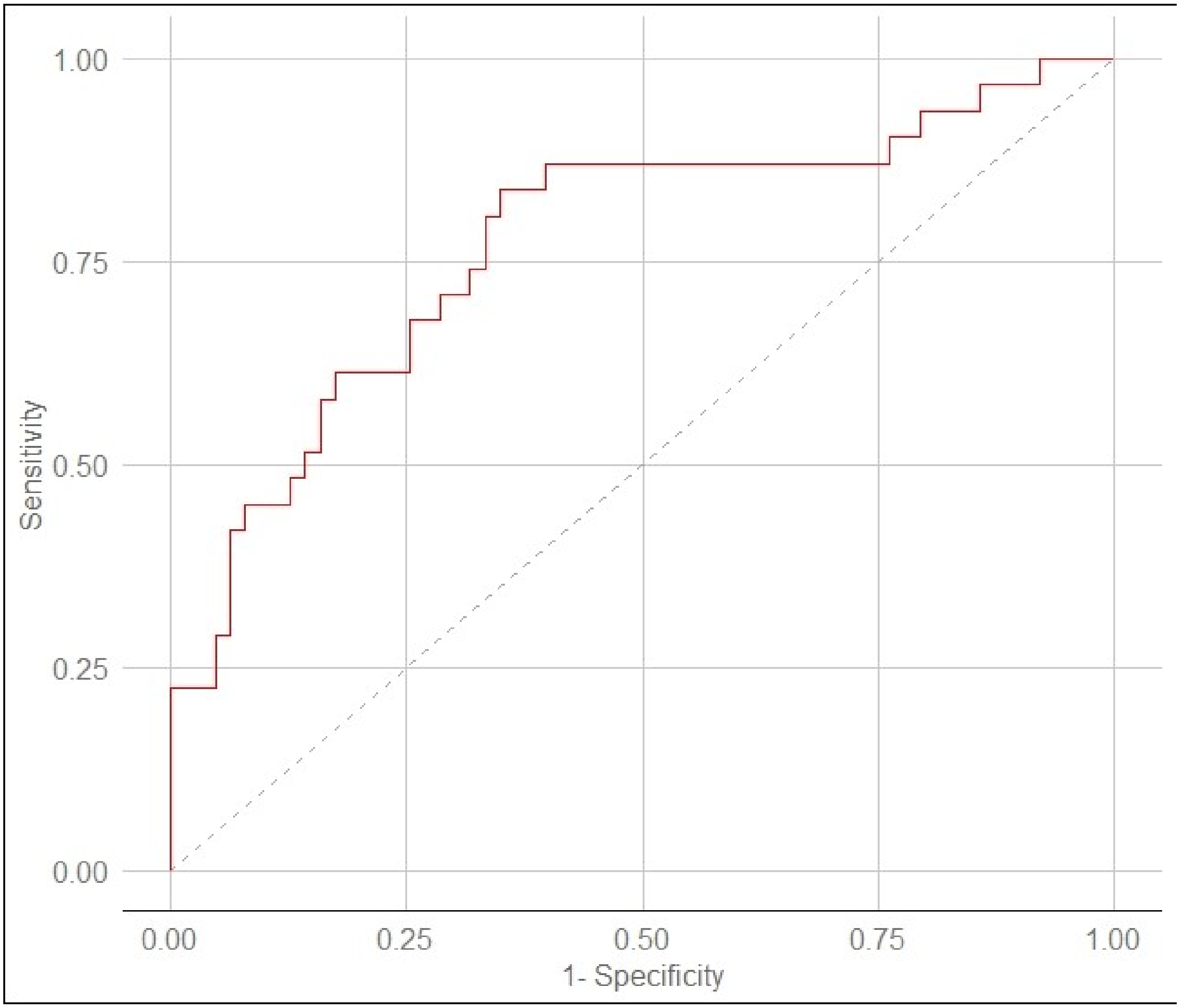
41. Zhang, S., Wu, X. & Zhu, M. Efficient missing data imputation for supervised learning. in *Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010* (2010). doi:10.1109/COGINF.2010.5599826

Appendix 1: ROC Curve: Violent vs. Sexual Offences



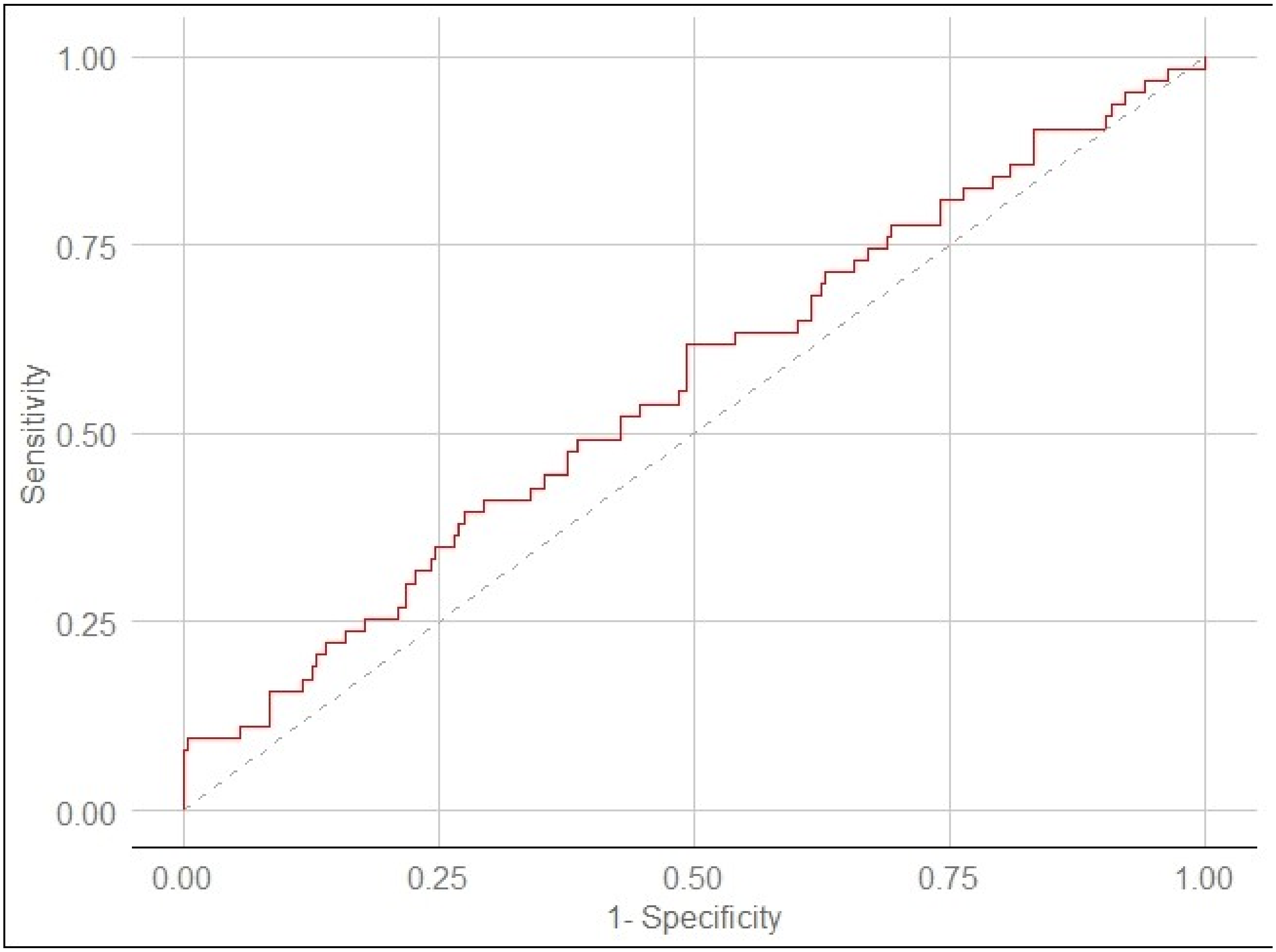
ROC curve is a plot of sensitivity versus 1-specificity (often called the false-positive rate) that offers a summary of sensitivity and specificity across a range of cut points for a continuous predictor.

Appendix 2: ROC Curve: Nonviolent vs. Sexual Offences



ROC curve is a plot of sensitivity versus 1-specificity (often called the false-positive rate) that offers a summary of sensitivity and specificity across a range of cut points for a continuous predictor.

Appendix 3: ROC Curve: Violent vs. Nonviolent Offences



ROC curve is a plot of sensitivity versus 1-specificity (often called the false-positive rate) that offers a summary of sensitivity and specificity across a range of cut points for a continuous predictor.