THE APPLICATION OF MACHINE LEARNING TO A GENERAL RISK–NEED ASSESSMENT INSTRUMENT IN THE PREDICTION OF CRIMINAL RECIDIVISM

MEHDI GHASEMI
University of Saskatchewan

DANIEL ANVARI
Kwantlen Polytechnic University

MAHSHID ATAPOUR
Capilano University

J. STEPHEN WORMITH

KEIRA C. STOCKDALE
Saskatoon Police Service
University of Saskatchewan

RAYMOND J. SPITERI
University of Saskatchewan

The Level of Service/Case Management Inventory (LS/CMI) is one of the most frequently used tools to assess criminogenic risk–need in justice-involved individuals. Meta-analytic research demonstrates strong predictive accuracy for various recidivism outcomes. In this exploratory study, we applied machine learning (ML) algorithms (decision trees, random forests, and support vector machines) to a data set with nearly 100,000 LS/CMI administrations to provincial corrections clientele in Ontario, Canada, and approximately 3 years follow-up. The overall accuracies and areas under the receiver operating characteristic curve (AUCs) were comparable, although ML outperformed LS/CMI in terms of predictive accuracy for the middle scores where it is hardest to predict the recidivism outcome. Moreover, ML improved the AUCs for individual scores to near 0.60, from 0.50 for the LS/CMI, indicating that ML also improves the ability to rank individuals according to their probability of recidivating. Potential considerations, applications, and future directions are discussed.

Keywords: LS/CMI; risk–need assessment; predictive accuracy; machine learning

Although efforts to predict criminal recidivism date back 90 years (Burgess, 1928), the last two decades have witnessed an explosion in the use of risk-assessment tools in criminal justice systems around the world. These tools vary dramatically in their length,

AUTHORS' NOTE: *The views expressed are solely those of the authors and do not necessarily reflect those of the Saskatoon Police Service. In addition, we wish to acknowledge support from the Ontario Ministry of Community Safety and Correctional Services, the Centre for Forensic Behavioural Science and Justice Studies at the University of Saskatchewan, the Saskatchewan Police Predictive Analytics Lab, Mitacs, and the Natural Sciences and Engineering Research Council of Canada.* *Correspondence concerning this article should be addressed to Raymond J. Spiteri, Department of Computer Science, University of Saskatchewan, 176 Thorvaldson Building, 110 Science Place, Saskatoon, Saskatchewan, Canada S7N 5C9; e-mail: spiteri@cs.usask.ca.*

CRIMINAL JUSTICE AND BEHAVIOR, 201X, Vol. XX, No. X, Month 2020, 1–21.
DOI: 10.1177/0093854820969753

scope, design, and method of calculating or appraising risk. They also vary in the type of forensic clientele for whom they are designed, the type of outcome they are meant to predict (e.g., types of recidivism), and the context in which they are applied (Andrews et al., 2006). Yet, they also tend to have some common characteristics. For example, most risk-assessment tools capture data about an individual's criminal history, a so-called static or historical factor and perhaps the most well-established risk factor of subsequent criminal behavior. Another characteristic that binds all forensic risk-assessment instruments is that they are ultimately intended to promote public safety by identifying individuals who are most likely to reoffend. It is then the responsibility of the criminal justice system (police, courts, correctional agencies, and community organizations) to use the results of forensic risk assessments to employ the appropriate means at their disposal to reduce or prevent further criminal behavior.

THE LEVEL OF SERVICE (LS) FAMILY OF RISK-ASSESSMENT TOOLS

The Level of Service/Case Management Inventory (LS/CMI; Andrews et al., 2004) is the latest version of a forensic risk–need assessment measure from a family of tools known as the LS scales. Versions of the LS scales have been used worldwide since the early 1990s, with increasing popularity over the last decade. For instance, by 2010, more than one million administrations were officially registered with the test publisher in a single year (Wormith, 2011). The popularity of the LS scales may be attributed to several important characteristics. First, unlike strictly actuarial measures, the LS scales were developed from well-established criminological and psychological theories (e.g., differential association theory, social learning theory), including a general personality and cognitive social learning theory of criminal behavior (e.g., Andrews & Bonta, 1994). Second, the LS scales have a rich tradition of research supporting its content and use in practical ways for correctional practitioners (Gendreau et al., 1996). This includes numerous validation studies and meta-analyses (e.g., Olver et al., 2014). Third, the LS scales have been found to have general applicability across many forensic populations. This includes adults and youth in custody or on community supervision, male and female populations, and various ancestral/ethnic backgrounds and cultures on diverse measures of recidivism, ranging from technical violations to criminal charges and convictions (e.g., Olver et al., 2009; Smith et al., 2009; Wilson & Gutierrez, 2014). Fourth, the LS scales have multiple applications in corrections. This includes not only the prediction of criminal recidivism but also the planning and delivery of forensic services and case management practices to prevent recidivism (e.g., Luong & Wormith, 2011), an attribute made possible because the scale includes dynamic risk factors, also known as criminogenic needs, as well as static risk factors, hence its status as a risk–need scale. Fifth, the LS scales strike a balance between comprehensiveness and simplicity. Ratings in applied settings require a skilled interview of forensic clientele, yet items are scored in a dichotomous (0–1) fashion and then summed. As such, it can easily be scored manually by a trained assessor.

A pilot version of the LS/CMI called the Level of Service Inventory–Ontario Revision (LSI-OR; Andrews et al., 1995) was introduced in Ontario, Canada, in 1995, and remains in use throughout this provincial jurisdiction. More than 20,000 administrations of this version are applied to forensic clientele in Ontario annually. For simplicity, in this study, we use the more generally known and widely used name for this version of the tool, the LS/CMI.

The LS/CMI (Andrews et al., 2004) consists of 43 items that are grouped into eight domains or subsections, commonly referred to as the "central eight" (Andrews & Bonta, 2010). They include criminal history (eight items), education/employment (nine items), family/marital (four items), leisure/recreation (two items), companions (four items), substance abuse (eight items), pro-criminal attitudes (four items), and antisocial pattern (four items). Although individual items are scored in the same dichotomous fashion, the domains are weighted by virtue of their differing number of items. Other sections of the LS/CMI are used in a checklist fashion, serving as flags for issues of particular concern, but they will not be reviewed here because they are not the focus of the current study.

Numerous studies have examined the predictive validity of the LS scales. A recent meta-analysis of 151 independent samples and 137,931 justice-involved individuals by Olver et al. (2014) demonstrated the predictive validity of the LS scales for any recidivism (mean random effects correlations of .30 for males and .31 for females). Although the predictive accuracy for general recidivism was consistently higher than the predictive accuracy for violent recidivism (overall mean random effects correlations r were .29 and .23, respectively), the LS/CMI generated higher correlations for both general and violent recidivism ($r = .42$ and .27, respectively) than the other LS variants ($r = .25$–$.30$ and .21–.28, respectively). However, the number of studies examining the LS/CMI means was modest ($k = 12$ and 11, respectively) because of the relative newness of this version of the tool.

Regardless, these investigations have evaluated the prescribed arithmetic scoring of the LS/CMI using traditional statistical approaches. It is possible that all predictor variables (risk factors) may not be of equal weight or demonstrate only linear relationships with criterion data (recidivism) as discussed by Garb and Wood (2019) in their recent review of methodological advances in statistical prediction. Newer statistical approaches that examine complex predictors in novel ways may yield important insights.

APPLICATIONS OF MACHINE LEARNING (ML) TO FORENSIC RISK ASSESSMENT

It is well known that humans generally do not have the best track record when it comes to making rational decisions and judgments (e.g., Grove et al., 2000). Even trained professionals do not fare nearly as well as basic actuarial algorithms when predicting human behaviors such as criminal recidivism (Andrews & Bonta, 2010), especially when faced with extensive information, limited feedback, and varying base rates for recidivistic events (Lin et al., 2020). The reasons for this are many, including the human mind's limits on working memory and human susceptibility to cognitive bias, emotion, fatigue, and so on (e.g., Dawes et al., 1989). Augmenting human capabilities with actuarial algorithms and computer-aided tools, including ML, may help to improve risk assessment and decision-making for correctional clientele.

ML is a branch of computer science that evolved from computational learning theory in artificial intelligence (e.g., Marsland, 2015; Murphy, 2012). It explores the analysis and construction of algorithms that can learn from, and make predictions about, relevant data. Because the amount of data available to scientists has recently seen unprecedented growth and ML techniques are "designed for the analysis of high-dimensional data with hundreds or thousands of predictors" (Garb & Wood, 2019, p. 1461), they have been attracting a great deal of attention. ML has been successfully applied to the solution of problems in diverse areas, including medicine and health care delivery systems, and has resulted in improved diagnostic accuracy and efficiency (e.g., Deo, 2015; Lavecchia, 2015; Topol, 2019).

Over the last decade, a growing debate also has mounted about the use of "big data" and ML algorithms in criminology and criminal justice generally and forensic risk assessment in particular (Berk & Bleich, 2013). Some have suggested that ML may improve the "hit rate" of extant risk-assessment tools (true positives [TPs] plus true negatives [TNs]) such as the LS scales (Wormith & Bonta, 2017, p. 135), whereas others have cautioned that the incremental validity of using ML approaches may be "modest," especially when the data are less complex (e.g., a data set containing scores on a single risk-assessment instrument; Garb & Wood, 2019, p. 1464). However, Duwe and Kim (2017) remind us that ML includes many different statistical techniques (e.g., decision tree [DT]–based algorithms, neural networks [NN], and support vector machines [SVMs]) and applications to the criminal justice field are still in their "infancy" (p. 597). Helpful overviews of ML techniques are provided by Tollenaar and van der Heijden (2013) and Duwe and Kim (2017). We limit the scope of our review to applications of ML to risk assessment, specifically predictive validity, given the importance of this type of validity in criminal justice decision-making and considering the majority of applications of ML to risk assessment has focused on predictive accuracy. However, we recognize that the "success" of a model strongly depends on the performance metric used, and predictive accuracy is only one of many important functions of risk-assessment tools. Although a range of prediction performance metrics exist, in addition to accuracy (ACC), we report the commonly reported area under the receiver operating characteristic curve (AUC) because of its frequency of use and intuitive application. Possible AUC values range from 0 to 1, representing the probability that a randomly selected recidivist would have a higher score than a randomly selected nonrecidivist (Rice & Harris, 2005).

In an early application of ML, Liu et al. (2011) compared logistic regression (LR), classification and regression trees (CART), and NN in the prediction of violent reoffending using a large sample of adult males in custody in the United Kingdom (N = 1,225). Prediction variables were taken from the Historical Clinical Risk Management–20 (HCR-20; Webster et al., 1997), a structured professional judgment (SPJ) approach to the assessment and management of violence, whereby assessors make clinical decisions based on the item data they collect, as opposed to quantitative estimates of risk. Although NN performed marginally better than LR and CART, the authors concluded that the improvement did not warrant the use of NN over traditional prediction schemes (with all AUCs ranging from 0.65 to 0.72 for violent recidivism).

In 2013, Tollenaar and van der Heijden compared the use of LR with several ML techniques, including multivariate regression spline, linear discriminant analysis, flexible discriminant analysis, recursive partitioning, adaptive boosting, logitBoost, NN, linear support vector networks, *k*-nearest-neighbors classification, and partial least squares, in the prediction of general, violent, and sexual recidivism. However, rather than using items from an existing risk-assessment tool, Tollenaar and van der Heijden used a host of available static variables that were available from offender databases (N = 20,000) in the Netherlands (e.g., criminal history counts). Overall, they found the most accurate model varied with the type of sample (e.g., offending subtypes and recidivism base rates) and the outcome being predicted (e.g., sexual, violent, and general reoffending), and ML approaches to the prediction of criminal recidivism generally were not superior to traditional regression-based approaches (with AUCs ranging from 0.708 to 0.776 for general recidivism). However, the conclusions drawn by Tollenaar and van der Heijden in 2013 were criticized at the time for being

premature (Berk & Bleich, 2013), and there were calls for further explorations of ML approaches (e.g., Brennan & Oliver, 2013; Bushway, 2013; Ridgeway, 2013).

More recent results have been mixed. For instance, Hamilton et al. (2015) compared the predictive accuracy of the Washington State Static Risk Assessment using traditional (LR) and ML methodologies (NN and random forest (RF) approaches) in a large sample of corrections clients reentering the community in the state of Washington ($N = 297,600$). AUCs ranged from 0.732 to 0.762 depending on the outcome of interest, with LR and ML approaches demonstrating comparable performance. However, using a sample of 40,000 individuals released from prison in Minnesota, Duwe and Kim (2016) found that prediction models developed with supervised learning classifiers outperformed classification techniques commonly used in risk–needs assessment tools (e.g., summative classification or the Burgess method).

To further investigate the performance of newer ML approaches relative to traditional methods in predicting recidivism, Duwe and Kim (2017) subsequently compared the predictive accuracy of 12 supervised learning algorithms. The data set used in the study was derived from that used to develop the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR; Duwe, 2014) and comprised 27,772 individuals released from prisons in Minnesota. The MnSTARR contains both static (e.g., criminal history) and dynamic items pertaining to institutional adjustment (e.g., disciplinary infractions, involvement in programming; Duwe, 2014; Duwe, 2019), and as such, both static and dynamic predictors were included in the data set. Newer ML approaches such as LogitBoost (AUC = 0.777), RFs (AUC = 0.781), and MulitBoosting (AUC = 0.775) were found to yield better results for general recidivism, albeit only modestly. Moreover, the methods yielding the best performance varied across 10 "scenarios" that varied by gender and type of recidivism.

As Duwe and Kim (2017) pointed out, and we concur, these results would seem to suggest that the type of statistical methods employed to assess risk for recidivism may depend on the purpose for which risk is being assessed and how they are being used. For example, a large correctional agency looking to automate risk classification within a geographic region or across institutions may require different technologies than an individual assessor who is looking to identify criminogenic needs and generate recommendations for case management. As with traditional methodologies, further "tuning" of ML models (e.g., calibration) is also required to address issues of diversity, including responsivity considerations (e.g., gender, ethnic/cultural background, age), location (e.g., region, country, institution vs. community settings), and time of assessment (e.g., intake, release, pre-/posttreatment). Risk variables may also change over time with or without intervention (e.g., cohort effects, treatment change).

In the last few years, there have been some promising and also concerning findings. For instance, using a large data set of predictors of offending in Texas ($N = 258, 248$), Curtis (2018) found that modern ML approaches predicted general arrest. Large effects were reported for RFs (AUC = 0.808) and XGBoost (AUC = 0.792). However, the majority of the top predictors were static predictors (e.g., criminal history), and one of the top predictors was the number of tattoos! An examination of 336 predictor variables in a sample of 3,061 juveniles in Florida with a history of sexual offenses by Ozkan et al. (2020) also found that RF models yielded strong findings with AUCs of 0.71 for an "all-predictors model" and 0.65 a for a "legal factors" model. Although comparable with AUCs reported in a recent

meta-analysis of tools used to access sexual recidivism in juveniles (AUCs = 0.64–0.67; Viljoen et al., 2012) to be included in the data set, youth were required to have been scored on the Positive Achievement Change Tool (PACT; Baglivio, 2009). It is unclear as to how these findings may compare with the predictive accuracy of the PACT alone in this sample or how best to interpret and apply this "black-box" all-predictors model.

Perhaps most interestingly, using only static, historical information about adult males who had been convicted of a sexual offense for the first time ($N = 756$), Lussier et al. (2019) were able to generate novel insights using ML approaches, specifically decision tree algorithms (DTAs), including chi-square automatic interaction detection (CHAID), Quick Unbiased Efficient Statistical Tree (QUEST), and CART. Although classic LR and DTA predictive models were not appreciably different, the use of DTAs (AUCs = 0.704–0.733) revealed the presence of different risk profiles for entry into sexual recidivism that were not revealed by classic LR (AUC = 0.746). Thus, there may benefit to combining traditional and modern approaches when assessing risk over time, development, and the course of a criminal career.

The purpose of the current study is to extend this body of work by applying modern ML approaches to a widely used, general risk–need assessment tool that is theoretically informed (in contrast to dustbowl empiricism), includes both static and dynamic factors, and can follow justice-involved individuals from intake through to case closure, often referred to as a "fourth-generation" tool (Andrews et al., 2006). Duwe and Kim (2017) have suggested that "fourth-generation risk-assessment instruments based on ML algorithms could potentially improve correctional practice" (p. 596; for example, access to programming). Moreover, the approach taken by Lussier and colleagues (2019) indicates that deductive (e.g., LR), inductive (i.e., ML), and combined approaches to risk modeling may contribute different theoretical and analytic insights. As a first exploratory step, we examine the performance of ML techniques relative to LS/CMI score in terms of predictive accuracy and the ability to rank individuals according to their probability of recidivating by means of a secondary analysis of two large data sets containing LS/CMI administrations for individuals in provincial custody in Ontario, Canada.

METHOD

DATA SETS

Provincial correctional policy in Ontario requires the administration of the LS/CMI to all individuals who are given a term of probation or sentenced to a period of incarceration of between 3 months and 2 years. Individuals who are sentenced to more than 2 years are transferred to the federal correctional authority, the Correctional Service of Canada; hence, they are not under provincial jurisdiction and are not administered the LS/CMI by the provincial correctional authority. With the introduction of an electronic data capture and scoring mechanism for the LS/CMI, the collection of large LS/CMI data sets became possible. In this article, we analyzed a combination of two data sets provided by the Ontario Ministry of Community Safety and Correctional Services (MCSCS).

The first data set (D1) consists of 72,725 records for individuals who were interviewed and assessed on the LS/CMI by MCSCS correctional staff during 2010 and 2011. This cohort included correctional clientele who had been sentenced to prison for a term of 3 to 24 months and then released from custody as well as those who were sentenced to a period of probation during this 2-year period. Individuals were then followed up at the end of 2013

through official records of readmission to the justice system in Ontario. Where applicable, dates of the first recidivistic event were used to determine which individuals criminally recidivated and how long it took them to do so. The average follow-up time during which individuals were eligible to reoffend was approximately 2.96 years (SD = 7.5 months), ranging from 1.04 to 4.5 years.

The second data set (D2) was retrieved using an earlier cohort from the same jurisdiction and under the same conditions as the 2010–2011 cohort; however, the second cohort spanned only a single year, 2004. LS/CMI data and recidivism outcomes were collected in the same manner as with the previous data set. A total of 26,450 individuals were then followed until January 2009, an average follow-up time of 4.54 years (SD = 3.5 months), ranging from 4.02 to 5.02 years.

It is important to note that these data sets represent two cohorts of consecutive admissions to the MCSCS system. Recidivism for both data sets was defined as any criminal offense for which an individual is returned into the MCSCS system on a reconviction, sentenced to either incarceration or community supervision.

Table 1 provides summary statistics for data sets D1 and D2, including information on mean total LS/CMI risk scores as well as recidivism rates for each gender and risk level. We note that the mean LS/CMI score for D1 (M = 14.30; SD = 8.91) is higher than the mean score for D2 (M = 12.53; SD = 8.79). We find this difference to be statistically significant, with a p value of less than 10^{-8}; however, we consider this p value to be an artifact of the large sample size. Effect sizes given in Table 1 using Hedges's g suggest small effect sizes ($\leq$0.27) between mean LS/CMI scores across databases (overall and sex stratified) and trivial effect sizes ($\leq$0.07) across databases stratified by risk levels. Figure 1 provides a more detailed visual presentation of the difference in the distribution of the LS/CMI scores for the two data sets. Referring to the boxplots in Figure 1A and B, it can be seen that the box containing the middle 50% scores for D1 is located at a higher level than that of D2. Also, D2 contains more outliers with high scores. We also note from Table 1 that the rate of recidivism is significantly higher (with p value of less than 10^{-8}) for D2, 36.01%, compared with 30.52% for D1. We also consider this p value to be an artifact of the large sample size. Effect sizes given in Table 1 using Cohen's h suggest small effect sizes ($\leq$0.26) between recidivism rates across databases (overall and stratified by sex and risk level). One possible explanation for this difference is the longer average follow-up time for D2 (4.54 years) compared with about 3 years for D1, allowing more opportunity for recidivism to occur and be recorded. Other summary statistics are given, including the mean, standard deviation, and median LS/CMI scores as well as AUCs (with 95% confidence interval) for the individual data sets according to gender and LS/CMI risk level.

Although data sets D1 and D2 differ in potentially important ways, the AUC values associated with the LS/CMI total scores are similar for D1 and D2 as well as for male and female subgroups (AUCs from 0.70 to 0.72). We make use of the combined data set (comprised of D1 and D2) for building and testing our predictive models to retain maximal data. The rationale for this approach is that our ultimate goal is to build a dynamic predictive model based on the maximum amount of available data. We create a training data set by selecting 50% of the records from each of D1 and D2 in a uniformly random fashion. These records are combined to form the training data set. The remaining records are combined to form the testing data set.

TABLE 1: Summary Statistics of LS/CMI Scores and Recidivism Rates for Data Sets D1 and D2

	D1 (n = 72,725)						D2 (n = 26,450)						D1 vs. D2	
Category	Prop. (%)	Recid. (%)	*M* (*SD*)	*Mdn*	AUC [95% CI]	Age	Prop. (%)	Recid (%)	*M* (*SD*)	*Mdn*	AUC [95% CI]	Age	Hedges's *g*	Cohen's *h*
Demographics														
Female	17.35	25.69	13.32 (8.45)	12	0.72 [0.69, 0.74]	33.81	18.32	28.90	11.05 (8.01)	9	0.71 [0.69, 0.73]	33.42	0.27	0.07
Male	82.65	31.54	14.51 (8.99)	13	0.72 [0.70, 0.73]	34.02	81.68	37.61	12.86 (8.93)	11	0.70 [0.68, 0.72]	33.23	0.18	0.13
All	100	30.52	14.30 (8.91)	13	0.72 [0.71, 0.73]	33.99	100	36.01	12.53 (8.79)	11	0.70 [0.68, 0.72]	33.27	0.20	0.12
LS/CMI risk level														
Very low (0–4)	13.67	7.51	2.51 (1.22)	3	—	—	19.88	11.73	2.44 (1.22)	3	—	—	0.06	0.14
Low (5–10)	25.90	15.00	7.49 (1.69)	7	—	—	29.25	22.57	7.38 (1.70)	7	—	—	0.06	0.19
Medium (11–19)	33.37	30.07	14.74 (2.59)	15	—	—	29.97	41.75	14.59 (2.59)	14	—	—	0.06	0.24
High (20–29)	20.32	52.51	23.93 (2.83)	24	—	—	15.44	64.85	23.99 (2.85)	24	—	—	0.02	0.25
Very high (30–43)	6.74	72.87	33.23 (2.78)	33	—	—	5.46	83.39	33.05 (2.72)	33	—	—	0.07	0.26

Note. Significant bivariate differences between D1 and D2 on mean LS/CMI scores and rates of recidivism. AUC = area under the receiver operating characteristic curve; CI = confidence interval; LS/CMI = Level of Service/Case Management Inventory; Prop. = proportion of data set size; Recid. = rates of recidivism for each LS/CMI risk level for the data sets D1 and D2; *M* (*SD*) = mean (standard deviation) LS/CMI score; *Mdn* = median LS/CMI score.

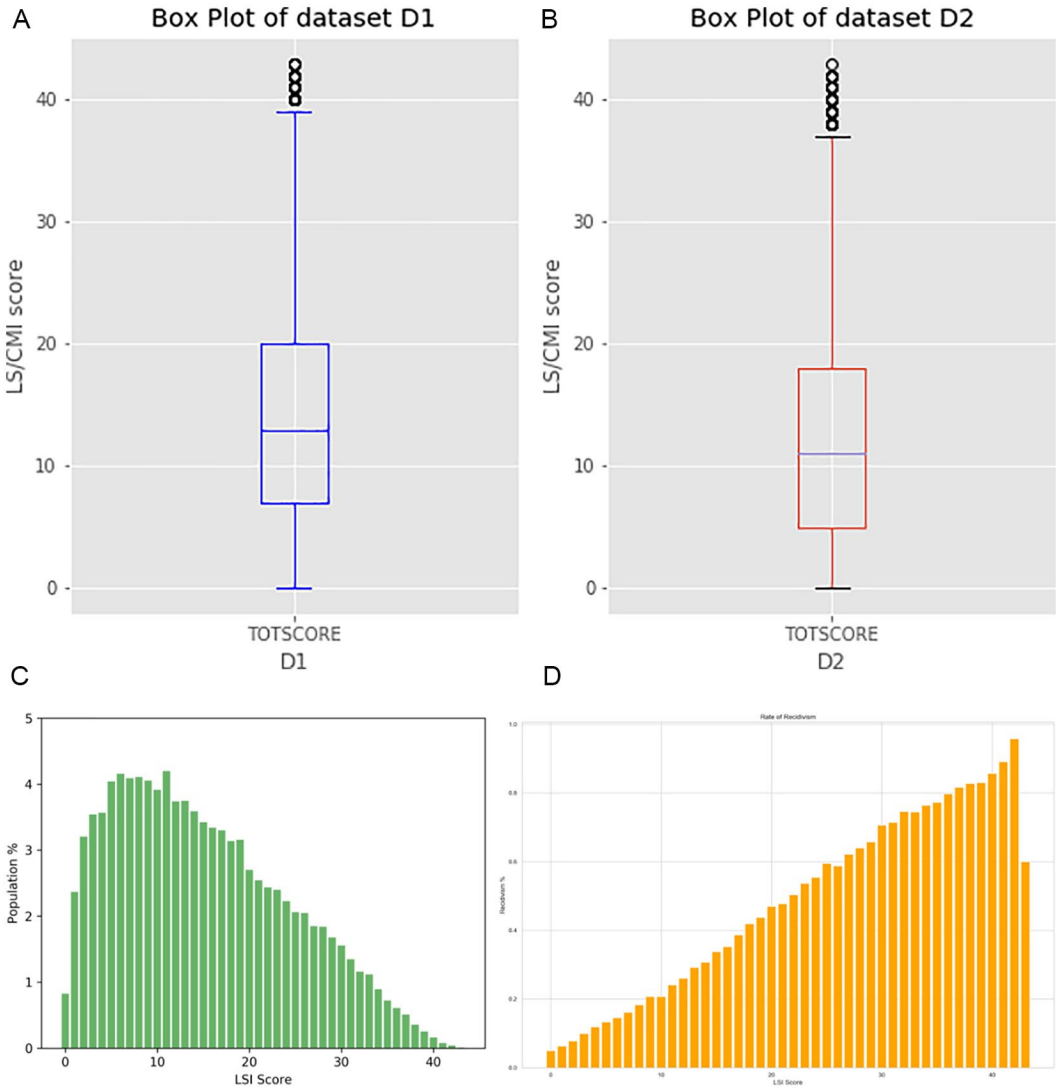

Figure 1: Visualizations of Different Aspects of the Distribution of LS/CMI Scores for Data Sets D1, D2, and the Combined Data Set: (A) Boxplot for Data Set D1 (n = 72,725); (B) Boxplot for Data Set D2 (n = 26,450); (C) Distribution of the LS/CMI Scores for the Combined Data Set; and (D) Rate of Recidivism for Each LS/CMI Score for the Combined Data Set

Note. LS/CMI = Level of Service/Case Management Inventory.

Additional details regarding these two data sets can be obtained from Wormith et al. (2012) and Wormith et al. (2015) as well as two master's theses (Hogg, 2011; Orton, 2014). Ethics approvals were obtained from the University of Saskatchewan for these projects as well as for a broader program of predictive research the current work sought to inform (BEH 16-166).

LS/CMI SCORES

The General Risk/Need Factors section of the LS/CMI consists of 43 risk–need items, A_i, scored dichotomously (0 = not present or 1 = present). Items are summed to provide a total score LS/CMI ranging from 0 to 43,

$$\text{LS/CMI} = \sum_{i=1}^{43} A_i,$$

and there are five risk levels associated with various ranges of scores. Table 1 gives the proportion of scores for each risk level as well as the corresponding rates of recidivism for data sets D1 and D2. In practice, a total LS/CMI score is obtained and compared with available norms to get a recidivism estimate. To algorithmically simulate this process in this study, LS/CMI is applied in the following way to predict recidivism. For a given data set, the recidivism rate for each score is calculated. If the recidivism rate for a given score is above 0.5, then any individual with that score is classified as likely to recidivate and otherwise not.

ML ALGORITHMS

There are various types of ML algorithms that can be applied to our data set, but because we have outcome or "target" data (i.e., data on whether an individual recidivated), we employ a class of ML algorithms generally known as supervised algorithms (Marsland, 2015). In supervised algorithms, the algorithm is fed by previously existing data (training data), where the target data are known, and the algorithm builds a model from these data. The goal is to enable the model to reliably predict target values on a new set of data (test data; that is, data on which the model has not been trained). More detailed descriptions of various ML algorithms can be found in Marsland (2015) and Murphy (2012). We now briefly describe the supervised ML algorithms used in this study.

DTs

An early example of a DT approach applied to risk assessment comes from the MacArthur Violence Risk Assessment study, in which Steadman and colleagues (2000) designed a method to predict violent recidivism among offenders with mental disorders. A DT learning approach refers to a predictive model that maps observations about an item to conclusions about the item's target value. Tree models, where the target variable can take a finite set of values, are called classification trees. In these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to those class labels. A tree can be constructed by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner (known as recursive partitioning). The recursion is completed when the subset at a node has the same value of the target variable or, when splitting, no longer adds value to the predictions (Marsland, 2015).

DTs may best be applied to those problems where instances are represented by attribute–value pairs, the target function has discrete output values, and the training data may contain errors. This makes DTs appropriate for our study because the target function (whether an individual recidivated) and all the input data (scores on LS/CMI items) are binary values. In

fact, the LS/CMI itself can be interpreted as a DT. For example, the LS/CMI classifies an individual based on their LS/CMI score into one of the five risk levels; then, based on the existing statistics for each risk level, it determines whether they are likely to recidivate or not.

RFs

RFs (e.g., Marsland, 2015) represent a set of classification algorithms that make predictions based on outputs of large number of decisions trees built on random subsets of features. RFs use the so-called bagging process to allow each individual tree to randomly sample from the training data set with replacement. This results in trees that are ultimately trained with different data and leads to more variation and diversification among the large number of trees in the forest. To increase the success of RF models, one needs to start with features that have a good level of predictive power and ensure these features are not highly correlated with each other. The overall idea is that if one tree can provide a good model, then many trees (a forest) should be able to do even better, provided there is enough diversity in the constituent trees.

SVMs

SVMs provide a state-of-the-art learning method that has been highly successful in a variety of applications. They are particularly effective when dealing with continuous data and data sets that are not linearly separable (Schölkopf & Smola, 2002). The SVM method has been developed based on two main ideas. The first idea is to map the feature vectors (data points) in a (nonlinear) way to a high (possibly infinite) dimensional space and then utilize linear classifiers in this new space. This mapping produces in nonlinear classifiers in the original space, thus overcoming the representational limitations of linear classifiers. However, the use of linear classifiers in the transformed space depends heavily on the computational methods for finding a classifier that performs well on the training data. The second idea is that, among the generally infinitely many hyperplanes that may separate the data, the linear classifier chosen is the one that maximizes the separation of the data (i.e., the one whose distance from it to the nearest data point on each side is maximized; Steinwart & Christmann, 2008). SVMs are suitable for classifying data of relatively high dimension. Because our data set consists of 43 LS/CMI variables, SVMs are a reasonable approach for classification.

K-FOLD CROSS-VALIDATION

The various ML models were all built in the following way using k-fold cross-validation with $k = 10$. First, the training set was randomly shuffled and divided into k equal parts (or folds). For each ML algorithm, k models were built using $k - 1$ of the folds as training data and the final fold as testing data. For any given performance metric, the results of the k models are averaged to provide the value reported.

SOFTWARE USED

The analytics presented in this article are the results from scripts written in Python programming language that use standard Python libraries for ML calculations and visualizations.

DATA ANALYSIS

Evaluating the Performance of Classification Methods

We examine three types of ML algorithms (DTs, RFs, and SVMs) to predict whether an individual is likely to recidivate or not. These algorithms can be thought of as classifiers. To evaluate a classifier, we need a way to compare the performance of each one (i.e., a measure that shows how well a given classifier predicts positive and negative cases of recidivism). A natural way to do this is to apply the classifier to a data set where the outcome is known, and hence, the performance of the classifier can be compared with existing data. We use the following quantities to compare performance of various classifiers: TP, the number of cases correctly predicted as positive; false positive (FP), the number of cases incorrectly predicted as positive; TN, the number of cases correctly predicted as negative; and false negative (FN), the number of cases incorrectly predicted as negative. All of these numbers are often summarized in the following matrix:

$$\begin{pmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TN} \end{pmatrix}.$$

The most popular (and yet arguably naïve when used exclusively) measure associated with a classifier is the accuracy (ACC), which is defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

The accuracy tells us that what portion of the testing data is correctly classified. However, such a measure is not without its shortcomings. For example, consider a hypothetical classifier that classifies everything as negative (i.e., no individual is predicted to recidivate). The accuracy of this all-negative classifier applied to our data set is in fact about 65% because the recidivism rate is about 35%. Despite its high success rate, this classifier is likely not acceptable because it never correctly classifies positives (i.e., individuals who recidivate). The high accuracy can be attributed to the low recidivism rate.

Hyperparameters

The hyperparameters of an ML algorithm refer to parameters of the algorithm that must be specified in addition to the data. We experimented with Gini and Entropy versions of DTs and RFs and SVMs with linear, quadratic, cubic, and radial basis function (RBF) kernels. The DT and RF methods with Gini (Marsland, 2015) and the SVM method with Cubic kernel (Steinwart & Christmann, 2008) resulted in the best performance and hence were chosen as hyperparameters and used in all the comparisons with LS/CMI score.

Comparing Overall Performance

We examine whether there is a difference in performance of the ML algorithms relative to LS/CMI score in predicting recidivism. We examine the predictive validity of the LS/CMI and ML algorithms for general recidivism using receiver operating characteristic

(ROC) analyses. ROCs generate an AUC value from 0 to 1, representing the probability that a randomly selected recidivist will obtain a higher score than a randomly selected nonrecidivist (Rice & Harris, 1995). We use the interpretive rubric of Rice and Harris (2005) in which the magnitude of AUC values is mapped to predictive effect sizes as follows: 0.55 to 0.63 (small/low), 0.64 to 0.70 (medium), and 0.71 and up (large/high). AUCs are evaluated by magnitude and in their ability to rank predictive models according to individual LS/CMI scores.

SENSITIVITY ANALYSIS FOR FEATURE SELECTION

ML algorithms such as DT and RF have the capability to identify and report the most influential features in the models they build. In this study, we use sensitivity analysis to elicit the LS/CMI items (or features) that have the most importance predicting recidivism. Sensitivity analysis is generally the study of how perturbations (small changes or uncertainties) in model inputs are propagated to uncertainties in model outputs. Specifically, when a small change to a model input leads to a large change in model output, we say that the model is *sensitive* to that input. There are a number of ways in which sensitivity can be measured. In this study, we consider three of the most popular methods: the Morris method, which performs global sensitivity analysis by making a number of local changes at different possible input values; the Sobol (or variance-based sensitivity analysis) method, which decomposes the variance of the output of the model into fractions and attributes them to inputs; and the moment-independent δ index, which measures the relative importance of an individual input in determining the uncertainty of model output by looking at the entire distribution range of model output.

As a basic usage of sensitivity analysis, one can consider a scenario when two individuals have the same scores and hence the same prediction and ranking based on LS/CMI. In this situation, the values of the top items can be used as additional information to rank and predict their future recidivism. For example, a positive value for top items indicates a higher probability for positive recidivism, and a negative value indicates a lower probability.

RESULTS

The overall rate of recidivism for the two data sets is 31.98%. Table 1 shows the distribution of individuals with respect to the five LS/CMI risk levels and their corresponding rates of recidivism for each data set. A typical interpretation of a row in Table 1 is, for example, that an individual classified as high in D2 is likely to recidivate with probability of 64.85% (and correspondingly will not recidivate with probability of 35.15%). In general, an individual is classified as likely to recidivate if more than 50% of individuals with the same score have done so; otherwise, the individual is classified as unlikely to recidivate. From this table, we also see that the data are imbalanced (i.e., the risk levels do not have equal representation). Figure 1C provides a visual representation of the skew in the population distribution of LS/CMI scores in the combined data set. Figure 1D shows the rate of recidivism for each LS/CMI score of the combined data set. As expected, the recidivism rate shows a steady increase as the LS/CMI score increases. The decrease in the recidivism rate and prediction accuracy for the maximum LS/CMI score (43) is likely due to insufficient data ($n = 7$).

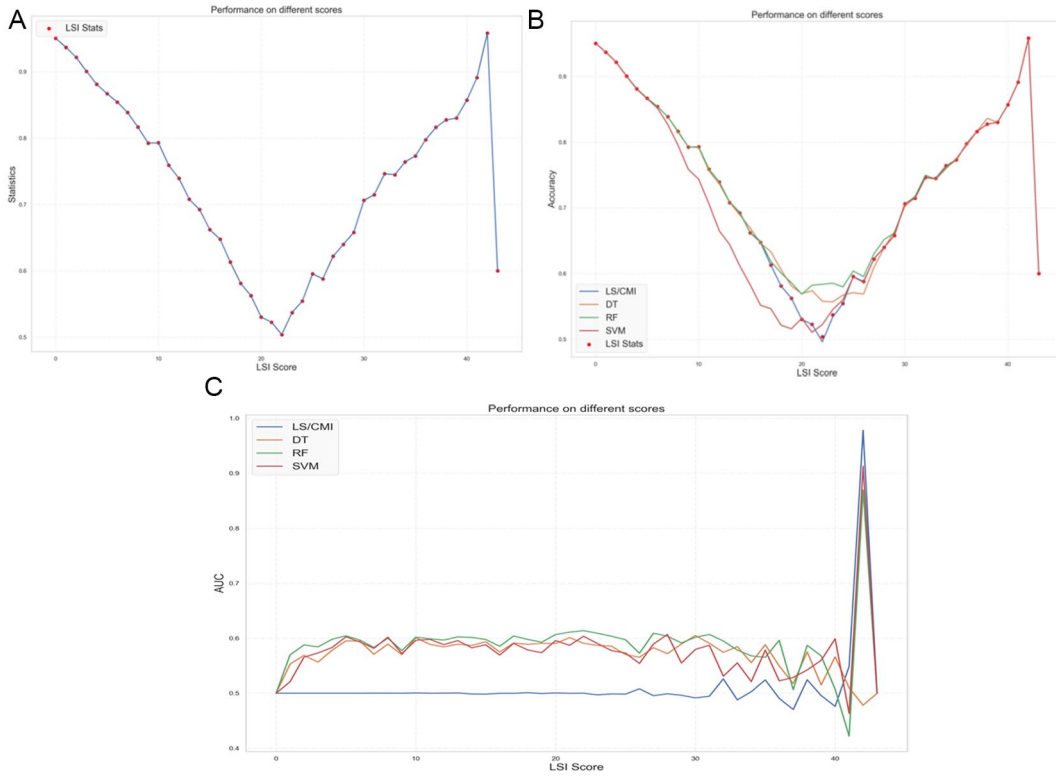

Figure 2: ACC and AUC Performance Metrics for the Four Methods Examined—LS/CMI, DT, RF, and SVM: (A) ACC of LS/CMI Method for Each Score; (B) ACC of the Four Methods, LS/CMI, DT, RF, and SVM, for Each LS/CMI Score; and (C) AUC of the Four Methods—LS/CMI, DT, RF, and SVM for Each LS/CMI Score

Note. LSI refers to LS/CMI = Level of Service/Case Management Inventory; DT = decision tree; RF = random forest; SVM = support vector machine; ACC = accuracy; AUC = area under the receiver operating characteristic curve.

The distribution of the prediction accuracy of LS/CMI scores is depicted in Figure 2A. From this figure, it can be observed that individuals classified as very high risk (LS/CMI scores between 30 and 43) can be confidently regarded as likely to recidivate, and those classified as low or very low risk (LS/CMI scores between 0 and 10) as unlikely to recidivate. However, for the relatively wide range of LS/CMI scores between 14 and 29, the LS/CMI predictive accuracy is below 70%. This decrease in the predictive properties of LS/CMI scores occurs for individuals classified as medium and high risk, and these two risk groups form over half (51.48%) of the individuals in the combined data set.

Next we calculate the performance measures for the prediction of recidivism for the LS/CMI, DT, RF, and SVM methods. Table 2 shows the overall predictive accuracy (ACC) as a weighted average of the accuracies for each score. It can be observed that the overall predictive accuracy of all four methods is comparable, with RF only slightly outperforming LS/CMI. From Figure 2B, it can be observed that all four methods behave similarly as a function of the LS/CMI score, showing high predictive accuracy at the extreme scores and relatively low predictive accuracy for the middle scores. It is noteworthy, however, that the RF method essentially outperforms the LS/CMI method in terms of predictive accuracy

TABLE 2: Performance Measures for Each Prediction Method

Method	ACC	AUC (95% CI)
LS/CMI	0.734	0.7517 [0.7511, 0.7524]
DT	0.695	0.7529 [0.7514, 0.7545]
RF	0.736	0.7531 [0.7519, 0.7545]
SVM	0.704	0.7545 [0.7528, 0.7562]

Note. ACC = accuracy; AUC = area under the receiver operating characteristic curve; CI = confidence interval; LS/CMI = Level of Service/Case Management Inventory; DT = decision tree; RF = random forest; SVM = support vector machine.

over the entire range of LS/CMI scores. In particular, the lowest value of predictive accuracy for RF is approximately 0.57, whereas for LS/CMI, it is slightly below 0.50.

The performance according to AUC is also shown in Table 2. According to the interpretive rubric of Rice and Harris (2005), these AUCs for all four methods correspond to large predictive effect sizes, with extremely small differences in magnitude between the methods. With this in mind, the AUC for LS/CMI total score was slightly lower than the other three methods, and its 95% CI does not overlap with that of SVM, which has the highest AUC, and represents a statistically significant difference.

Figure 2C shows the distribution of AUC values for the different methods tested at each possible individual LS/CMI score. This figure illustrates that ML algorithms do a better job in discriminating recidivists from nonrecidivists compared with traditional LS/CMI summative methods (AUC values often around 0.6 compared with 0.5 for LS/CMI summative score) for a broad range of scores from low to moderately high. From their construction, ML algorithms take into account the way in which the individual items are combined to produce a given total score, and this leads to improved performance compared with simple consideration of total scores. RF has the second-highest AUC and seems to be the most effective method overall in terms of both the ACC and AUC performance metrics.

Figure 3 shows a heatmap of the sensitivities of the LS/CMI items according to the three different sensitivity metrics discussed above and sorted in decreasing order by Sobol index. These analyses can be used to inform prediction of future recidivism because they demonstrate which factors have the most influence in predicting recidivism. From Figure 3, we see that Items A18 (charge laid, probation breached, or parole suspended during prior community supervision), A14 (three or more present offenses), and A423 (could make better use of time), and to a lesser extent, A735 (current drug problem) are the most sensitive items in the LS/CMI according to the Sobol and moment-independent δ indices.

DISCUSSION

ML is often misconstrued as "pitting human minds against the machine" (Ahuja, 2019; Norman, 2018) or equated with completely "automated offender risk assessment" (Wormith, 2017). Neither represents accurate or complete understandings of applications of ML. In our view, ML is a tool or set of techniques that can potentially augment current risk-assessment approaches and assist in understanding behavioral patterns relevant to criminal justice (e.g., criminal recidivism). The results of this study build upon the limited previous findings demonstrating that ML algorithms can perform as well or better than summative scores on validated risk-assessment tools (e.g., Duwe & Kim, 2016); a novel contribution of this

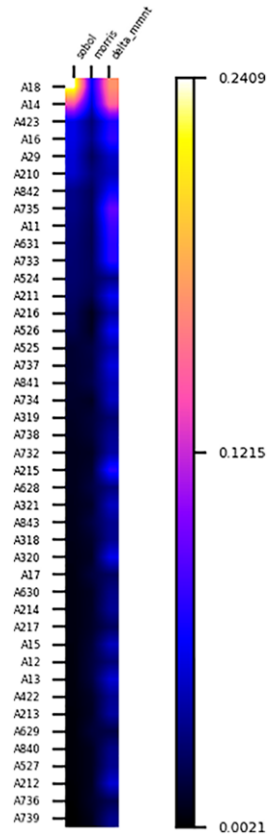

Figure 3: Sensitivities of LS/CMI Items According to Different Metrics
Note. LS/CMI = Level of Service/Case Management Inventory and individual LS/CMI items are numbered along the left-hand side (e.g., A18, A14) with methods used for sensitivity analyses across the top, including Sobol, Morris, and moment-independent δ index.

study is that it features a theoretically driven, fourth-generation, general risk–need tool, the LS/CMI.

Interestingly, classification accuracy as measured by ACC was found to improve significantly for the middle LS/CMI scores using RFs. As an example, for an LS/CMI score of 22, the ACC increased from 0.51 to 0.59 using RFs. The accuracy of the LS/CMI for these middle scores is near 0.50, making it difficult to assess as to whether an individual is likely to recidivate. Because as many as 15% to 20% of individuals may fall into this portion of the "High" risk band, these results suggested that ML algorithms may help us to increase the predictive capability for recidivism among individuals in lower-confidence, higher-density risk classification groupings.

Examination of a more sensitive performance metric (AUCs), which accounts for fluctuations in base rates and considers the relative rankings of scores, also revealed that ML algorithms performed equally as well as total LS/CMI scores, with SVM slightly outperforming LS/CMI score (AUCs = 0.7517 for LS/CMI to 0.7545 for SVM; Table 2), with prediction magnitudes consistent with previous meta-analytic reviews (Olver et al., 2014).

This pattern was observed across all LS/CMI scores with the exception of those in the very high risk range, owing to insufficient *n* at the most extreme scores (e.g., maximum score of 43). Moreover, all the ML approaches investigated consistently improved AUC by approximately 10 percentage points for the majority of individual LS/CMI scores (Figure 2C). These AUCs may be further improved with incorporation of additional risk-relevant and dynamic data beyond LS/CMI scores as recommended by Garb and Wood (2019).

It is important to note that LS/CMI summative scoring makes an identical prediction for two individuals with similar scores, say a score of 22. However, mathematically, the number of distinct ways an individual can be assigned the summative score of 22 is approximately 1 trillion. In fact, the LS/CMI method does not distinguish among any of these different cases, whereas an ML algorithm like SVM can provide a more nuanced analysis and could differentiate between individuals with a summative score of 22 as likely to recidivate, and others not, depending upon how the score was reached. Thus, these preliminary results suggest that there may be underlying patterns or different combinations of scores that are more predictive of recidivism. Further analyses of these patterns may be fruitful, not only in terms of predictive accuracy but also to identify clusters and weightings of criminogenic needs that may separate recidivists from nonrecidivists with similar LS/CMI scores, including frequently obtained and seemingly less predictive risk scores. Results of exploratory sensitivity analyses have begun to identify dynamic factors or criminogenic risk–needs that may have the most influence in the prediction of recidivism (e.g., poor use of time, current drug problem). Our plan for future work is to conduct additional mathematical and statistical analyses on the most common combinations of these items and the number of unique paths resulting in a specific summative LS/CMI score as well as to include available features beyond LS/CMI items.

Mere "prediction" should not be the primary goal of any risk assessment. Rather, prevention is the primary purpose (e.g., risk reduction). Thus, all predictive technology is perhaps best viewed as preventive technology and this includes ML. To our knowledge, there are few studies that have evaluated "real world" applications of ML. In one such illustrative study, Berk (2016) examined the impact of ML "risk forecasts" on Parole Board decisions in Pennsylvania. Although some evidence for "smarter decision making" was reported, it was difficult to ascertain the full implications of the evidence because standard practices and ML approaches were drawing upon much of the same information by virtue of the fact that ML "forecasts were meant to supplement the information available to the Board, not replace it" (p. 22). This is one of many possible applied uses of ML data to augment, not replace, other decision-making tools and mechanisms. No one tool, computer-aided or not, should be used in a standalone fashion to inform criminal justice decision-making, but rather use of validated tools as part of comprehensive and contextualized assessments is required as part of best practices. Moreover, as we seek to integrate additional information into risk assessments, such as individual strengths or protective factors or changes in dynamic risk over time, ML approaches may make better use of enriched information (e.g., repeated or multiple assessments incorporating risk, protective factors, and change information), and they could also uncover relationships between risk-relevant variables and outcomes that may differ across groups, settings, and time.

Furthermore, ML may also provide new insights that can inform intervention practices. For instance, Lussier et al. (2019) recently used DT algorithms to identify risk factors for entry into sexual reoffending. Future work may be able to employ ML to test hypotheses

regarding changes in risk over time and test causal inferences (Barabas et al., 2018), including the effects of intervention such as correctional programming and resultant changes in risk–need scores. When used in this manner, ML approaches could have utility in preventing criminal or aggressive behavior, including identifying situations that may lead to violence toward self and others. For example, Bala and Truatman (2019) have recently explored the applicability of ML approaches to promote identification and intervention for individuals in custody who may be at risk for engaging in self-harming behaviors.

STUDY LIMITATIONS AND FUTURE DIRECTIONS

Like all tools, we would encourage further evaluation of ML approaches and advocate for their responsible use. As cautioned by Barabas et al. (2018), some ML approaches "transform the space of input features into a higher order space that is often difficult to interpret" (p. 8). Clear interpretations must be advanced and tested, and such analyses should include local validation and updated models. There is also the important issue of algorithmic fairness (Berk et al., 2018; Corbett-Davies & Goel, 2018). It has been argued that both risk-assessment tools and models may be biased for a variety of reasons (e.g., label bias, feature bias, sample bias, and calibration issues); however, the very data used to train and test ML algorithms may also be compromised, and one must avoid retrenching biases (Corbett-Davies & Goel, 2018). Although ML approaches require large sample sizes, "big data" are not necessarily "deep." The current study examined data from a single source (i.e., LS/CMI scores). Integrating data from other sources may improve models, advance understanding, and reduce inherent biases. Recently, Menger and colleagues (2019) endeavored to predict in-patient violence from clinical notes in patient electronic health records. As such, novel data sources and data elements (e.g., text data) could be integrated with traditional data sources (e.g., risk scores) to enhance statistical models and further our understanding of behavioral patterns relevant to criminal justice outcomes (e.g., recidivism and desistance).

Finally, although two large data sets of LS/CMI administrations were utilized to retain maximal data for exploratory analyses, it is recognized that as field research, there is a lack of uniformity between the samples (e.g., differences in available follow-up time and mean LS/CMI score). However, healthy sampling variance was observed, and use of techniques robust to fluctuations in base rate was employed. This said, more nuanced analyses beyond the scope of the present work could examine the associations between important moderators and recidivism that may have bearing on LS/CMI score and outcome. For instance, ML approaches may further contribute to our understanding of the mechanisms that may underlie the relative superiority of discrimination of item patterns by gender, age, or other risk moderators. Further examinations of underlying risk–need patterns (e.g., pathways of criminogenic needs) using ML may also assist practitioners to refine prevention and correctional strategies based on the patterns observed in the data.

CONCLUSIONS FOR THE "NEAR FUTURE" OF RISK–NEED ASSESSMENT

Always looking to push the field of risk assessment forward, Andrews, Bonta, and Wormith discussed "The Recent Past and Near Future of Risk and/or Need Assessment" in their seminal 2006 paper. Over 10 years later, Wormith (2017) provided additional glimpses into the future of risk assessment in his policy paper titled "Automated Offender Risk

Assessment: The Next Generation or a Black Hole?" Risk assessment in the digital era, use of artificial intelligence in criminal justice, and "smart prisons" are no longer the near future—they are the present. ML approaches are now making significant contributions to health care, not to mention business and entertainment, and there have been calls to build "fair algorithms" to assist criminal justice decision-making (Corbett-Davies & Goel, 2018). Recent preliminary findings suggest that although ML approaches can contribute meaningfully to risk assessment, management, and reduction, they should be developed with care. With smart, automated technologies advancing at "warp speed" (Wormith, 2017, p. 281), research and statistical methodologies must keep pace to support ethical, effective, and cost-efficient correctional practices; promote innovation in risk assessment and management; and ultimately, better, safer outcomes for criminal justice clients and communities.

ORCID iD

Raymond J. Spiteri https://orcid.org/0000-0002-3513-6237

REFERENCES

Ahuja, A. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *7*, Article e7702. https://doi.org/10.7717/peerj.7702

Andrews, D. A., & Bonta, J. (1994). *The psychology of conduct* (1st ed.). Anderson.

Andrews, D. A., & Bonta, J. (2010). *The psychology of conduct* (5th ed.). LexisNexis.

Andrews, D. A., Bonta, J., & Wormith, J. S. (1995). *Level of Service Inventory–Ontario Revision (LSI-OR): Interview and scoring guide*. Ontario Ministry of the Solicitor General and Correctional Services.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *Level of Service/Case Management Inventory (LS/CMI): An offender assessment system. User's guide*. Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or needs assessment. *Crime and Delinquency*, *52*, 7–27. https://doi.org/10.1177/0011128705281756

Baglivio, M. T. (2009). The assessment of risk to recidivate among a juvenile offending population. *Journal of Criminal Justice*, *37*(6), 596–607. https://doi.org/10.1016/j.jcrimjus.2009.09.008

Bala, N., & Truatman, L. (2019). "Smart" technology is coming for prisons, too. *Slate*. https://slate.com/technology/2019/04/smart-ai-prisons-surveillance-monitoring-inmates.html

Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *Proceedings of Machine Learning Research*, *81*, 1–15.

Berk, R. A. (2016). *An impact assessment of machine learning risk forecasts on parole board decisions and recidivism* (Working Paper No. 2016-4.0). University of Pennsylvania.

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, *12*, 513–544. https://doi.org/10.1111/1745-9133.12047

Berk, R. A., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research*. Advance online publication July 2018. https://doi.org/10.1177/0049124118782533

Brennan, T., & Oliver, W. L. (2013). The emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions. *Criminology & Public Policy*, *12*, 551–562. https://doi.org/10.1111/1745-9133.12055

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce (Ed.), *The workings of the indeterminate sentence law and the parole system in Illinois* (pp. 221–234). Illinois State Board of Parole.

Bushway, S. D. (2013). Is there any logic to using logit: Finding the right tool for the increasingly important job of risk prediction. *Criminology & Public Policy*, *12*, 563–567. https://doi.org/10.1111/1745-9133.12059

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. https://arxiv.org/abs/1808.00023.

Curtis, J. (2018). *On using machine learning to predict recidivism* [Unpublished doctoral dissertation, Texas Tech University].

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*, 1920–1930. https://doi.org/10.1161%2FCIRCULATIONAHA.115.001593

Duwe, G. (2014). The development, validity, and reliability of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR). *Criminal Justice Policy Review*, *25*, 579–613. https://doi.org/10.1177%2F0887403413478821

Duwe, G. (2019). Better practices in the development and validation of recidivism risk assessments: The Minnesota Sex Offender Screening Tool-4. *Criminal Justice Police Review*, *30*, 538–564. https://doi.org/10.1177/0887403417718608

Duwe, G., & Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections Policy, Practice, and Research*, *1*, 155–176. https://doi.org/10.1080/23774657.2016.1178083

Duwe, G., & Kim, K. (2017). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review*, *28*, 570–600. https://doi:10.1177/0887403415604899

Garb, H. N., & Wood, J. M. (2019). Methodological advances in statistical prediction. *Psychological Assessment*, *31*, 1456–1466. http://dx.doi.org/10.1037/pas0000673

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of predictors of adult offender recidivism: What works! *Criminology*, *34*, 401–433. https://doi.org/10.1111/j.17459125.1996.tb01220.x

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hamilton, Z., Neuilly, M., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology*, *11*, 299–318. https//doi.org/10.1007/s11292-014-9221-8

Hogg, S. M. (2011). *The Level of Service Inventory (Ontario Revision) scale validation for gender and ethnicity: Addressing reliability and predictive validity* [Unpublished master's thesis, University of Saskatchewan].

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today*, *20*, 318–331. https://doi.org/10.1016/j.drudis.2014.10.012

Lin, Z., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*, Article eaaz0652. https://doi/org/10.1126/sciadv.aaz0652

Liu, Y. Y., Yang, M., Ramsey, M., Xiao, S. L., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*, *27*, 547–573. https://doi.org/10.1007/s10940-011-9137-7

Luong, D., & Wormith, J. S. (2011). Applying risk/need assessment to probation practice and its impact on the recidivism of young offenders. *Criminal Justice and Behavior*, *38*, 1177–1199. https://doi.org/10.1177/0093854811421596

Lussier, P., Deslauriers-Varin, N., Collin-Santerre, J., & Bélanger, R. (2019). Using decision tree algorithms to screen individual at risk of entry into sexual recidivism. *Journal of Criminal Justice*, *63*, 12–24. https://doi.org/10.1016/j.jcrimjus.2019.05.003

Marsland, S. (2015). *Machine learning: An algorithmic perspective*. (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17476

Menger, V., Spruit, M., van Est, R., Nap, E., & Scheepers, F. (2019). Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *Journal of the American Medical Association Network Open*, *2*, Article e196709. https://doi.org/10.1001%2Fjamanetworkopen.2019.6709

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Norman, A. (2018, January 31). Your future doctor may not be human. This is the rise of AI in medicine. *Futurism*. https://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwjGv5jWjofpAhXKQc0KHTm-CN8QFjAAegQIARAB&url=https%3A%2F%2Ffuturism.com%2Fai-medicine-doctor&usg=AOvVaw3W-WB1mvsjKAtpwL1ag4XV

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior*, *36*, 329–353. http://doi.org/10.1177/0093854809331457

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the level of service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*, 156–176. https://doi.org/10.1037/a0022200

Orton, L. C. (2014). *An examination of the professional override in the Level of Service Inventory–Ontario Revision (LSI-OR)* [Unpublished master's thesis, University of Saskatchewan].

Ozkan, T., Clipper, S. J., Piquero, A. R., Baglivio, M., & Wolff, K. (2020). Predicting sexual recidivism. *Sexual Abuse*, *32*, 375–399. https://doi.org/10.1177/1079063219852944

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*(5), 737–748. https://doi.org/10.1037/0022-006X.63.5.737

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, *29*, 615–620. https://doi.org/10.1007/s10979-005-6832-7

Ridgeway, G. (2013). Linking prediction and prevention. *Criminology & Public Policy*, *12*, 545–550. https://doi.org10.1111/1745-9133.12057

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. The MIT Press.

Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy*, *8*, 183–208. https://doi.org/10.1111/j.1745-9133.2009.00551.x

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P., Robbins, P. C., Mulvey, E. P., Grisso, T., Roth, L. H., & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, *24*, 83–100. https://doi.org/10.1023/A:1005478820425

Steinwart, I., & Christmann, A. (2008). *Support vector machines* (1st ed.). Springer. https://doi.org/10.1007/978-0-387-77242-4

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society*, *176*, 565–584. https://doi.org/10.1371%2Fjournal.pone.0213245

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Viljoen, J. L., Mordell, S., & Beneteau, J. (2012). Prediction of adolescent sexual reoffending: A meta-analysis of the J-SOAP-II, ERASOR, JSORRAT-II, and Static-99. *Law and Human Behavior*, *36*, 423–438. https://doi.org/10.1037/h0093938

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR 20. Assessing risk for violence. Version 2*. Mental Health, Law and Policy Institute.

Wilson, H. A., & Gutierrez, L. (2014). Does one size fit all?: A meta-analysis examining the predictive ability of the Level of Service Inventory (LSI) with aboriginal offenders. *Criminal Justice and Behavior*, *41*(2), 196–219. http://doi.org/10.1177/0093854813500958

Wormith, J. S. (2011). The legacy of D. A. Andrews in the field of criminal justice: How theory and research can change policy and practice. *International Journal of Forensic Mental Health*, *10*, 78–82. https://doi.org/10.1080/14999013.2011.577138

Wormith, J. S. (2017). Automated offender risk assessment: The next generation or a black hole? *Criminology & Public Policy*, *16*, 281–303. https://doi.org/10.1111/1745-9133.12277

Wormith, J. S., & Bonta, J. (2017). The Level of Service (LS) instruments. In J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais, & Z. Hamilton (Eds.), *Handbook of recidivism risk/need tools* (pp. 117–145). John Wiley.

Wormith, J. S., Hogg, S. M., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, *39*, 1511–1538. https://doi.org/10.1177/0093854812455741

Wormith, J. S., Hogg, S. M., & Guzzo, L. (2015). The predictive validity of the LS/CMI with Aboriginal Offenders in Canada. *Criminal Justice and Behavior*, *42*, 481–508. https://doi.org/10.1177%2F0093854814552843

Mehdi Ghasemi works as a senior scientist at Edmonton Police Service and an adjunct professor in the Department of Math & Stats at the University of Saskatchewan. His scientific activities mainly include mathematical modeling, optimization, and advanced data analysis.

Daniel Anvari is a faculty member in the Department of Mathematics and Statistics at Kwantlen Polytechnic University. His areas of research are applications of dynamical systems and machine learning in biology, health, and social sciences.

Mahshid Atapour is a faculty member in the Department of Mathematics and Statistics at Capilano University. Her areas of research are applied probability and applications of statistics and machine learning in health and social sciences.

J. Stephen Wormith (now deceased) began his career as a psychologist and researcher in various correctional jurisdictions in Canada. He then became a professor in the psychology department at the University of Saskatchewan and also the director of the Centre of Forensic Behavioural Science and Justice Studies. Over his long career, he made fundamental contributions to offender risk and psychological assessment, offender treatment, sexual offenders, and crime prevention. He was a fellow of the Canadian Psychological Association (CPA) and represented the CPA on the National Associations Active in Criminal Justice (NAACJ).

Keira C. Stockdale is a registered clinical psychologist currently employed by the Saskatoon Police Service and an adjunct professor in the Department of Psychology at the University of Saskatchewan. Her research and clinical activities include risk assessment and treatment for justice-involved youth and adults.

Raymond J. Spiteri is a professor in the Department of Computer Science at the University of Saskatchewan. His areas of research are numerical analysis, scientific computing, and high-performance computing with specialization in time-stepping methods for differential equations.