Journal of Classification 35:147-171 (2018)
DOI: 10.1007/s00357-018-9252-3

The Lack of Cross-Validation Can Lead to Inflated Results and Spurious Conclusions: A Re-Analysis of the MacArthur Violence Risk Assessment Study

Ehsan Bokhari

University of Illinois at Urbana-Champaign

Lawrence Hubert

University of Illinois at Urbana-Champaign

Abstract: Cross-validation is an important evaluation strategy in behavioral predictive modeling; without it, a predictive model is likely to be overly optimistic. Statistical methods have been developed that allow researchers to straightforwardly cross-validate predictive models by using the same data employed to construct the model. In the present study, cross-validation techniques were used to construct several decision-tree models with data from the MacArthur Violence Risk Assessment Study (Monahan et al., 2001). The models were then compared with the original (non-cross-validated) Classification of Violence Risk assessment tool. The results show that the measures of predictive model accuracy (AUC, misclassification error, sensitivity, specificity, positive and negative predictive values) degrade considerably when applied to a testing sample, compared with the training sample used to fit the model initially. In addition, unless false negatives (that is, incorrectly predicting individuals to be nonviolent) are considered more costly than false positives (that is, incorrectly predicting individuals to be violent), the models generally make few predictions of violence. The results suggest that employing cross-validation when constructing models can make an important contribution to increasing the reliability and replicability of psychological research.

Keywords: Classification trees; Cross-validation; Replicability; Misclassification costs; Random forests; Violence prediction.

Ehsan Bokhari is now a Senior Analyst with the Los Angeles Dodgers in Los Angeles, California.

Corresponding Author's Address: 1000 Vin Scully Ave, Los Angeles, CA 90012, e-mail: ebokhari@msn.com.

Published online: 26 March 2018

1. Introduction

Cross-validation is an important part of constructing a behavioral predictive model. A failure to cross-validate may lead to inflated and overly-optimistic results, as Meehl and Rosen (1955) noted some sixty years ago: "If a psychometric instrument is applied solely to the criterion groups from which it was developed, its reported validity and efficiency are likely to be spuriously high" (p. 194). The Classification of Violence Risk (COVR; Monahan et al., 2001) assessment tool is an actuarial device designed to predict the risk of violence in psychiatric patients. The COVR is a computer-implemented program based on a classification tree construction method that has been praised for its "ease of administration" (McDermott, Dualan, and Scott, 2011, p. 4). When constructed, however, the COVR was not cross-validated; thus, the results from the construction sample may be overly optimistic (for example, see McCusker, 2007).

The research presented in this paper reanalyzes data from the MacArthur Violence Risk Assessment Study (VRAS) used to develop the COVR. We begin by describing a widely-applied method for cross-validation, commonly called K-fold cross-validation. Data are then presented from the MacArthur VRAS. Several classification tree models are built from the VRAS dataset demonstrating the importance of cross-validation. In addition, we show how differing cutscores (see Appendix A in Supplementary Material online) implicitly affect the costs associated with false negatives and positives. The COVR implicitly assumes that false negatives (incorrect classifications of violent individuals) are more costly than false positives (incorrect classifications of nonviolent individuals).

2. A Brief Introduction to Cross-Validation

Cross-validation is an important tool for prediction, allowing the researcher to estimate the accuracy of a prediction tool in practice. Assessing the accuracy of a model with the same data used to create the model will give overly optimistic estimates of accuracy because a model is typically fit by minimizing some measure of inaccuracy; thus, the model reflects both the true data pattern as well as error. Cross-validation is a strategy to separate these two entities.

Assume we have a dataset $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ is an $n \times p$ matrix containing n observations measured across p predictor variables, and $\mathbf{y}$ is an $n \times 1$ vector containing n observations measured on a single outcome variable (for example, the outcome of whether an act of violence was committed). In this scenario, the outcome variable is known, and the construction of a model to predict the known values of $\mathbf{y}$ is typically referred to as *supervised learning*.

In prediction, interests generally center on modeling $\mathbf{y}$ as a function of $\mathbf{X}$, where it is assumed that for some function f,

$$\mathbf{y} = f(\mathbf{X}) + \varepsilon;$$

here, ε represents an $n \times 1$ vector of random error terms, assumed to have mean 0, finite variance, and be uncorrelated with the set of predictor variables. The primary goal is to estimate $f(\mathbf{X})$ so that a practical classification function,

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}),$$

is constructed. The total error in prediction, $\mathbf{y} - \hat{\mathbf{y}}$, can be divided into two types of components: *reducible error* and *irreducible error*. Decomposing the mean-squared error gives

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] =& \ \mathbb{E}\left[(f(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X}))^2\right] \\
=& \ \mathbb{E}\left[(f(\mathbf{X}) - \hat{f}(\mathbf{X}))^2\right] + 2\mathbb{E}\left[(f(\mathbf{X}) - \hat{f}(\mathbf{X}))\varepsilon\right] + \mathbb{E}(\varepsilon^2) \\
=& \ \underbrace{\mathbb{E}\left[(f(\mathbf{X}) - \hat{f}(\mathbf{X}))^2\right]}_{\text{reducible error}} \ + \ \underbrace{\mathbb{V}(\varepsilon),}_{\text{irreducible error}}
\end{aligned}$$

where $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ represent the expected value and variance, respectively. These two types of error determine the accuracy of predictions. Typically $\mathbb{V}(\varepsilon)$ is unknown and hence cannot be reduced; the reducible error can be minimized, which is of course the goal of predictive modeling. When the predicted function perfectly matches the true function, the total mean squared error equals $\mathbb{V}(\varepsilon)$; thus, $\mathbb{V}(\varepsilon)$ represents a lower bound for the total error. In practice, $f(\mathbf{X})$ is not known, so one can only strive to get close to this lower bound by constructing predictive models that produce the smallest total error values.

In constructing a predictive model, there are several available measures for assessing how well the model fits the data (for example, the mean squared error, the coefficient of determination [R^2], the proportion of predictions correct). It is important to note that this assessment of error is often made with the sample used to construct the model and not on predictions with an independent sample. When a model is constructed for purposes of prediction, the model's predictive accuracy on new data is most relevant. Suppose there is a given measure of accuracy, say γ, for assessing the model and this measure is obtained with the same data used to construct the model. A way of evaluating a model's predictive ability is to collect new data and measure how accurate the predictions are; that is, a new accuracy measure γ' is obtained. The difference between γ and γ' represents the drop in how well the model predicts (assuming that a larger γ is associated with better accuracy, typically, $\gamma - \gamma' > 0$); this drop is known as *shrinkage*. Rather than assessing predictive accuracy with the same data relied on to build the model, the original data can be randomly split into two parts: the *training data* and the *testing data*. The training data is for constructing the model; the testing data is for estimating the predictive accuracy of the model. This process is typically more efficient in terms of time and cost than collecting new data after the model is developed, and can help prevent overfitting.

2.1 K-fold Cross Validation

Given n observations, cross-validation involves splitting the data so that a specified proportion, say q, of the data is present in the training set and the rest of the observations are in the testing set (that is, qn observations are in the training set and $(1-q)n$ are in the testing set). This can be carried out multiple times, choosing a different training and testing set each time.

K-fold cross validation is one of the more popular cross-validation strategies and will be the only one discussed. K-fold cross-validation involves splitting the data into K subsets; the training set consists of the union of $K-1$ subsets, and the testing set is defined by the remaining observations. This process is repeated so that each subset acts once as the testing sample. Because each replication of this process will produce results that vary, it is common to compute an average across all replications. The simplest form of K-fold cross validation is to let $K = 2$: the training set contains half of the observations, and the testing set the other. The most computationally costly form is to let $K = n$, so that each observation acts as the testing sample; this is commonly known as *leave-one-out cross-validation*. In addition to the computational costs, another disadvantage of leave-one-out cross-validation is that the variance of the estimate can be relatively large compared to other estimates—it is, however, approximately unbiased; setting $K < n$ provides an estimate of the test error with less variance but more bias (see Hastie, Tibshirani, and Friedman, 2009, pp. 242–243). A reasonable choice for K is commonly considered to be ten (Breiman and Spector, 1992). Appendix B (in Supplementary Material online) presents a thorough discussion of what is known as the “bias-variance trade-off” in prediction, and how cross-validation can help determine the equilibrium point that simultaneously minimizes both the bias and the variance of a model.

3. Decision Trees

Decision trees, commonly referred to as *Classification and Regression Trees* (Breiman et al., 1984), are popular statistical learning techniques generally used for prediction. Consider an $n \times p$ data matrix, $\mathbf{X}$, containing n observations measured across p predictor variables, and an $n \times 1$ vector $\mathbf{y}$ containing n observations measured across a single outcome variable. The outcome variable contained in $\mathbf{y}$ is the variable of interest with respect to prediction. When the outcome variable is continuous, regression trees are constructed; when categorical, classification trees are constructed. For violence prediction in general, the outcome variable is binary (that is, categorical with two classes) representing the presence or absence of an act of violence; because of this specification, our focus is solely on classification trees.

Classification trees are constructed by first splitting the data into two disjoint subsets based on one of the p predictor variables. Within each subset, further partitioning of the data is done, and within the resulting subsets this process continues until some user-specified stopping criterion is reached; the complete procedure is known as *recursive partitioning*. Recursive partitioning is a *top-down greedy* algorithm: top-down because the algorithm begins with a tree with no splits and works "down" to a tree with many splits (and once a split is made, it remains); greedy because each split made is the "best" conditioned on the given splits (and not on possible future splits). An observation that falls into a subset with no further splits (called a *terminal node*) is classified based on all the observations within that subset, which is typically the modal observation (that is, the most prevalent outcome within the node), but other choices are possible, as will be discussed shortly.

The first split occurs at the *root node* of the tree; extending *branches* from the root node lead to subsample nodes, called *leaves*. As mentioned, the splits continue until a specified criterion is met, such as a constraint on the minimum number of observations in a given leaf, or a criterion based on significance testing. After a tree is constructed, it can be *pruned* to reduce the number of branches, eliminating those that add less to the tree's predictive ability. The notion behind pruning is to create a subtree that has better predictive accuracy on new data, and thus, the level of pruning is commonly determined by cross-validation.

The data are subjected to splitting with the goal of grouping the observations so as to minimize the number of observations incorrectly classified. There are several ways to assess goodness of fit in classification, one being the *Gini index* (Breiman et al., 1984; Gimli, 1912). Given R classes for an outcome variable, the Gini index for a group of observations is given by

$$G = \sum_{r=1}^{R} g_r(1 - g_r) = 1 - \sum_{r=1}^{R} g_r^2,$$

where g_r ($r = 1, \ldots, R$) is the proportion in the group from the rth class. The Gini index can be thought of as a measure of impurity. Note that when

$g_r = 1$, $1 - g_r = 0$, and $G = 0$ indicating perfect purity. When $g_1 = \cdots = g_R = 1/R$ (that is, the proportion of observations are evenly divided among the R classes), the index is maximal at $G = 1 - \frac{1}{R}$. The Gini index is commonly used for determining the splits of the trees, where the split minimizing the Gini index is chosen at each step of the partitioning.

Decision trees are popular because they are easy to interpret, but they are not the best statistical learning method in terms of predictive accuracy. Predictive accuracy can be enhanced by various ensemble methods such as *tree bagging* and *random forests* (Breiman, 2001; these are more generally known as *random subspace methods*). *Bagging* (the term *bag* is a short-hand phrase for bootstrap aggregation; see Breiman, 1996) is an ensemble learning method designed to avoid the overfitting of a model and is commonly used with classification trees (that is, tree bagging). Suppose a dataset, $\mathbf{X}$, contains n observations on p predictors. Similar to the bootstrap method, B training sets of size L are generated, where $1 \leq L \leq n$, by randomly sampling (with replacement) from $\mathbf{X}$; each training set is fit by the model and, after aggregating, an average over the B replications provides a predicted response for each observation.

Note that some of the observations in the ith training set, $\mathbf{B}^{(i)}$, may be duplicate. The larger L is, the more likely there will be at least one duplicate observation; the probability of such an event is $1 - \frac{n!}{n^L(n-L)!}$. The probability that any given observation is not selected is $(1 - \frac{1}{n})^L$. If $L = n$ and as $n \to \infty$, the probability approaches $e^{-1} \approx .37$. For a large enough n and when the training sample is equal to n, it would be expected that on average, about 63% of the bootstrap sample consists of unique observations. The 63% represents a probabilistic lower bound; for $L < n$, one would expect more than 63% of the sample to be unique (for example, in the trivial case where $L = 1$, there are no duplicate observations). The approximately 37% of the observations not used in fitting the model on the ith replication are called *out-of-bag* (OOB); thus, the OOB observations are the testing set and can be used to assess predictive accuracy. For any given observation, by aggregating over the subset of B replications—where the observation was not used to fit the model—the average OOB prediction accuracy can be calculated and compared to the cross-validated error; this comparison gives us the average OOB error difference. The OOB errors can also be used to assess the importance of predictors by randomly permuting the OOB data across variables one at a time, and estimating the OOB error after permutation—a large increase in the OOB error indicates the variable's importance in the model.

Random forests (Breiman, 2001) are tree bagging methods that randomly select a subset of predictors to be used at each split. The advantage here is that it can "decorrelate" the trees by preventing a single variable from dominating the analysis; for instance, if one predictor is very strong, it will likely be the root node for a majority of the trees constructed; the subsequent nodes will be similar as well (that is, the trees will be highly correlated). By convention (and default in statistical software such as R), $\sqrt{p}$ predictors are randomly selected at each node for classification trees.

The advantage of ensemble tree models is that they tend to reduce the variance found in single decision tree models, leading to more accurate results by aggregating over a number of single decision trees. For more information on decision trees and other statistical learning models, the reader is referred to Kuhn and Johnson (2013), Hastie et al. (2009), James et al. (2013); the latter two references are freely available online.

Here, a classification tree was first developed similar to that done in Monahan et al. (2001); next random forests were built to create better classification tree models with better predictive accuracy. All classification tree models were constructed in R (R Core Team, 2014). Before proceeding to the results, however, it is necessary to discuss how a classification tree classifies observations and how this process is related to the costs of false positives and negatives.

4. Misclassification Costs

In general, there are two types of misclassifications that are of concern: false positives and false negatives. A false negative incorrectly predicts the absence of whatever the model is designed to predict (for example, predicting nonviolence in a violent individual); a false positive incorrectly predicts the presence (for example, predicting violence in a nonviolent individual). These two types of misclassifications can have drastically different consequences, and one may assign differing costs to each.

Suppose in a given terminal node there are n observations, of which n_r are from class r ($r = 1, \ldots, R$). An observation is classified into class r based on the modal class at that given terminal node; thus, if $n_r > n_{r'}$ for all $r \neq r'$, all observations within the terminal node are classified as belonging to class r. The empirical posterior probability for each class can be defined as the number of observations in the terminal node coming from a particular class divided by the total number of observations; thus, the estimated posterior probability is

$$\hat{P}(r|\mathbf{x}) = \frac{n_r}{n},$$

where $\mathbf{x}$ is a vector of predictor variables associated with the observation. Given this definition, an observation is classified as coming from class r when $\hat{P}(r|\mathbf{x}) > \hat{P}(r'|\mathbf{x})$ for all $r \neq r'$.

As an addition to the classification process, costs can be assigned to misclassifications; the cost function is labeled $C_s(r)$ and represents the cost of classifying an observation into class s when it truly belongs in class r (note that $C_r(r) = 0$). By including a cost function, an observation is classified into class r by minimizing

$$\sum_{r=1}^{R} \hat{P}(r|\mathbf{x})C_s(r)$$

across all s. Note that when $C_s(r)$ is the same for all $r = 1, \ldots, R$ (that is, the costs are equal across all classes), the previous situation obtains and an observation is classified based on the modal class.

Given two classes (that is, $r = 1, 2$, where 1 could represent nonviolent individuals and 2, those who are violent), an observation is classified into class $r = 2$ when

$$\hat{P}(2|\mathbf{x})C_1(2) > \hat{P}(1|\mathbf{x})C_2(1).$$

With respect to classification of nonviolent and violent individuals, $C_1(2)$ and $C_2(1)$ are, respectively, the costs associated with a false negative and a false positive. Alternatively, the above inequality can be written as

$$\frac{C_1(2)}{C_2(1)} > \frac{\hat{P}(1|\mathbf{x})}{\hat{P}(2|\mathbf{x})}.$$

The lower bound, $\frac{\hat{P}(1|\mathbf{x})}{\hat{P}(2|\mathbf{x})}$, is the conditional odds in favor of the event 1 because $\hat{P}(2|\mathbf{x}) = 1 - \hat{P}(1|\mathbf{x})$; for example, the odds in favor of an individual not being violent, given the data.

If $C_1(2) = C_2(1)$, an observation is classified as coming from class 2 when $\hat{P}(2|\mathbf{x}) > \hat{P}(1|\mathbf{x})$, or equivalently,

$$\frac{\hat{P}(2|\mathbf{x})}{\hat{P}(1|\mathbf{x})} > 1.$$

Bayes Theorem allows this to be rewritten as

$$\frac{\frac{\hat{P}(\mathbf{x}|2)\hat{P}(2)}{\hat{P}(\mathbf{x})}}{\frac{\hat{P}(\mathbf{x}|1)\hat{P}(1)}{\hat{P}(\mathbf{x})}} = \frac{\hat{P}(\mathbf{x}|2)\hat{P}(2)}{\hat{P}(\mathbf{x}|1)\hat{P}(1)} > 1.$$

Considering $\hat{P}(\mathbf{x}|2)$ and $\hat{P}(\mathbf{x}|1)$ fixed, the classification cutscore can be changed by adjusting $\hat{P}(1)$ and $\hat{P}(2)$; these probabilities are the sample base rates (note that for $r = 1, 2$, $\hat{P}(2) = 1 - \hat{P}(1)$). Thus, adjusting the prior probabilities is an equivalent way of adjusting costs.

As will be discussed shortly, Monahan et al. (2001) suggested the cutscore for classification of high-risk individuals as twice the sample base rate of violence (approximately .37). This implies that an individual is classified as violent when the individual belongs to a terminal node where $\hat{P}(2|\mathbf{x}) > .37$, and implicitly assigns unequal costs to false positives and negatives. Explicitly, let $\hat{P}(2|\mathbf{x}) = .37$ (and consequently, $\hat{P}(1|\mathbf{x}) = .63$) so

$$\frac{C_1(2)}{C_2(1)} = \frac{\hat{P}(1|\mathbf{x})}{\hat{P}(2|\mathbf{x})} = \frac{.63}{.37} = 1.67.$$

By lowering the cutscore to .37 for classification of violence, the authors imply that, given the specified prior probabilities, incorrectly classifying an individual as nonviolent (a false negative) is 1.67 times worse than incorrectly classifying an individual as violent (a false positive).

Most authors of actuarial measures for violence risk assessment are reluctant to discuss the costs of false positives versus false negatives (Mossman , 2006, 2013: Vrieze and Grove, 2008); an exception to this is Richard Berk. In his book, *Criminal Justice Forecasts of Risk: A Machine Learning Approach* (Berk, 2012), he suggests that

> *the costs of forecasting errors need to be introduced at the very beginning when the forecasting procedures are being developed* [original emphasis]. Then, those costs can be built into the forecasts themselves. The actual *forecasts* [original emphasis] need to change in response to relative costs. (p. 20)

In the examples that Berk provides (regarding parole release), he suggests that the ratio of false negatives to false positives be as high as twenty to one (also, see Berk, 2011). The reasoning for such an extreme ratio, as justified by Berk (2012), is that the agency the model was built for was “very concerned about homicides that could have been prevented” (p. 5). Thus, the “agency” was willing to accept that a large number of potentially nonviolating parolees were not granted parole; the proportion of those predicted to fail that actually did was only about .13 for the sample data used in the text (see Table 1.1 in Berk, 2012)).

5. The MacArthur Violence Risk Assessment Study

The Classification of Violence Risk (COVR; Monahan et al., 2006) is an assessment instrument developed from the MacArthur Violence Risk Assessment Study (VRAS). The COVR is computer-implemented and designed to estimate the risk of violence in psychiatric patients; given the appropriate credentials, it is available for purchase from Psychological Assessment Resources (http://www.parinc.com). The COVR assigns patients

Table 1. The five risk categories for the Classification of Violence Risk (COVR) assessment tool along with point estimate risks (in probabilities) and respective confidence intervals (CI) (Monahan et al., 2006).

Category	Risk	Point Estimate	95% CI
5	Very High	.76	[.65, .86]
4	High	.56	[.46, .65]
3	Average	.26	[.20, .32]
2	Low	.08	[.05, .11]
1	Very Low	.01	[.00, .02]

to one of five risk groups defined by the "likelihood that the patient will commit a violent act toward another person in the next several months" (Monahan et al., 2006, p. 728). Table 1 gives the five risk groups defined by their best point estimates and 95% confidence intervals.

The development of the COVR is detailed in Monahan et al. (2001) (also see Steadman et al., 2000, Monahan et al., 2000, and Banks et al., 2004 for less detailed reviews). The COVR was based on a sample of 939 recently-discharged patients from acute inpatient psychiatric facilities in three locations within the United States: Pittsburgh, Pennsylvania; Kansas City, Missouri; and Worcester, Massachusetts. Patients were restricted to those who were white, African-American, or Hispanic; English-speaking; between the ages of 18–40; and charted as having thought, personality, or affective disorder, or engaged in substance abuse.

According to the original MacArthur study (Monahan et al., 2001), violence is defined as "acts of battery that resulted in physical injury; sexual assaults; assaultive acts that involved the use of a weapon; or threats made with a weapon" (p. 17). A second category of violent incidents was labeled as "other aggressive acts" (Monahan et al., 2001, p. 17) including non-injurious battery; verbal threats were not considered. The outcome variable of violence is dichotomous—either the patient committed an act of violence or did not. It does not consider the number of violent acts or their severity. The patients were interviewed once or twice during the twenty weeks after discharge. Of the 939 patients, 176 were considered violent; thus, the base rate for violence in this sample is .187.

The authors identified 134 potential risk factors, listed in detail in Monahan et al. (2001). Using SPSS's CHAID (chi-squared automatic interaction detection) algorithm (SPSS, Inc., 1993), the authors developed a classification tree based on the given risk factors. The final classification model was constructed by an iterative classification tree (ICT) process: after an initial classification tree was developed, those who were still unclassified (that is, those within .09 to .37 estimated probabilities of committing violence according to the model) were reanalyzed using the same CHAID algorithm. After four iterations, 462 patients were classified as low risk (less than .09 probability of committing violence), 257 were classified as high risk (greater than .37 probability of committing violence), and 220 remained unclassified. The cutoffs of .09 and .37 were chosen because they represent, respectively, one half and twice the base rate of violence in the sample.

The authors' goal was to create an actuarial tool that was "clinically feasible"; thus, it was to include only risk factors that could be computed easily. Of the 134 original risk factors, 28 were eliminated that "would be the most difficult to obtain in clinical practice" (Monahan et al., 2001, p. 108), as determined by the length of the instrument measuring the risk factor (more than twelve items was considered too long), or the risk factor not being readily or easily ascertainable by mental health professionals. After doing so, the same ICT method was applied to the 106 remaining risk factors using three iterations.

The correlation between the predictions made by the clinically-feasible and original ICT models was .52; the authors noted the low correlation:

> The fact that these [two] prediction models are comparably associated with the criterion measure, violence (as indicated by the ROC analysis), but only modestly associated with each other [as indicated by the correlation coefficient], suggested to us that each model taps into an important, but different, interactive process that relates to violence. (p. 117)

The authors then constructed nine additional ICT models using the 106 clinically-feasible variables; for each of the nine trees the authors "forced a different initial variable" (p. 118; that is, the root nodes for the ten trees differed). The ten models led to ten classifications for each individual (high, average, or low) and each individual was assigned a score corresponding to their risk level (1, 0, or -1, respectively); the scores were then summed to create a composite score ranging from -10 to 10. The authors remarked, "As two models predict violence better than one, so ten models predict violence better than two (that is, the area under the ROC curve was .88 for ten models compared to .83 for two models)" (p. 122).

The authors questioned whether ten models were necessary; to determine this empirically, they performed stepwise logistic regression and concluded that only five of the ten were needed, leading to composite scores ranging from -5 to 5 (the AUC remained the same, .88). The composite scores were divided into five distinct groups based on the following ranges: $[-5, -3]$, $[-2, -1]$, $[0, 1]$, $[2, 3]$, and $[4, 5]$ (these five groups correspond to, respectively, the very-low, low, average, high, and very-high risk groups found in Table 1; the probabilities represent the proportion of those violent within each group).

The authors did not cross-validate their model. As Monahan et al. (2001) state on page 106, “Dividing the sample leaves fewer cases for the purpose of model construction” and, quoting Gardner et al. (1996), “wastes information that ought to be used estimating the model.” When their ICT models were constructed in the late 1990s and early 2000s, computing power was not what it is now, but cross-validation on a dataset of 939 was certainly possible (although possibly not in the version of SPSS relied on). With today’s computing power there is little reason not to cross-validate a model or to argue that cross-validation “wastes” data. As will be shown, cross-validated error can be drastically different from what is called the resubstitution error for the initially constructed model.

As noted, Monahan et al. (2001) cited the Gardner et al. (1996) source when they made their remark claiming cross-validation wastes information, so this reasoning did not necessarily originate with them. Looking at the Gardner, Lidz, Mulvey, and Shaw (1996) article referenced in the quote above, a footnote on page 43 states that a bootstrap cross-validation was performed on the authors’ logistic regression model, a perfectly reasonable alternative. Although Monahan et al. performed a bootstrap analysis to estimate the variability of the predictions (and where 1,000 bootstrap samples helped estimate 95% confidence intervals for the probability-of-violence point estimates given in Table 1), the base predictive model was not cross-validated.

5.1 VRAS Dataset

To illustrate the process of cross-validation, we used the data from the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) to construct several decision-tree models. As noted, the data were obtained from 939 patients discharged from inpatient psychiatric facilities based in Pittsburgh, Kansas City, and Worcester, MA. The ages of the patients range from 18 to 40 (Mean $=$ 29.9; Median $=$ 30.0). Of the 939 patients, 538 (57%) were male; 645 (69%) were White, 273 (29%) were African-American, and 21 (2%) were Hispanic.

The response variable (`Violence`) is a binary outcome variable representing whether an act of violence took place within the follow up period (`Violence` $= 1$ if an act of violence occurred; `Violence` $= 0$ if not). Thirty-one predictor variables were included based on the results from the main effects logistic regression and iterative classification tree models in Monahan et al. (2001). The data are available for download through the MacArthur Research Network website (http://www.macarthur.virginia.edu/risk.html); the dataset for the present analysis was obtained directly from the MacArthur researchers—it is a “cleaned-up” version from the statistician on

Table 2. Pearson product-moment correlations of predictor variable with response variable, Violence, in reanalyzed dataset (r) and reported correlations in Monahan et al. (2001) (r').

Variable	r	r'	Variable	r	r'
Age	−.07	−.07	HeadInj	.03	.06
BISnp	.05	.05	LegalStatus	.11	.11
BPRSa	−.08	−.08	NASb	.17	.16
BPRSh	.08	.08	NegRel	.05	.06
BPRSt	−.04	−.04	PCL	.26	.26
ChildAbuse	.14	.14	PCS	.03	.03
Consc	.09	.10	PriorArr	.24	.24
DadArr	.15	.15	PropCrime	.11	.11
DadDrug	.14	.16	RecViol2	.14	.14
DrugAbuse	.16	.17	Schiz	−.12	−.12
Emp	−.05	−.05	SNMHP	−.10	−.10
FantEsc	.13	.13	SubAbuse	.18	.18
FantSing	.10	.10	Suicide	−.01	−.01
FantTarg	.12	.12	tco	−.09	−.10
Function	−.01	−.01	Threats	.06	.06
GranDel	−.01	−.01			

the project. The best attempt was made for preprocessing the data to match that in the Monahan et al. analysis. All software code, including the preprocessing as well as variable descriptions, can be found in the Supplementary Material appended to our report online.

As a way of comparing how close our variables match those of the MacArthur authors, Pearson product-moment correlation coefficients between the predictor variables and Violence were compared to those found in Chapter 5 of Monahan et al. (2001; see Tables 5.2, 5.3, and 5.5). Table 2 displays the estimated correlations for each predictor variable with the response variable as well as the reported correlations from Monahan et al. Although not a foolproof method for confirming that the variables were preprocessed in a similar manner, it certainly does indicate discrepancies that may exist. There is a lot of agreement (at least to two decimal places), but it is not complete. Seven of the 31 correlations disagree, six only by one percent. The largest discrepancy is between prior head injury (r = .03 vs. r' = .06).

6. VRAS Classification Tree Model

The number of observations at each node in a classification tree is referred to as the *leaf size*; the minimum leaf size is a constraint provided by the modeler (the default in R is 7; the minimum leaf size set by Monahan et al., 2001 was 50). Rather than using the default setting, we determined the minimum leaf size using cross validation. The minimum leaf size is plotted against the leave-one-out cross-validated error, shown in Figure 1 (solid line). Several minimum leaf sizes give a cross-validated error less than the base rate, implying that the expected error on new data is less than the error when simply predicting all individuals to be nonviolent (that is, what is often called "base-rate prediction"; see Appendix A in Supplementary Material online for further details); the minimum cross-validated error of .179 was obtained at the minimum leaf size of 48. As a comparison, the dotted line in Figure 1 gives the resubstitution error at each minimum leaf size (that is, the misclassification error on the same data used to construct the model); as the tree becomes less complex, or less flexible, (that is, the minimum leaf size increases), the resubstitution error increases toward the base rate. Trees with a resubstitution error equal to the base rate represents those without any branches (that is, trees with only a root node so that no partitions are being made). The observation that the resubstitution error increases as the trees become less complex, and that the cross-validated error decreases and then increases, is an example of the trade-off between bias and variance in predictive models (see Appendix B in Supplementary Material online; also Hastie et al., 2009).

Based on a minimum leaf size of 48, the initial classification tree constructed is shown in Figure 2. The resubstitution error (with a cutscore of .50) was .181, implying the misclassification of 170 observations; the cross-validated error was slightly lower, .179. Both measures indicate the model was outperforming base-rate prediction (the base rate for violence in the sample was .187 so base-rate prediction would be to predict all individuals to be nonviolent resulting in a misclassification error of .187; see Appendix A in Suppementary Marterial online for more details). Based on a .50 cutscore, 48 individuals were classified as violent (27 of whom were) and the rest nonviolent.

The same analyses were repeated but the cost matrix was set so the cutscore for classifying individuals was twice the baserate (that is, .37); thus, false negatives were considered to be about 1.67 times more costly than false positives. The minimum leaf size was determined to also be 48 (in a similar fashion to that done using equal costs); this led to the same tree being produced.

If we adhered to Berk's (2012) 20:1 false negative to false positive ratio, the resubstitution error (with a minimum leaf size of 30) is .570 and the cross-validated error is .586. Because of these results and the fact that it is difficult to justify a 20:1 ratio for the VRAS definition of violence, this cost ratio is not considered in the remaining analyses.

Suppose one decided not to empirically determine a minimum leaf size but let the minimum leaf size be one, the default setting in some soft-

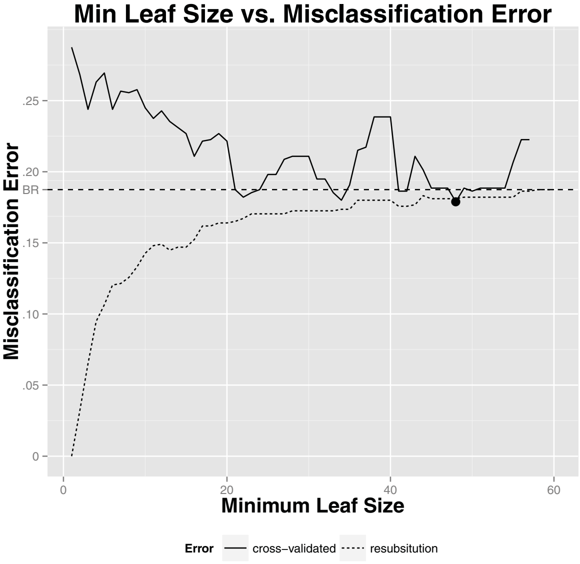

Figure 1. Determining the minimum leaf size for a classification tree with leave-one-out cross-validation error. The dotted line displays the resubstitution error; the solid line, the cross-validated error. The minimum leaf size was determined to be 48.

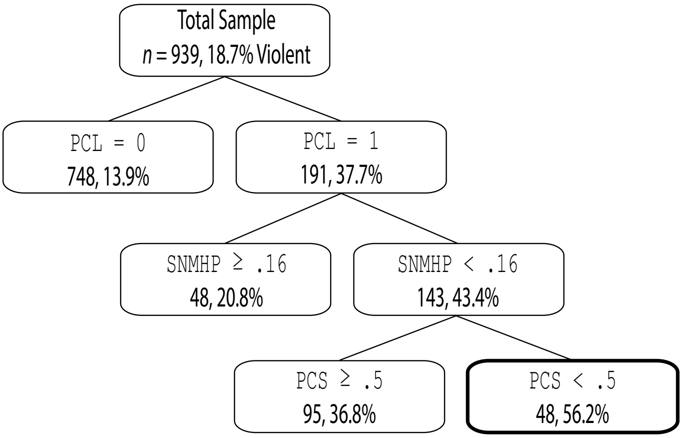

Figure 2. Classification tree with a minimum leaf size of 48 and equal costs. The bold box represents the node where predictions of violence are made. Note that PCL is the Psychopathy Checklist (Hare, 1980); SNMHP is the proportion of social network members who are also mental health professionals; and PCS is the Perceived Coercion Scale (Monahan et al., 2001).

ware (for example, MATLAB). In doing so, the resubstitution error for such a tree was .009; only 8 of the 939 patients were misclassified. The sensitivity of the model (that is, the proportion of violent individuals predicted to be violent) was .966; the specificity (that is, the proportion of nonviolent individuals predicted to be nonviolent) was .997; the positive predictive value (PPV; that is, the proportion of violent predictions that were correct) was .988; the negative predictive value (NPV; that is, the proportion of nonviolent predictions that were correct) was .992; the AUC (the area under the receiver operating characteristic curve) was .999. No method in the literature for predicting violence comes close to these accuracy measures. Without cross-validating this model, however, one blindly believes that an extremely capable model for predicting violence has been found; the cross-validated error was .296, providing strong evidence that the model overfit the data. Based on a cutscore of .37 rather than .50, the model with a minimum leaf size of one misclassified 12 individuals (resubstitution error of .013) but the cross-validated error was .279. Again, this exemplifies the overfitting of a model and the importance of cross-validation, and provides an example of how a more flexible model (smaller minimum leaf size) performs well on the data for which the model was fit but far worse on new data. A good fit does not imply a good model (Roberts and Pashler, 2000).

7. VRAS Random Forest Model

Random forest models were next implemented for the VRAS dataset. A subset of the VRAS dataset was randomly selected as the training sample with the remaining observations representing the testing sample. The testing sample contained 30% of the original data (282 observations); the training data contained the remaining 657 observations (the base rate for violence in the training sample was .181, and .202 in the testing sample). The random forest model was fit to the training sample for $B = 1000$ trees and a minimum leaf size of 10. After fitting 1000 trees, the random forest model was used to predict violence in the training set (that is, the observations used for fitting the model); the predictions were perfect. The 1000 trees generated were aggregated to estimate the probability an individual will be violent by computing the proportion of times the individual is classified as violent (an individual was classified as violent if the predicted probability exceeded .50; that is, costs were considered equal here).

The results discussed thus far are, as noted, based on the training data. The greater concern is with how well violence can be predicted in new observations with the random forest model; this is evaluated with the testing data. Of the 282 observations, two were predicted—one incorrectly—to be violent (see Table 3). The cross-validated error was .202, equal to the base

Table 3. Predicting violence with a random forest model using the testing data. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	1	1	2
	No ($\bar{B}$)	56	224	280
	Column Totals	57	225	282

Table 4. Predicting violence with a random forest model. If a patient had a predicted probability greater than twice the base rate (.37), a prediction of violence was made; otherwise a prediction of no violence was made.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	15	8	23
	No ($\bar{B}$)	42	217	259
	Column Totals	57	225	282

rate. The sensitivity and specificity were, respectively, .018 and .996; the positive and negative predictive values were, respectively, .500 and .800.

The next analysis was the same as the previous one except that the cost ratio of false negatives to false positives was set to 1.67. At the individual tree level an observation was classified as violent when it belonged to a terminal node where the proportion of violent individuals was greater than .37; at the aggregate level (that is, across all 1000 trees) an individual was predicted to be violent when classified as violent in more than 37% of the trees. The results for the training data were again perfect; for the testing data, the results can be found in the bottom of Table 4.

As expected, more predictions of violence were made. For the testing data, the model classified 8.2% of the sample as violent—compared to 0.4% when costs were equal. The random forest model appears to be performing fairly well on the testing data (cross-validated error: .177).

7.1 Out-of-Bag Prediction

Rather than splitting the data prior to fitting the ensemble method, the entire dataset can be used and cross-validation error estimated from the OOB observations; this maximizes sample size (that is, nothing is "wasted") and a cross-validated error is still obtained. The results produced by a cutscore of .50 are displayed in the top of Table 5. The OOB error was .192, slightly more than the baserate. The sensitivity of the model was .057; the specificity

Table 5. Predicting violence with a random forest model on the entire sample. The top of the table uses a cutscore of .50: If a patient had a predicted probability greater than twice the base rate (.50), a prediction of violence was made; otherwise a prediction of no violence was made. The bottom table uses a cutscore of .37.

.50 cutscore

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	10	14	24
	No ($\bar{B}$)	166	749	915
	Column Totals	176	763	939

.37 cutscore

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	13	15	28
	No ($\bar{B}$)	163	748	911
	Column Totals	176	763	939

was .982; the positive predictive value was .417; the negative predictive value was .819; and the AUC was .75.

Carrying out the same analysis but setting the classification cutscore to .37 produced similar results. As the bottom of Table 5 shows, the model classified 3.0% of individuals as violent. The sensitivity of the model was .074; the specificity is .980; the positive predictive value is .464; the negative predictive value is .821. The OOB error was .190 with an AUC of .75.

7.2 Variable Selection

Thus far the decision trees have included all thirty-one variables. Out-of-bag observations allow the quantification of variable importance to the classification trees. For each variable, the values are randomly permuted and the increase (or decrease) in the OOB error calculated (that is, the difference in OOB error before and after permutation). This is carried out for every tree and normalized with the standard deviations of the differences. Variables with larger average differences can be quantified as more important than variables with smaller averages. A dot plot displaying the variable importance is given in Figure 3.

From Figure 3, several variables appear to be more important than others (in particular `PCL`), and several actually *decrease* the OOB error after permutation. It is interesting to note that `Schiz` is among the more important variables but `Age` is not. The decision for removing variables is conservative; only variables with average OOB error differences near and less than zero are removed (`PCS`, `Suicide`, `Consc`, `Age`, `DadArr`, `HeadInj`, and `Threats`); thus, the final model consists of twenty-four predictor variables.

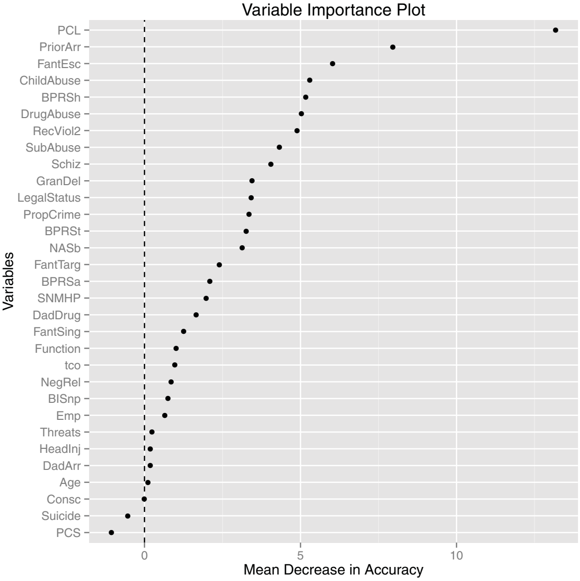

Figure 3. Measure of importance for each of the thirty-one variables. The importance measure is the average of the differences in out-of-bag error before and after permutation across all trees.

Table 6. Predicting violence with the final random forest model with equal costs.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	17	16	33
	No ($\bar{B}$)	159	747	906
	Column Totals	176	763	939

7.3 Final Model

The final model was estimated with 1000 trees omitting the variables discussed in the previous section and with equal costs. The estimated cross-validated error using OOB observations for the final model is .186, a slight improvement upon the random forest model that included all thirty-one variables. The results are summarized in Table 6. The sensitivity of the model is .097; the specificity is .979; the PPV is .515; and the NPV is .825. The ROC plot is given in Figure 4; the AUC for the final model is .75.

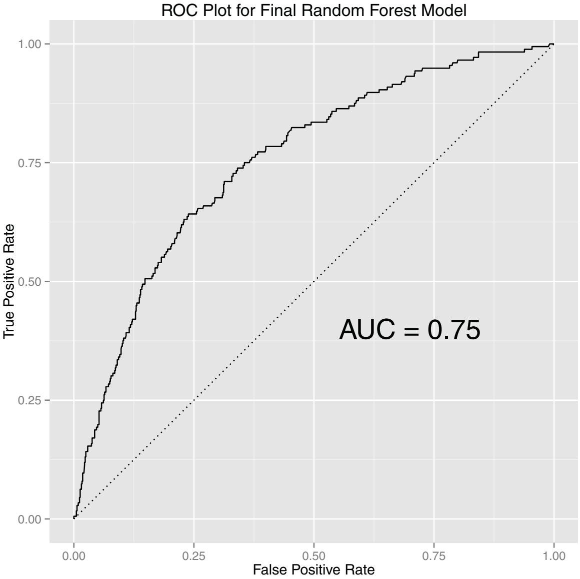

Figure 4. ROC plot for the final random forest model.

8. Conclusion

The results given in this paper reiterate the argument that predicting violent behavior is extremely difficult. Unless unequal costs regarding false positives and negatives are assumed—particularly when false negatives are considered to be more costly than false positives—the models made a small number of predictions of violence which consequently, led to very low sensitivity. Therefore, unless the model explicitly states false negatives as more costly than false positives, predictions of violence are infrequent and the sensitivity is abysmally low. Even when false negatives were stated to be 1.67 times greater than false positives, the sensitivity was far from adequate in the models. Harris and Rice (2013) claim "it can be reasonable for public policy to operate on the basis that a miss (for example, failing to detain a violent recidivist beforehand) is twice as costly as a false alarm (for example, detaining a violent offender who would not commit yet another violent offense)" (p. 106). Whether this is true, it is ethically questionable to assume that costs are anything but equal unless public policy explicitly states otherwise.

The analyses presented also demonstrate the importance of cross-validation. Without cross-validating a model, results and conclusions regarding accuracy can be misleading and overly optimistic. Each type of classification tree model constructed performed quite well when assessed on the data used to fit it, and almost always outperformed base-rate prediction. But when cross-validation methods were implemented, the results were dramatically different. Rarely did the model outperform base-rate prediction on the testing sample, and the resubstitution error was often higher, indicating that the models even failed to outperform base-rate prediction on the training data.

The random forest model was chosen for two reasons: (a) it is a decision tree model and the COVR is based on a decision tree model and (b) random forests have been shown to be highly accurate models in real world applications (Fernández-Delgado et al., 2014). To be sure, a logistic regression model and linear and quadratic discriminant analysis models were also fit for comparison (see the supplementary material for the full details). The results were similar, but the logistic regression model performed best and slightly outperformed the final random forest model in terms of cross-validated error and AUC. But because of the reasons just given, the random forest model was chosen to represent the final model. The final model also did not incorporate unequal costs; as mentioned, unless public policy explicitly states that false negatives should be considered more costly than false positives with respect to the type of violent behavior being predicted in the VRAS sample, we feel that costs do not warrant adjustment. If costs are considered unequal, they should be determined a priori and not from an optimization based on the data; we agree with the quote from Berk (2012) when he says, “the costs of forecasting errors need to be introduced at the very beginning when the forecasting procedures are being developed.” Optimizing a model based on differing cost ratios could lead to unethical decision making and unintended consequences.

Table 7 is from the VRAS study (derived from Table 6.7 in Monahan et al., 2001). When individuals who fall into the very high- and high-risk groups are classified as violent and all others as nonviolent, the model correctly classifies 86.0% of individuals, better than nearly every model presented in the current analysis. At this cutscore the model has a sensitivity and specificity of $\frac{(48+57)}{176} = .60$ and $\frac{(135+229+339)}{763} = .92$, respectively; the positive and negative predictive values are $\frac{(48+57)}{(63+102)} = .64$ and $\frac{(135+229+339)}{(183+248+343)} = .91$, respectively. Recall that the COVR was a combination of ten ICT models, five of which were kept. The authors claim that this “multiple model approach minimizes the problem of data overfitting that can result when a single ‘best’ model is used” (p. 127). Because the authors did not cross-validate, it is impossible to determine how much the model overfits the data, but it certainly seems that it does. As McCusker (2007) says,

> One could wonder whether the iterative classification tree methodology (a technique that involves repetitive sifting of risk factors) that

Table 7. COVR risk groups from Monahan et al. (2001, cf. Table 6.7, p. 125).

		Violence			
		Yes	No	Row Totals	Proportion Violent
Risk Group	Very High	48	15	63	.76
	High	57	45	102	.56
	Average	48	135	183	.26
	Low	19	229	248	.08
	Very Low	4	339	343	.01
	Column Totals	176	763	939	.19

> was used to create the COVR ended up, in a sense, fitting the data in the development sample too specifically. Perhaps as a very carefully tailored garment will be expected to fit one individual perfectly but most other people not as well, so the algorithms of the COVR ought to be anticipated to classify other samples less exactly than they categorized the members of the development sample. (p. 682)

In November 2012, the journal *Perspectives on Psychological Science* released a special issue dedicated to the lack of replicability in psychological research. The issue begins with a question from the editors: “Is there currently a crisis of confidence in psychological science reflecting an unprecedented level of doubt among practitioners about the reliability of research findings in the field?” (Pashler and Wagenmakers, 2012, p. 528); they immediately follow their question with an answer: “It would certainly appear that there is” (p. 528). The recent “replicability crisis” in psychology has given many reasons to question and doubt the results published in psychological journals. Because of its important role in reducing predictive error, cross-validation to construct predictive models can be expected to contribute to improved replicability. Our results, when compared with cross-validation studies of the COVR, provide an illustration of this assertion.

As important as cross-validation is for developing a model, the true test lies in how well the model does with an independent sample; that is, can the results be replicated? Therefore, the next step in validating the model is to assess the accuracy with an independent sample. To date, five studies have attempted to validate the COVR (Doyle et al., 2010; McDermott et al., 2011; Monahan et al., 2005; Snowden et al., 2009; Sturup, Kristiansson, and Lindqvist, 2011). Table 8 displays a summary of the original study, the five validation studies, and the present study; many of the measures are more closely represented by the results found in the cross-validated models presented here. For instance, the AUC from the original study (Monahan et al., 2001) is .88 whereas the AUC for all validation studies are between .58 and .77 (the AUC for our final random forest model is .75). The sensitivity in the original study is .60; in four of the five validation studies the sensitivity is

Table 8. Summary of COVR studies.

			Study					
	Original	Present	1	2	3	4	5	Average
Violence Base Rate	.19	.19	.23	.52	.24	.15	.06	.16
Violence Selection Ratio	.17	.04	.35	.21	.05	.14	.03	.13
Reported AUC	.88	.75	.70	.73	.58	.73	.77	—
Sensitivity	.60	.10	.75	.37	.00	.41	.21	.39
Specificity	.92	.98	.77	.96	.93	.91	.98	.92
PPV	.64	.52	.49	.91	.00	.45	.36	.48
NPV	.91	.82	.91	.59	.75	.90	.95	.89
Misclassification Error	.14	.19	.24	.35	.29	.16	.07	.17

Note. Aside from *Reported AUC*, all statistics are calculated using data in the form of 2 × 2 contingency tables where a COVR score of 4 or 5 leads to a prediction of violence and all lower scores do not.
Original: Monahan et al. (2001); Present: Results are based on the final random forest model with equal costs (see Table 6; Study 1: Monahan et al. (2005); Study 2: Snowden, Gray, Taylor, and Fitzgerald (2009); Study 3: Doyle, Shaw, Carter, and Dolan (2010); Study 4: McDermott, Dualan, and Scott (2011); Study 5: Sturup, Kristiansson, and Lindqvist (2011); Average: Weighted average of the five validation studies.

below .50 (for our final random forest model it is .10). The positive predictive value for the construction study is .64 whereas four of the five validation studies have a PPV below .50 (the PPV for the final random forest model is .52).

The results from Table 8 suggest that the five validation studies did not replicate the findings of Monahan et al. (2001). Rather, the validation results give more evidence of the results presented here; all measures provided in Table 8 are closer to the results from the final random forest model than the original study's model they are based on, aside from the sensitivity and specificity which is a direct result of the choice to not apply unequal costs in the final model. Thus, we conclude that the lack of cross-validation in a prediction model should also be reason for skepticism.

References

BANKS, S., ROBBINS, P.C., SILVER, E., VESSELINOV, R., STEADMAN, H.J., MONAHAN, J., and ROTH, L.H. (2004), "A Multiple-Models Approach to Violence Risk Assessment Among People With Mental Disorder", *Criminal Justice and Behavior, 31*, 324–340.

BERK, R. (2011), "Asymmetric Loss Functions for Forecasting in Criminal Justice Settings", *Journal of Quantitative Criminology, 27*, 107–123.

BERK, R. (2012), *Criminal Justice Forecasts of Risk: A Machine Learning Approach*, New York, NY: Springer.

BREIMAN, L. (1996), "Bagging Predictors", *Machine Learning, 26*, 123–140.

BREIMAN, L. (2001), "Random Forests", *Machine Learning, 45*, 5–32.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth and Brooks.

BREIMAN, L., and SPECTOR, P. (1992), "Submodel Selection and Evaluation in Regression. The *X*-Random Case", *International Statistical Review*, 291–319.

DOYLE, M., SHAW, J., CARTER, S., and DOLAN, M. (2010), "Investigating the Validity of the Classification of Violence Risk in a UK Sample", *International Journal of Forensic Mental Health, 9*, 316–323.

FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S., and AMORIM, D. (2014), "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?", *The Journal of Machine Learning Research, 15*, 3133–3181.

GARDNER, W., LIDZ, C.W., MULVEY, E.P., and SHAW, E.C. (1996), "A Comparison of Actuarial Methods for Identifying Repetitively Violent Patients with Mental Illnesses", *Law and Human Behavior, 20*, 35–48.

GINI, C. (1912), *Variability and Mutability: Contribution to the Study of Distributions and Report Statistics*, Bologna, Italy: C. Cuppini.

HARE, R.D. (1980), "A Research Scale for the Assessment of Psychopathy in Criminal Populations", *Personality and Individual Differences, 1,* 111–119.

HARRIS, G.T., and RICE, M.E. (2013), "Bayes and Base Rates: What is an Informative Prior for Actuarial Violence Risk Assessment?", *Behavioral Sciences and the Law, 31*, 103-124.

HASTIE, T. , TIBSHIRANI, R., and FRIEDMAN, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), New York, NY: Springer.

JAMES, G., WITTEN, D., HASTIE, T., and TIBSHIRANI, R. (2013), *An Introduction to Statistical Learning*, New York, NY: Springer.

KUHN, M., and JOHNSON, K. (2013), *Applied Predictive Modeling*, New York, NY: Springer.

MCCUSKER, P.J. (2007), "Issues Regarding the Clinical Use of the Classification of Violence Risk (COVR) Assessment Instrument", *International Journal of Offender Therapy and Comparative Criminology, 51*, 676–685.

MCDERMOTT, B.E., DUALAN, I.V., and SCOTT, C.L. (2011), "The Predictive Ability of the Classification of Violence Risk (COVR) in a Forensic Psychiatric Hospital", *Psychiatric Services, 62*, 430–433.

MEEHL, P.E., and ROSEN, A.(1955), "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores", *Psychological Bulletin, 52*, 194–215.

MONAHAN, J., STEADMAN, H.J., APPELBAUM, P.S., GRISSO, T., MULVEY, E.P., ROTH, L.H., and SILVER, E. (2006), "The Classification of Violence Risk", *Behavioral Sciences and the Law, 24*, 721–730.

MONAHAN, J., STEADMAN, H.J., ROBBINS, P.C., APPELBAUM, P.S., BANKS, S., GRISSO, T., and SILVER, E. (2005), "An Actuarial Model of Violence Risk Assessment for Persons with Mental Disorders", *Psychiatric Services, 56*, 810–815.

MONAHAN, J., STEADMAN, H.J., ROBBINS, P.C., SILVER, E., APPELBAUM, P.S., GRISSO, T., and ROTH, L.H. (2000), "Developing a Clinically Useful Actuarial Tool for Assessing Violence Risk", *The British Journal of Psychiatry, 176*, 312–319.

MONAHAN, J., STEADMAN, H.J., SILVER, E., APPELBAUM, P.S., ROBBINS, P.C., MULVEY, E.P., and BANKS, S. (2001), *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*, New York, NY: Oxford University Press.

MOSSMAN, D. (2006), “Critique of Pure Risk Assessment or, Kant Meets *Tarasoff*”, *University of Cincinnati Law Review, 75*, 523–609.

MOSSMAN, D. (2013), “Evaluating Risk Assessments Using Receiver Operating Characteristic Analysis: Rationale, Advantages, Insights, and Limitations”, *Behavioral Sciences and the Law, 31*, 23–39.

PASHLER, H., and WAGENMAKERS, E.J. (2012), “Editors’ Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?”, *Perspectives on Psychological Science, 7*, 528–530.

POLLACK, I., and NORMAN, D.A. (1964), “A Non-Parametric Analysis of Recognition Experiments”, *Psychonomic Science, 1*, 125–126.

R CORE TEAM (2014), *R: A Language and Environment for Statistical Computing (Version 3.1.1)*, Vienna, Austria, http://www.R-project.org/.

ROBERTS, S., and PASHLER, H. (2000), “How Persuasive is a Good Fit? A Comment on Theory Testing”, *Psychological Review. 107*, 358–367.

SNOWDEN, R.J., GRAY, N.S., TAYLOR, J., and FITZGERALD, S. (2009), “Assessing Risk of Future Violence Among Forensic Psychiatric Inpatients with the Classification of Violence Risk (COVR)”, *Psychiatric Services, 60*, 1522–1526.

SPSS, INC. (1993), *SPSS for Windows* (Release 6.0), Chicago, IL: SPSS, Inc.

STEADMAN, H.J., SILVER, E., MONAHAN, J., APPELBAUM, P.S., ROBBINS, P.C., MULVEY, E.P., and BANKS, S. (2000), “A Classification Tree Approach to the Development of Actuarial Violence Risk Assessment Tools”, *Law and Human Behavior, 24*, 83–100.

STURUP, J., KRISTIANSSON, M., and LINDQVIST, P. (2011), “Violent Behaviour by General Psychiatric Patients in Sweden: Validation of Classification of Violence Risk (COVR) Software”, *Psychiatry Research, 188*, 161–165.

VRIEZE, S.I., and GROVE, W.M. (2008), “Predicting Sex Offender Recidivism. I. Correcting for Item Overselection and Accuracy Overestimation in Scale Development. II. Sampling Error-Induced Attenuation of Predictive Validity over Base Rate Information”, *Law and Human Behavior, 32*, 266–278.