Predictive Power of Dynamic Risk Factors in the Finnish Risk and Needs Assessment Form Compared to Static Predictors

Benny Salo[1], Toni Laaksonen[1], Pekka Santtila[1,2]

[1]Åbo Akademi University, Faculty of Arts, Psychology and Theology

[2]New York University Shanghai, Faculty of Arts and Sciences

Author Note

Corresponding author: Benny Salo

This work was supported by the Criminal Sanctions Agency in Finland. It was also supported by personal funding: for the first author by the National Doctoral Programme of Psychology in Finland, the Finnish Cultural Foundation, the Åbo Akademi University Foundation, Svensk-Österbottniska Samfundet and Waldemar von Frenckells stiftelse; and for the third author by the Academy of Finland Project 287800.

Abstract

We compared the predictive potential of dynamic and static risk factors of recidivism, via machine learning methods. The data contained 746 men that had, and 746 men that had not, reoffended during follow-up periods between 179 and 1332 days. Static predictors included the crime committed, prison history, and age. Dynamic predictors were 50 items from the Finnish Risk and Needs Assessment Form (RITA). Static risk factors strongly predicted both general and violent recidivism. Dynamic predictors performed slightly worse—they added little beyond static risk factors to the prediction of general recidivism and minimally to the prediction of violent recidivism. All the predictive models had good discriminative power with AUC between .70 and .80 and good calibration. Using static predictors, however, produced a wider range of estimated probabilities. Results show that these dynamic risk factors, as assessed, do predict recidivism but, in our opinion, risk assessments should primarily use static predictors.

Keywords: risk and needs assessment, dynamic risk factors, recidivism, machine learning, RITA, Finland

Predictive Power of Dynamic Risk Factors in the Finnish Risk and Needs Assessment Form Compared to Static Predictors

Risk assessment in correctional settings has two important, and partly competing, roles: to assess the risk of new crimes and thereby assign prisoners to the appropriate level of supervision, and to identify needs of the prisoner to provide treatments that can facilitate successful reintegration into society (Monahan & Skeem, 2016). As a tool for meeting these challenges, *risk- and needs assessment instruments* that consider so called *dynamic risk factors* are increasingly being employed (Cording, Beggs Christofferson, & Grace, 2016). Dynamic risk factors (risk factors that, at least theoretically, can change during the prison sentence) are set in contrast to *static risk factors* that cannot be changed through interventions (e.g. number of previous sentences or age). Dynamic risk factors (such as alcohol use and employment problems) have received increasing attention as important parts of risk and needs assessment. While static variables can be recorded cheaply and reliably, dynamic variables have shown to be able to predict recidivism on similar level as static variables and are viewed to have the added benefit of simultaneously indicating targets for intervention (Campbell, French, & Gendreau, 2009; Yang, Wong, & Coid, 2010). In this study we examine the relative predictive power of dynamic and static risk factors in the applied setting of Finnish prisons. We also put the predictive power of these predictors, known at the start of the sentence, in relation to the potential predictive power of considering events during the sentence.

It is worth noting that *assessing risk* and *identifying needs* correspond to *predicting* and *explaining* recidivism, which are tasks that put different demands on both the variables and the statistical methods used (Shmueli, 2010). In reviewing the current state of research on dynamic

risk factors both Cording and collegues (2016) and Ward (2016) pointed out that dynamic variables that have their origin in prediction tools often are multi-dimensional composite constructs, rather than causal. This is currently very common, and a lot of work is needed to turn dynamic variables used for prediction into construct valid risk factors that also show an actual causal link to recidivism. All this is needed for dynamic variables to, truly, be considered effective targets for intervention. The present study focuses on prediction, and we explain why below.

Regarding statistical methods, the goals of accurate prediction and of interpretable explanation may be incompatible in the sense that maximizing one often entails compromises on the other (Kuhn & Johnson, 2013). When we try to explain how recidivism can be predicted we focus on the function that links the predictors to the outcome. This function is likely to be complex and, to be able to describe it, we often need to simplify it into a manageable statistical model with a set of assumptions. An alternative is to ease our requirement to understand our model in favor of maximizing how well it duplicates the results of the true function (Breiman, 2001b). It can be argued that this latter approach, treating the relationship between predictors and outcome as a *black box,* is unsatisfactory. Focusing on prediction rather than explanation does, however, have several scientific uses, among them the aim to quantify to what extent an outcome is predictable with the information at hand (Shmueli, 2010).

Quantifying the predictive potential of a set of variables is valuable for prediction in its own right. Even if we, ultimately, choose to use a more transparent or more easily applied scoring technique, we are interested in how much is lost in terms of predictive power in this simplified application. Regarding the hope to eventually find targets for intervention, the

predictive potential tells us something about the potential, of causal constructs that might be related to the measured predictors, to explain recidivism.

The present article aims to contribute to the discussion of the role of dynamic predictors of recidivism in two ways. For one, we perform an explicit comparison of the predictive potential of static and dynamic predictors through methods that are designed to optimize predictive performance rather than interpretability. In addition, we examine the predictive potential of dynamic variables as they are assessed in an applied setting rather than under the strict control of scale developers. This is in contrast to much of the research on dynamic risk factors (Cording et al., 2016). We do all this in the setting of risk assessment in Finnish prisons.

Risk Assessment in Finnish Prisons

In Finland, in accordance with the Finnish Imprisonment Act (*Vankeuslaki, 767*, 2005), a sentence plan for how the term should be served is drawn up, including placement and planned activities during the sentence. In this plan, available information about previous sentences, working and functional ability, criminality, and other circumstances are considered. The plan is drawn up in one of the regional assessment and allocation units and amended in the prison unit where the person is eventually placed. In some cases (typically for persons with sentences longer than one year) the assessment unit uses a structured assessment form, here termed the *Finnish Risk and Needs Assessment Form* (*Riski- ja tarvearvio,* RITA). The RITA contains two opening section on the details of the committed crimes for the current sentence and eight sections covering dynamic variables. These sections are: *Accommodation and managing activities of daily living; Income and managing financial situation; Education, employment and skills supporting education and employment; Social bonds and life-style, Alcohol use; Drug use; Thinking and behavior;* and *Attitudes*.

RITA is based on the *Offender Assessment System* (OASys) used in England and Wales (Debidin, 2009). It was translated and adapted for the Finnish circumstances in 2004 (Lilja, 2014) and taken into common use by the Criminal Sanctions Agency in Finland in 2006. The form has no formal status in any process of correctional decision-making, however. For example, there are no regulations as to what kind of RITA profile an inmate should have in order to be placed in an open institution (Criminal Sanctions Agency in Finland, 2004).

Dynamic Versus Static Actuarial Predictors

As long as causal links are not established, it is useful to view both static and dynamic variables as *markers* of unknown causal factors for recidivism (Ward, 2016). Even if we limit our view of dynamic variables to them being risk markers, their use is motivated by the possibility to, in fact, measure *changes* in risk. This, however, is conditional on that dynamic variables actually can change—not only in theory, but in practice—and that these changes are associated with changes in risk. There is some evidence that this is true for at least some dynamic variables (Clarke, Peterson-Badali, & Skilling, 2017; Howard & Dixon, 2013). In addition to this requirement, basing risk assessment on dynamic variables should not entail a compromise on predictive performance, and preferably add to predictive accuracy to motivate the added cost in development and application of dynamic variable measurements. The present study focuses on the requirement of actual predictive power.

There is ample evidence for actuarial assessment generally having higher predictive power than clinical judgement (Andrews et al., 1990; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Quinsey, 2009; Ægisdóttir et al., 2006). Beyond that, there is research suggesting that it does not much matter for predictive accuracy what type of items are included in these actuarial assessment for predicting recidivism. Kroner, Mills, and Reddon

(2005) randomly picked items from four well-established risk assessment instruments and showed that the resulting instruments predicted recidivism as well as the original instruments. In meta-analyses, risk assessment instruments that include, and those that do not include, dynamic risk factors showed similar predictive performance (Campbell et al., 2009; Yang et al., 2010), and there are several instruments developed to measure dynamic risk factors that perform *on par* with other risk assessment instruments (See Table 1 in Cording et al., 2016). Coid and collegues (2011) suggest that there seems to be a "glass ceiling" for predictions of predicting violence just above an AUC-value of .7 and that only retrospective studies tend to reach AUC-values of .8 or above.

Information Unavailable to Actuarial Assessment at the Start of the Sentence

Actuarial assessment at the start of the sentence may be unable to perfectly capture what is knowable at that early stage about the risk of recidivism, but it is also worth considering that there is information that is revealed, or events that occur, later during the sentence. To get an indication of how much the actuarial assessments at the beginning of the sentence misses, and how much information is potentially available by considering how the sentence unfolds, we examine the added predictive power of including four additional variables; *release from open prison*, *conditional release*, *supervised parole*, and *crime during the prison sentence*. We consider these to be indicators of the collective predictive power of personnel judgement, prisoner behavior, and treatment during the served sentence. We do not propose a way to incorporate these variables into risk assessment, or make any claims that it would be useful. Our purpose is, rather, to put the predictive performance of the actuarial information in relation to the maximum available information in the data that we analyze.

Quantifying Predictability

To help us think about how to quantify predictability, let us introduce some concepts commonly used in the field of *machine learning*. The inaccuracy of predictions can be divided into *reducible* and *irreducible error* (Hastie, Tibshirani, & Friedman, 2009). The reducible part is due to misspecification of the model, and reduction of this error is, thus, to a large part in the hands of the modeler. The irreducible part is out of the control of the modeler. It is due to *unreliability of predictors*, *unreliability of outcome measurement*, and *unmeasured predictors*. Considering these sources of inaccuracy, it becomes clear that if we can put predictive models on equal footing regarding reducible error, we can isolate the differences in the irreducible error. If we, in addition, can exclude differences in unreliability of the outcome measurement, we are able to attribute differences in the predictive performance of different models to what predictors are considered in those models. The differences that are attributable to predictors will be both due to their information value and to the level of reliability of the measurement of these predictors.

To control for unreliability of the outcome measure, we simply use the same outcome in all models. To control for reducible error is less straightforward. The natural approach is to minimize reducible error for all models. However, the exact proportion of reducible and irreducible error is never known (Hastie et al., 2009). The best we can do is to choose algorithms that have proven to do a good job of minimizing reducible error and assume that the remaining differences between models are minimal.

There exist a multitude of algorithms that aim to minimize reducible error and produce maximally accurate predictions – none of which can be considered universally superior to others (Hastie et al., 2009; Kuhn & Johnson, 2013). While more complex models often are able to extract more predictive power than more interpretable models, the predictive performance of

various methods depends on the nature of the predictors, the outcome, and the relationship between them. What model will perform best is generally unknown before testing. To put models with different predictors on equal footing, it is thus worth considering more than a single statistical method for prediction.

Evaluating Predictive Performance: Discrimination and Calibration

There are two parts to the predictive performance of a model: discrimination and calibration (Helmus & Babchishin, 2017). Discrimination is the ability of the test to assign higher estimated probabilities of recidivism for individuals that do in fact reoffend, compared to individuals that do not reoffend. On the other hand, a test is well calibrated if it assigns a level of estimated probabilities to a group of individuals that corresponds well with the actual recidivism rate in that group.

While good predictors are likely to improve both, discrimination and calibration are very different concepts. A test can have good discrimination without having good calibration. As long as the estimated probability is relatively higher among individuals that eventually reoffend than among individuals that do not, the discrimination can be considered good, regardless of what those estimated probabilities are. Calibration requires that the estimated probabilities correspond to actual recidivism rates. We evaluate both forms of predictive performance in this study.

Research Questions

In the present study we examine the predictive performance of different sets of predictors. We did this both for predictions of general and of violent recidivism. Specifically, we asked the following questions:

1. Do dynamic predictors, in the form of RITA-items, have discriminative predictive power similar to the available static predictors, with both sets recorded at the start of the sentence?
2. Do dynamic predictors and static predictors add to the discriminative predictive power of the other set?
3. What is the added predictive potential of also considering information gathered during the prison sentence? Put another way, how much do we estimate that the models miss (in respect to predictive power) by only considering information known at the start of the sentence?
4. Are any of the prediction models considering dynamic or static items useful for risk assessment from the perspectives of discrimination and calibration?

Method

Sample

Information used in the present study was retrieved from the *National Prisoner Database* (in Finnish: *Vankitietojärjestelmä*), upheld by the Criminal Sanctions Agency in Finland. The database contains records of current and former inmates, including information about the offences they have been convicted for, temporary releases or parole, possible disciplinary reports as well as assessment reports (if completed).

This sample consisted of a total of 1564 individuals, released from a Finnish prison between 2007 and 2011. Subjects were considered if they had the Finnish Risk and Needs Assessment Form (*Riski- ja tarvearvio*, RITA) fully completed. Of these, half (n = 782) were chosen based on the fact that they had been sentenced to a new prison term at the time when the data were collected, during 2012. The period between release and the new sentence ranged from

179 to 1332 days. Each of these 782 reoffending individuals were matched by an individual of the same sex that had been released at approximately the same date but had not been sentenced to a new term by the time the data were collected. The average difference in release date between pairs of reoffending and non-reoffending individuals was 3 days.

Between 2007 and 2011, an average of 4400 prisoners were released from Finnish prisons each year (excluding those imprisoned for remand or as substitute for unpaid fines). Out of these, 6.9% were women (Blomster, Linderborg, Muiluvuori, Salo, & Tyni, 2012). In our sample, the proportion of women was 4.3%. Because of the low number of female prisoners in our sample (n = 68), we decided to limit our examination to the 1496 male prisoners.

Predictors and Outcomes

Static predictors

Static predictors included categories for the crime comitted for the current crime, prison history, and age. The offence for which the current prison was served was coded with 15 binary variables; *homicide*, *assault*, *sexual offence*, *offence against official authorities*, *other offence against the person*, *robbery*, *theft*, *auto theft*, *other property offence*, *criminal damage*, *narcotic offence*, *offence related to possession of weapon*, *traffic offence*, *white-collar offence*, and *other offences* (e.g. relating to animal maltreatment or waste mismanagement). The categories are not mutually exclusive. For example, violent resistance towards a police officer would be coded both as offence against official authorities and as assault.

Four variables pertaining to number of previous sentences were included: number of *prison terms, community service terms*, *remand terms*, and *terms as substitute for unpaid fines*. For 45 men this information was missing. For these, the number of previous sentences were

coded as 0 and a new variable indicating missing information regarding previous sentences was introduced and coded as 1 for these cases.

Other variables pertaining to prison history were a binary variable indicating any *escape, unlawful absence or attempt thereof* during any of the noted terms, and the person's *age when the first term was noted in the prison database*. Current age was defined as the age at release.

Age at first term was missing for 426 individuals. Missing values were replaced by the current age (age at release) rounded down to the nearest integer (age at release was recorded as years in decimal numbers, age at first term was recorded as an integer), and like for the number of previous sentences a variable indicating missing information was introduced. Sentence length was not available in the data.

Dynamic predictors

The considered dynamic predictors were 50 items from the following sections of the RITA: *Accommodation and managing activities of daily living; Income and managing financial situation; Education, employment and skills supporting education and employment; Social bonds and life-style, Alcohol use; Drug use; Thinking and behavior;* and *Attitudes* plus the item *Takes responsibility for the current offence.* The items are all coded on a three-point scale with 0 = risk factor not present, 1 = the risk factor is somewhat present, or 2 = the risk factor is evidently present. Items can be found in the Appendix to Salo, Laaksonen, and Santtila (2016)

Variables related to the prison term

The variables we use to estimate the predictive power stemming from information available after the initial assessment are *placement in open prison, granting conditional release, supervision of parole,* and *crime during the sentence*.

After considering the risk of absconding or crime an individual can be placed to serve part of sentence in an open institution with minimum security. In addition to practical considerations, the individual needs to agree to abstain from drug use and, if considered needed, take recurring drug tests. The individual can be placed back into a closed institution if required. The variable we used as predictor for recidivism was *placement at release.* (Of the males, 664 concluded their sentence in an open prison). This variable is thus a predictor that is a combination of the prison personnel's view of risk, the willingness of the imprisoned individual to meet requirements, and the individual's ability to fulfill these requirements. The same is true for granting *conditional release,* where after similar considerations, the last six months can be served with technical surveillance, but in freedom. For this variable we had information available on whether conditional release was granted (n = 337) and whether it was eventually revoked (75 of those granted conditional release)—this was dummy coded into two separate predictors.

In the majority of cases prisoners are released on parole. This parole is generally supervised if the remaining sentence is longer than 18 months and the supervision may take several forms. Our variable on *supervision of parole* is thus primarily an indicator of sentence length.

Finally, the variable *crime during sentence* indicates whether there was suspicion of a committed during any lawful or unlawful absence from the prison. This also includes later reports from a new conviction that at least a part of the crime was committed during the previous sentence. As with the other variables in this category, the usability of this as a predictor in an applied setting is dubious or unclear, but we see it as an indicator of possibly knowable information before release.

Outcomes

Recidivism was in the present study operationalized as a new prison sentence. For an individual to be considered as having reoffended this prison sentence had to be recorded in the prisoner database. For these new convictions the new offence was coded according to the same categories as the current offence. A new sentence of any kind (including violent offences) was coded as *general recidivism*. New offences were considered *violent recidivism* if they included homicide or assault. We did not consider the category of the previous crime when coding for general and violent recidivism.

Subsets for Cross-Validation

We used two methods to validate the models, repeated cross-validation within a *training set* and validation in a separate *test set*. First we split the sample into a training set (n = 1122) and a test set (n = 374) with three quarters of all individuals in the training set. Within the training set we used cross-validation to choose the values of so called tuning parameters (see below) and to compare the discrimination of models containing different sets of predictors. This cross-validation was done by splitting the training set into ten folds with each of the folds, in turn, serving as a validation set. This ten-fold cross-validation was repeated ten times for added accuracy of estimates of discrimination. After choosing the best values for the tuning parameters for each model (based on resulting AUC), the model was trained on the entire training set and then used to predict recidivism in the test set (once per model).

Comparing models through cross-validation has high power since the confidence intervals around the average performance can be made very narrow by repeating the cross-validation enough times. This is very useful for determining the best set of predictors in the current sample but says less about the generalizability to other samples. The generalizability of

predictive performance is more conservatively evaluated in the test set. None of the observations in the test set have been involved in training or tuning the model. Validation in the test set, on the other hand, has lower statistical power. Statistical power can be increased by allocating more individuals to the test set. However, as this automatically decreases the size of the training set, it diminishes the accuracy of the trained model. In prediction modelling, the balance is often struck with more individuals in the training set (Hastie et al., 2009; Kuhn & Johnson, 2013).

Machine Learning Methods

There has been considerable development of so called machine learning methods over the last few decades (Hastie et al., 2009). The subclass of methods for *supervised learning* (of which logistic regression is an example) is pertinent for classification tasks such as actuarial risk assessment. There has been limited use of advanced supervised learning methods for risk assessment with some researchers finding limited benefit from the methods (Hamilton, Neuilly, Lee, & Barnoski, 2015; Tollenaar & van der Heijden, 2013). However, what method will prove most effective depends on the relationship between predictors and outcome and is unknowable before model training (Berk & Bleich, 2013; Kuhn & Johnson, 2013).

We employed two often used methods for supervised learning with differing strengths: *elastic net logistic regression* and *random trees*. A strength of both algorithms is that they are robust against the effect of non-informative predictors which allows us to include all potentially predictive variables in a given category.

Elastic net logistic regression

Penalized regression methods such as *ridge regression* (Hoerl & Kennard, 1970) and *lasso regression* (Tibshirani, 2011) seek to increase predictive power in a regression model (in this case logistic regression) by increasing parsimony. Parsimony is increased by imposing a

penalty parameter, λ, and thus shrinking the estimated regression coefficients towards 0. Ridge regression imposes this penalty in relation to coefficient size and thus shrinks large coefficients more aggressively and no coefficient all the way to 0. Lasso regression on the other hand imposes an equally strong penalty to all coefficients, all the way down to 0 if λ is large enough. In both ridge and lasso regression, when parameter $\lambda = 0$ the model is an ordinary logistic regression. As λ is increased, the model approaches a null model where all coefficients are 0. In both cases the appropriate value for λ is treated as a *tuning parameter*, decided through testing different values and choosing the value that produces the best results in cross-validation.

Ridge and lasso regression perform well under different conditions and what penalty will give the best result is unknown before testing the models. Often the best solution is a mix of the two types of penalties in what is called an *elastic net* (Zou & Hastie, 2005). The elastic net introduces a second tuning parameter, α, a mixing parameter, which describes the nature of the penalty imposed. Parameter α, takes a value between 0 and 1, from pure ridge to pure lasso regression.

Elastic net logistic regression performs well under similar conditions as logistic regression, that is, when there are linear relationships between predictors and the log-odds of the outcome that are parsimoniously captured by logistic regression coefficients. The elastic net, however, is less susceptible to the effects of multicollinearity thanks to the penalty imposed on the coefficients.

Random forest

Classification trees (Breiman, Friedman, Olshen, & Stone, 1984) splits the sample repeatedly into subsets based on the available predictors. In this case, whatever predictor that can best be used to separate reoffending individuals from those individuals that do not reoffend, is

chosen first. This is evaluated based on how *pure* (i.e. homogenous in respect to recidivism status) the resulting groups are. A new split is added based on what predictor can achieve the biggest improvement in purity in either of the current groups. Successively a *tree* is *grown* by again splitting the current subgroups.

Random forest (Breiman, 2001a) summarizes the prediction of many classification trees and introduces variability in trees via two processes. First the sample on which the tree is grown is a new bootstrapped sample for each new tree. Second, the number of predictors that the algorithm considers for each split in the tree is constrained to a limited number of randomly selected predictors. Randomly excluding some predictors has the result of allowing weaker predictors, that otherwise would be out-powered, to contribute to the prediction. Introducing variability in trees, in combination of averaging predictions over many trees, has proved to result in very good predictive performance.

The number of random predictors to consider at each split is defined by the modeler as a tuning parameter, m_{try}. The optimal value for m_{try} is not known before testing and is best chosen via cross-validation. Though the value of m_{try} generally does not affect predictive performance very much, a lower value has the effect of giving a bigger role to relatively weak predictors and models with lower m_{try} performs well when these weaker predictors have unique contributions.

The strength of the random forest algorithm lies in its flexibility. By growing trees in multiple layers it can accommodate for complex interactions between predictors. However, while it splits predictors at optimal points, it does not capture linear relationships with the outcome as well as logistic regression. From an interpretability standpoint it is also inferior to logistic regression (Hastie et al., 2009; Kuhn & Johnson, 2013).

Choosing Tuning Parameters

We chose the tuning parameters via testing multiple models, with different tuning values, for their discrimination in the training set. Parameter λ can be quickly tested over a big range of values—we ultimately tested values between 0 and 3 in increments of 0.01. For α and m_{try} we started with a broad range of possible values for the tuning parameters. Based on the results we defined a new range spanning the closest tested values above and below the best tested value. We tested new values within that range and iteratively narrowed in on the final chosen value for the tuning parameter. For parameter alpha we started with the values 0, .25, .50, .75 and 1, and ultimately settled on a value of an increment of 0.02. For m_{try} we started with a set of values including 1 and multiples of 3 up to the number of predictors in the model and narrowed further testing to increments of 1.

Evaluation Criteria of Predictive Performance

One of the most common metrics for evaluating discrimination is the *Area Under the Curve* (AUC) from a *Receiver Operating Characteristic* (ROC) analysis. The AUC considers the specificity and sensitivity of the test at all possible levels of cutoff.

Calibration statistics have unfortunately not been used as much in the risk assessment literature and the methods for examining calibration vary (Hanson, 2017; Helmus & Babchishin, 2017). Commonly calibration is evaluated in some form by comparing the expected rate of recidivism to the observed rate. One illustrative form of doing this is a calibration plot. In our calibration plot, we divided individuals into quintiles based on the estimated probability of recidivism, calculated the average estimated probability in that group, and plotted that against the de facto recidivism rate of the group. A well calibrated model follows a diagonal line where the expected and the observed recidivism rates are equal.

Statistical Software

We used the software environment *R* (Version 3.4.3; R Core Team, 2017) for all analyses. For training the model and cross-validating them in the training set we used the package *caret* (Version 6.0-78; Kuhn, 2008). Together with caret we used the packages *glmnet* (Version 2.0-13; Friedman, Hastie, & Tibshirani, 2010) for elastic net regression and *randomForest* (Version 4.6-12; Liaw & Wiener, 2002) for random forest. AUC values and their bootstrapped confidence intervals were computed using the package *pROC* (Vesrion 1.10-0; Robin et al., 2011). Plots were built using the package *ggplot2* (Version 2.2.1; Wickham, 2009).

Results

Discrimination

The discriminative power of the predictive models measured through AUC are presented in Figure 1. It illustrates the difference in the statistical power of our two methods of comparing predictive models. The test set validations (x-axis) have much wider and largly overlapping confidence intervals with few statistically significant differences. They do overlap the estimates from the training set, but suggest that we should be somewhat careful when generalizing those results. However, to claim that non-significant results indicate that there is no difference between sets of predictors would also be wrong. The differences are more clearly examined in the training set using repeated cross-validation (y-axis). To compare the predictive potential of different sets of predictors we focus on these results.

The pattern regarding the predictive power of sets of predictors is different for general and violent recidivism. For general recidivism models with dynamic and static items perform on similar levels. Static predictors (mean AUC = .772; 95 % CI = [.763, .781]) performed only just better than the dynamic (mean AUC = .763; 95 % CI = [.753, .772]) . Combining the two sets of

predictors produced a slight gain in discrimination (mean AUC = .799; 95 % CI = [.791, .808]). Including term variables indicated that these items, are close, but do not fully reach the potential of what is learnable before release. Using all predictors gave a mean AUC of .830 (95 % CI = [.821, .838]).

Predicting violent recidivism the difference between static and dynamic predictors was more pronounced. The dynamic predictors (mean AUC = .731; 95 % CI = [.720, .742]) performed worse than they did for general recidivism. At the same time static predictors using the random forest algorithm (mean AUC = .791; 95 % CI = [.780, .801]) performed at a similar level to combining static and dynamic (mean AUC = .797; 95 % CI = [.787, .806]) and even adding term related variables (mean AUC = .805; 95 % CI = [.795, .814]) .

The differences we observe in predictive potential are small but not unmeaningful. Using the table comparing effect sizes of different types reported by Rice and Harris (2005) we conclude that the difference between using dynamic predictors alone and combining them with static items corresponds to a difference of Cohen's $d = 0.17$ for general recidivism and $d = 0.30$ for violent recidivism. The discriminative power is also at the high end of what is typically seen for predicting recidivism (Coid et al., 2011). The test set validations for using both dynamic and static items in an elastic net show AUC of .780 (95% CI = [.734, 826]) for general and .803 (95% CI = [.744, .858]) for violent recidivism.

Calibration

Figure 2 shows the calibration of the predictive models. All models show good calibration with observed recidivism rates corresponding fairly closely to the average expected recidivism rate in respective quintile of estimated probabilities. The differences between models are mainly in respect to the range of the estimated probabilities. The estimated probabilities for

violent recidivism is markedly narrower with few estimations over a probability of 50%. This may partly be attributable to the lower base rate of violent recidivism. For both general and violent recidivism the range of predictions is a little narrower when using only dynamic items. In both cases the more conservative range of predictions seem to be motivated. The individuals that the models are able to put in respective quintile recidivate at rates corresponding to the average estimation.

Discussion

In the introduction of the present paper we defined four questions that we aimed to answer regarding the predictive performance of different sets of predictors. We first turn to discussing how we see that the results answer these questions. Our answers are different for prediction of general and prediction of violent recidivism. We then shortly discuss what the performance of the two different supervised learning algorithms tell us about the relationship between predictors and recidivism. After that we discuss what the results suggest about the role of dynamic variables in predicting recidivism. We conclude with policy implications, limitations of the study, and future directions of research.

Do Dynamic Predictors Perform as Well as Static Predictors?

The relative performance of dynamic and static predictors differed depending on whether we predicted general or violent recidivism. The results suggest that dynamic predictors, assessed in the beginning of the sentence, predict general recidivism quite well, and on par with static items. When it comes to violent recidivism there is an advantage of the static items corresponding to Cohen's *d* of 0.24. Our dynamic items predict both general and violent recidivism acceptably well, but violent recidivism slightly worse. The static items, on the other hand, predict violent recidivism better than they do general recidivism.

Do Dynamic Predictors Add to the Predictive Power of Static Items?

Adding dynamic items to the static items does in fact improve discrimination for general recidivism, but not very much (corresponding to Cohen's $d = 0.13$). For violent recidivism the improvement from adding dynamic predictors to the static was minimal. One interpretation of this is that using static predictors alone comes close to filling the predictive potential of the information available at the beginning of the sentence. It comes very close when it comes to predicting violent recidivism and dynamic items therefore have little to add. This interpretation fits well the idea of Coid et al. (2011) about a 'glass ceiling'. The glass ceiling in the present study however lies around AUC = .8.

Is There More Information Knowable Before Release?

Adding variables related to the unfolding of the prison term breaks this glass ceiling with a small improvement of the prediction of general recidivism. Interestingly, we do not see the same with violent recidivism. It is worth noting that what is measured here is discrimination. A lack of increase in discriminative power should not be interpreted as meaning that what happens, or what is learned, during the sentence does not have an effect on the risk of violence. No change in discrimination means no change in the ranking according to estimated risk of recidivism. Thus, a better interpretation might be that extra care is put in the supervision decisions regarding violent offenders and that high risk individuals are treated in line with readily available information about risk. For example, even if placement in open prison would be a protective factor against recidivism, adding placement as a predictor in the prediction model does not affect the discriminative power if placement in open prison is primarily granted to individuals with low risk of recidivism in line with the actuarial assessment.

Are the Models Useful for Risk Assessment?

All models, that use dynamic, static, or both sets of predictors, discriminate between high and low risk individuals reasonably well. This was validated in a separate test set where all models had an AUC between .7 and .8, (however, considerable range of the 95% confidence interval). Calibration was also good for all the models but the range of estimated probabilities was more limited for violent recidivism. The models were not able to identify any individual with a probability of committing a new violent crime much higher than 50%. This may also correspond to the true predictability of violent offences. While for general recidivism there seem to be individuals who are caught in a circle of repeat offending, violent recidivism is less predictable. Using only dynamic items also offered a narrower range of estimated probabilities for both general and violent recidivism making models that include static items alone, or in combination with dynamic items, more useful.

Comparing the Two Machine Learning Methods

We would expect the random forest model to outperform the elastic net when the relationship between the predictors used and the outcome is complex, involving multiple layers of interactions. In our case, the elastic net algorithm performed slightly better in all cases except one: the random forest model produced better discrimination when using only static items to predict violent recidivism. An interpretation of this is that multiple items make a small univarite contribution that is best used as simply added to each other (which an elastic net does effectively). Only when limiting the available information to the static items does the ability of random forest to account for complex interactions make a difference. The typical improvement of the elastic net over a unpenalized logistic regression was a difference in AUC around .05.

The Role of Dynamic Predictors

As laid out in the introduction we aimed to put models on equal footing in all respects than the set of predictors used and their measurement reliability. One concern about the applied use of dynamc risk factors has been that they, because of diminished reliability, do not necessarily perform as well when applied in prisons as when they are developed and tested by the researchers that constructed them (Cording et al., 2016). The RITA form, from which the dynamic items used in the present study are extracted, is administered with considerable freedom of interpretation of the case workers and it could have been assumed that the predictive power would be limited due to lacking measurement reliability. In fact, these items predict recidivism at an acceptable level, though not as well as static items. The issue is more a question of whether the added predictive power outweighs the cost of administration.

The dynamic items only add to the information of static variables when it comes to prediction of general recidivism. One could attribute this to the fact that dynamic variables are changable, that the initial assessment does not necessarily reflect the conditions at release and that therefore the dynamic items loose some of their predictive power over time. Static items are, however, also suspeptible to loose predictive power over time so the relative comparison of dynamic and static predictors is still of interest. Furthermore, we see limited added predictive power from adding the variables that would reflect change during the sentence.

The patterns of discriminative power of the models, in fact, tell different stories about the predictability of general and violent recidivism. Dynamic items and how the sentence unfolds seems to be able to tell us more about the risk for general recidivism than about the risk for violent recidivism. For violent recidivism we seem to be able to learn, what there is to learn, simply from the static variables. This suggest that general recidivism is more circumstance

driven than violent recidivism, and while circumstances can be stubbornly stable in themselves, the stable personal characteristics that have resulted in previous criminal behavior are likely to have a relatively bigger role when explaining possible future violence.

Policy Implications

The results of this study do not give reason to base risk assessment on the dynamic RITA-items rather than on static items. While RITA-items can be used to reach an acceptable level of prediction, the static items perform better and are easier to collect. The predictions can be made using the computer algorithm or converted into an analog tool. Because the logistic regression based algorithms generally performed well an analog tool can be based on the logistic regression coefficients in these models. This conversion might entail simplifying coefficients, deleting low information items and possibly complementing static items with the best dynamic items to reach a parsimonious model that performs close to the models including all static and dynamic items and using computerized algorithms.

It is also worth reminding ourselves that predictions work better on a group level than on an individual level. The calibration plots show that predictions are quite accurate averaged over the members of a certain risk group. The model can allocate an individual to the wrong risk category without severely damaging the performance on a group level.

Dynamic items might serve best, as the RITA is currently applied, for identifying needs and to plan intervention that will improve the prisoner's situation. This has value in itself but might have limited efficacy for decreasing the risk for recidivism, especially when it comes to violent recidivism. The result of interest when it comes to violent recidivism is the high discriminative power of the static items, that suggest that much of can be known about who are at highest risk of reoffending is knowable even before the sentence starts. To be clear, the present

study makes no claim of being able to assess the efficacy of any specific intervention or prison treatment in general. How the prison term unfolds may have important consequences. The high discriminative power of static items simply imply that those who start off with the highest risk, remain the ones with the highest risk. This will be true even if low risk individuals have their risk lowered further, or if the risk of recidivism is lowered collectively for the prison population.

Limitations and Further Research

This study examines only the static and dynamic variables available to us. Other static items, and especially other dynamic items might have given added predictive power. The strength of this study is however in its applied setting.

Even more so, the items that aim to capture what happens during the sentence was limited to only four variables. With repeated assessment of dynamic items more could be said about how dynamic risk factors change and how the risk of recidivism change. With the information already at hand we could go beyond discriminative power and also investigate the effect of the practices of placement in open prison and conditional release on the risk of recidivism. By matching groups on their propensity for being placed in open prison or granted conditionals release we could examine if the level of supervision change the level of risk within a risk group.

References

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendrau, P., & Cullen, F. T. (1990). Does Correctional Treatment Work? A Clinically Relevant and Psychologically Informed Meta-Analysis. *Criminology*, *28*(3), 369–404. http://doi.org/10.1111/j.1745-9125.1990.tb01330.x

Berk, R. A., & Bleich, J. (2013). Statistical Procedures for Forecasting Criminal Behavior. *Criminology & Public Policy*, *12*(3), 513–544. http://doi.org/10.1111/1745-9133.12047

Blomster, P., Linderborg, H., Muiluvuori, M.-L., Salo, I., & Tyni, S. (2012). *Rikosseuraamuslaitoksen tilastoja 2011 [Statistics of the Criminal Sanctions Agency 2011]*. Helsinki: Rikosseuraamuslaitos.

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. http://doi.org/10.1214/ss/1009213726

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Wadsworth International Group.

Campbell, M. A., French, S., & Gendreau, P. (2009). The Prediction of Violence in Adult Offenders: A Meta-Analytic Comparison of Instruments and Methods of Assessment. *Criminal Justice and Behavior*, *36*(6), 567–590. http://doi.org/10.1177/0093854809333610

Clarke, M. C., Peterson-Badali, M., & Skilling, T. A. (2017). The Relationship Between Changes in Dynamic Risk Factors and the Predictive Validity of Risk Assessments Among Youth Offenders. *Criminal Justice and Behavior*, *44*(10), 1340–1355. http://doi.org/10.1177/0093854817719915

Coid, J. W., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Farrington, D., & Rogers, R. (2011).

Most items in structured risk assessment instruments do not predict violence. *Journal of Forensic Psychiatry & Psychology*, *22*(1), 3–21. http://doi.org/10.1080/14789949.2010.495990

Cording, J. R., Beggs Christofferson, S. M., & Grace, R. C. (2016). Challenges for the theory and application of dynamic risk factors. *Psychology, Crime & Law*, *22*(1–2), 84–103. http://doi.org/10.1080/1068316X.2015.1111367

Criminal Sanctions Agency in Finland. (2004). *Sijoittajayksikkötoiminnan käsikirja. Rikosseuraamusalan käsikirjoja 1/2004 [Operational handbook of the assessment and allocation unit. Handbooks of the criminal sanctions field 1/2004]*.

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. http://doi.org/10.1126/science.2648573

Debidin, M. (Ed.). (2009). *A compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009*. Ministry of Justice, United Kingdom. Retrieved from http://webarchive.nationalarchives.gov.uk/20110201125714/http://www.justice.gov.uk/publications/docs/research-analysis-offender-assessment-system.pdf

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

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*(1), 19–30. http://doi.org/10.1037/1040-3590.12.1.19

Hamilton, Z., Neuilly, M.-A., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology*, *11*(2), 299–318. http://doi.org/10.1007/s11292-014-9221-8

Hanson, R. K. (2017). Assessing the Calibration of Actuarial Risk Scales. *Criminal Justice and Behavior*, *44*(1), 26–39. http://doi.org/10.1177/0093854816683956

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. http://doi.org/10.1007/b94608

Helmus, L. M., & Babchishin, K. M. (2017). Primer on Risk Assessment and the Statistics Used to Evaluate Its Accuracy. *Criminal Justice and Behavior*, *44*(1), 8–25. http://doi.org/10.1177/0093854816678898

Hoerl, : Arthur E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. http://doi.org/10.1080/00401706.1970.10488634

Howard, P. D., & Dixon, L. (2013). Identifying change in the likelihood of violent recidivism: causal dynamic risk factors in the OASys violence predictor. *Law and Human Behavior*, *37*(3), 163–74. http://doi.org/10.1037/lhb0000012

Kroner, D. G., Mills, J. F., & Reddon, J. R. (2005). A Coffee Can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry*, *28*(4), 360–374. http://doi.org/10.1016/j.ijlp.2004.01.011

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5). http://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. http://doi.org/10.1007/978-1-4614-6849-3

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://cran.r-project.org/doc/Rnews/

Lilja, T. (2014). *Valvotun koevapauden peruuntumisen ennustettavuus [The predictability of*

annulment of conditional release]. Laurea University of Applied Sciences, Tikkurila, Finland. Retrieved from https://www.theseus.fi/bitstream/handle/10024/85732/Lilja_Tiina.pdf?sequence=1

Monahan, J., & Skeem, J. L. (2016). Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology*, *12*(1), 489–513. http://doi.org/10.1146/annurev-clinpsy-021815-092945

Quinsey, V. L. (2009). Are we there yet? Stasis and progress in forensic psychology. *Canadian Psychology/Psychologie Canadienne*, *50*(1), 15–21. http://doi.org/10.1037/a0014401

R Core Team. (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, *29*(5), 615–620. http://doi.org/10.1007/s10979-005-6832-7

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. http://doi.org/10.1186/1471-2105-12-77

Salo, B., Laaksonen, T., & Santtila, P. (2016). Construct validity and internal reliability of the Finnish risk and needs assessment form. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, *17*(1). http://doi.org/10.1080/14043858.2016.1161940

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. http://doi.org/10.1214/10-STS330

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal*

of the Royal Statistical Society: Series B (Statistical Methodology), *73*(3), 273–282. http://doi.org/10.1111/j.1467-9868.2011.00771.x

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: Series A (Statistics in Society)*, *176*(2), 565–584. http://doi.org/10.1111/j.1467-985X.2012.01056.x

Vankeuslaki [Impisonment Act], 767 (2005). Finland.

Ward, T. (2016). Dynamic risk factors: scientific kinds or predictive constructs. *Psychology, Crime & Law*, *22*(1–2), 2–16. http://doi.org/10.1080/1068316X.2015.1109094

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *13Yang*, *M.*(5), 740–767. http://doi.org/10.1037/a0020473

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. http://doi.org/10.1111/j.1467-9868.2005.00503.x

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., … Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, *34*(3), 341–382. http://doi.org/10.1177/0011000005285875

Tables

Table 1. Tuning parameters used for the prediction models

	General recidivism			Violent recidivism		
	Elastic net		Random forest	Elastic net		Random forest
Predictors	α	λ	m_{try}	α	λ	m_{try}
Dynamic items	0.08	0.41	5	0.02	0.22	19
Static items	0.96	0.02	2	0.04	0.21	3
Dynamic and static items	0.98	0.02	31	0.06	0.20	29
All predictors	0.88	0.02	61	0.06	0.19	31

Figures

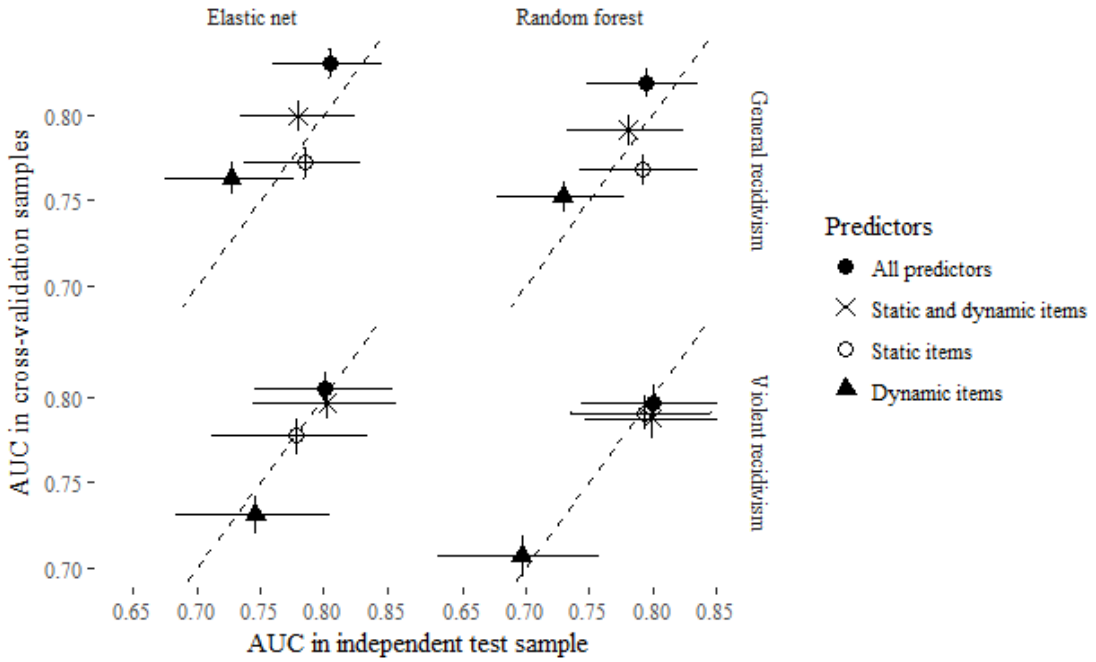

Figure 1. Discrimination of the predictive models measured in Area Under the Curve (AUC). AUC estimated in the test set on the is plotted on the x-axis against AUC estimated through 10-fold cross-validation repeated 10 times in the training set on the y-axis. Lines around points represent 95% confidence interval of the respective estimate. The diagonal dashed line represents equal performance evaluated via both methods of validation. Note the range of both axes are truncated to make the small differences visible.

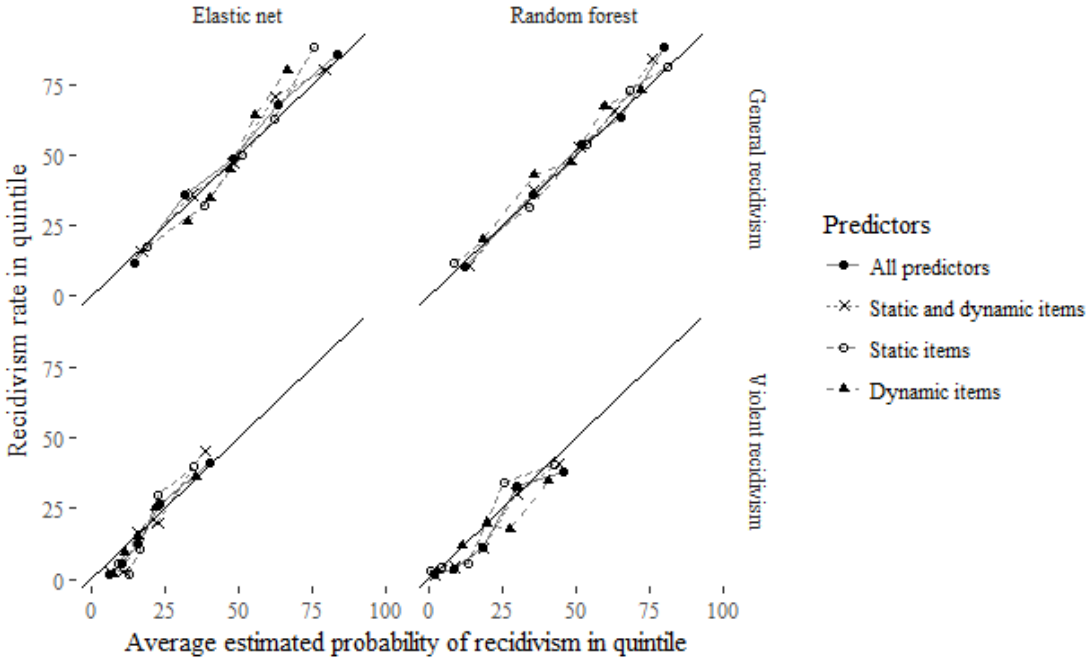

Figure 2. Calibration plot for the predictive models. The test sample (n = 374) is divided into quintiles based on the estimated probability of recidivism. The average estimated probability in respective quitile is plotted on the x-axis against the de facto recidivism rate in the quintile on the y-axis. The diagonal line represents perfect calibration where estimated and de facto recidivism rates are equal. Note that individuals may belong to different quintiles depending on model.