Statrec - Performance, Validation and Preservability of a Static Risk Prediction Instrument

by

Nikolaj Tollenaar[1] (n.tollenaar@minjus.nl)
(Research and Documentation Centre, WODC)
B. S. J. Wartna
(Research and Documentation Centre, WODC)
P. G. M. Van der Heijden
(Utrecht University and University of Southampton)
Stefan Bogaerts
(Tilburg University)

Résumé

Statrec - Performance, validation et conservabilité d'un instrument de prédiction du risque statique. StatRec est un instrument de prédiction du récidivisme utilisant seulement un nombre limité de facteurs statiques. Dans cet article, nous discutons et actualiser le barème Statrec pour les ré-condamnations en quatre ans et d'évaluer ses performances prédictive sur plusieurs dimensions: le temps, la région et des sous-échantillons non aléatoires. En outre, en utilisant les données de dossier pénal, nous étudions dans quelle mesure l'ajout de facteurs dynamiques améliore la performance prédictive. La performance prédictive de l'échelle se révèle être relativement stable au fil du temps et comparable aux échelles OGRS en Angleterre et au Pays de Galles. La précision des probabilités estimées est le seul indicateur qui diminue légèrement à mesure que le temps passe. Bien que l'échelle ne fait pas usage de facteurs dynamiques et situationnels liés au risque de récidive, l'analyse des dossiers pénaux montrent que l'ajout n'améliorer le pouvoir prédictif que légèrement. L'échelle pourrait jouer un rôle dans la validation des instruments de risque dynamiques plus spécifiés.

Abstract

StatRec is a prediction instrument for recidivism making use of only a limited number of static factors. In this paper, we discuss and update the Statrec scale for four-year reconviction and evaluate its predictive performance over several dimensions: time, region and non-random subsamples. Additionally, using criminal file data we investigate to what extend adding dynamic factors improves the predictive performance. The predictive performance of the scale proves to be relatively stable over time and comparable to the England and Wales' OGRS-scales. The precision of the estimated probabilities is the only indicator that decreases slightly as time passes. Though the scale does not make use of dynamic and situational factors related to the risk of re-offending, criminal file analyses show that adding these enhance the predictive power only slightly. The scale could play a role in the validation of more specified dynamic risk instruments.

Mots clefs

Evaluation des risques; Récidive; Validation; ASC; Étalonnage, Statistique juridique

[1] Corresponding author: Nikolaj Tollenaar, Research and Documentation Centre (WODC),
Email: n.tollenaar@minjus.nl

Keywords
Risk Assessment; Recidivism; Validation; AUC; Calibration, Justice Statistics

Introduction

Nowadays in many countries risk assessment is a recurrent element in the administration of criminal justice. Curbing the danger presented by individual offenders has become a major objective in the imposition and implementation of sanctions. The discrimination of the high risks offenders from the low risk offenders has become a central task for those working in the field of criminal justice. The same applies to the Netherlands where risk assessment takes place at various stages of the criminal proceedings. As police officers consider the application of pre-trial detention, when the prosecutor and the judge contemplate the options for administering an appropriate penalty, and when probation or prison officers are asked to advice upon matters of release, at each step of the execution of penal law one needs to know how likely it is that the individual will re-offend. In earlier days, in these situations a behavioural expert would typically be engaged. This expert drew up his recommendations based on interviews with the client in a verbatim report. Nowadays however, placement decisions concerning individual offenders are increasingly based on the use of formal methods of risk assessment, standard questionnaires or checklists that are used to calculate an individual risk score. Research has shown that the application of structured methods of risk assessment provides better results than the mere clinical observations of a behavioural expert (see for example Mossman, 1994; Trout and Bishop, 2002; Aegisdottir et al., 2006).

Risk Assessment Methods

The systematic use of formal risk assessment is fairly new to the Netherlands. The investigation into the practical value of the instruments is still ongoing in many cases. Most instruments are examples of *Structured Clinical Risk Assessment* (SCRA), a method whereby an expert completes a scientifically developed checklist. Examples of SCRA are the PCL-R, the HCR-20 and the SVR-20 (respectively Hare et al., 2000; Webster et al., 1997; Boer et al., 1997). The checklist is filled in by experts: psychologists, orthopedagogues, (juvenile) psychiatrists and probation and aftercare employees. They oversee the dimensions of the questionnaire and base their final conclusions on the scores they have given. The weighing of the scored factors is left up to the expert.

In another form of risk assessment the risk scores are made up according to a fixed, mathematical formula typically from a statistical regression model. In such cases, fixed norms apply to the method of scoring the questions on the list, and the total score is calculated automatically. Here, we refer to an *actuarial* instrument.

Another important distinction in this field is the difference between static and dynamic predictors. Some predictors relate to the here and now and are, in principle, 'treatable' or changeable. These are *dynamic* factors, for instance behavioural disorders and the lack of work or adequate schooling. These factors are related to the *prevention* of recidivism, and can be influenced. On the other hand, *static* factors cannot. Example of static predictors are gender and the number of penal cases. These cannot be manipulated through treatment or intervention. These are factors that cannot be changed. Albeit irreversible, they turn out to be powerful predictors of future criminal behaviour.

In practice, actuarial instruments are more often based on static predictors, whereas SCRA instruments also contain dynamic factors. Besides a recidivism risk assessment, structured clinical risk assessment tools also provide starting points for training, treatment and guidance of offenders. Moreover, dynamic factors can be incorporated into a criminological theory. These factors can be placed in a causal model, which indicates how they are connected to (repeated) criminal behaviour. Actuarial instruments are not deployed to explain re-offending behaviour, but are aimed purely at prediction. Its indicators are correlated with the probability of recidivism, but no causality is implied. the nature of the relationship, the manner in which they influence criminal behaviour is mostly irrelevant. Prediction and explanation are two separate things, as has been shown by Cook and Campbell (1979) who refer to an example from psychology. The number of times a patient underwent psychotherapy proved to be positively related to the probability of suicide. One would almost believe that psychotherapy leads to suicide. However, no causality is involved but only empirical association.

Added Value of SCRA Instruments

Structured Clinical Risk Assessment instruments are made in the interest of 'treatment' of offenders. They contribute more strongly to the formation of theories on criminal behaviour. For these reasons they might be preferred over actuarial methods, which are solely aimed at predicting the risk of recidivism. Some authors claim that the method of Structured Clinical Risk Assessment also provides more accurate predictions than actuarial scales (see, for example, de Vogel et al., 2004). This could well be possible, because the experts play a decisive role when filling in the questionnaire. As in the actuarial approach, empirical connections form the starting point, but with their professional insights, experts are able to recognise the exceptions to the rule, which could well have a positive influence on the predictive powers of the system.

As yet, there is not much empirical evidence to support this claim (Philipse, 2005). The method of Structured Clinical Risk Assessment has advantages over the actuarial approach, but these of course only apply if the prediction of recidivism is at least equally successful. If a specialised SCRA instrument predicts less well than a purely statistical method on the basis of mere, empirical associations, this might mean that criminological notions that are at the basis of the design of the SCRA instrument are incorrect. The treatment envisaged could be less successful than was anticipated. For if the dynamic factors do not display a convincing empirical relation with criminal behaviour the treatment of these factors will not reduce recidivism in the target group.

This argument can be reversed and stated in more positive terms: the quality of a SCRA instrument can be tested by comparing its performance with that of an actuarial instrument. The Research and Documentation Center of the Dutch Ministry of Justice developed an actuarial prediction instrument based on judicial data that assesses the chance an offender will again come into contact with the judicial authorities. This scale can be used to benchmark the predictive validity of SCRA instruments. In this paper we discuss its origin, construction and performance.

Origin of the StatRec Scale

StatRec (Wartna, Tollenaar and Bogaerts, 2009), which is an acronym for *stat*ic risk of *rec*idivism, came into being as a component of QuickScan, which is a screening instrument that is used during the early care contact between the probation and aftercare services and a suspect (De Ruiter and de Jong, 2006). The QuickScan focuses on static *and* dynamic factors. Employees from the probation and aftercare

services use the instruments to record their first impressions of their clients. StatRec forms the first subscale of the QuickScan, but can also be used in isolation. In itself, StatRec is a prediction instrument with static predictors derived from one single source, judicial documentation. Other instruments for establishing the static risk of recidivism are VRAG (Harris and Rice, 1997) and Static-2002 (Hanson and Thornton, 2003). These, however, were developed on samples of violent or sexual offenders. The use of StatRec is limited to adult offenders of all types of crime.

StatRec bears great similarity to a scale applied in England and Wales: the *Offender Group Reconviction Scale* (OGRS). Like StatRec, the OGRS is an actuarial, 'low theory' instrument that focuses exclusively on static factors. Its implementation does not require clinical assessment of the suspect's personal circumstances. Scoring OGRS does not require a behavioural expert. Anyone authorised to inspect judicial documentation can fill in the list. The scale now plays an important role in the execution of British criminal justice. In time, it underwent several adjustments. The first version of OGRS was developed by Copas and Marshall (1998) using data from the so-called offender's index, a database of convictions. It was updated by including more predictors by Taylor (1999) to improve the predictions for violent and sexual offenders. Next it was adjusted to incorporate multiple time points of recidivism (Maden et al, 2005). Most recently, the OGRS switched from conviction data of the Offenders Index to data from the police national computer (PNC), so it would more closely model offending behaviour (Howard, Francis, Soothill & Humphreys, 2009). However, contrary to the offender index, the PNC data are 'weeded', which means that in time older records are removed in order to keep the database size manageable (Francis. Soothill & Humphreys, 2007: 5).

Like OGRS StatRec has its focus on *all* suspects of crimes and like OGRS it concerns *general* recidivism, meaning that no distinction is made on the basis of the nature or the seriousness of the new offence. StatRec can be applied to suspects aged 18 or older. Because the original StatRec scale was over five years old, doubts arose whether the predictions were still valid. Additionally, there was a need to move over to a new system of offence classification. Therefore the StatRec was revised.

This study aims to answer the following questions:

- Does the predictive performance of the StatRec scale change substantially over time?
- What is the form of the re-estimated StatRec scale on 2005 data?
- What is the predictive performance of the re-estimated StatRec scale and how does it compare to the OGRS?
  - Is the predictive performance of the StatRec scale consistent over different regions?
  - Does the scale work well in specific subsamples with a different case-mix?
  - Does the absence of dynamic factors have consequences for the predictive performance?

Method

Data Used

For the development of StatRec, we made use of the data from the Dutch offender's index (DOI), the research and policy database for judicial documentation in the Netherlands. The DOI is an anonymised version of the judicial documentation system (JDS), the legal registration of criminal cases dealt with by the Prosecutor's Office and/or the courts. It holds the complete actualised record of the criminal history of all offenders registered since 1997. The use of registration data implies that only those offences that were detected by the police and handled by the judiciary are is included in the prediction.

An extract from the JDS provides a chronological overview of the criminal case history of a person. Officials from the police and the judicial authorities can request these extracts ex officio; researchers can use them for the investigation of criminal careers. The extracts contain the data required to score StatRec. On a case by case basis, information is registered concerning which public prosecutor's office has received the criminal case, which offences the person was suspected of and which agency decided on the case. All offences are included in the investigation, therefore not only those pursuant to the Dutch Criminal Code, but also the offences that come under special legislation such as the Road Traffic Act, the Opium Act and the Weapons and Ammunition Act.

We choose to use the data of persons convicted in 2005 as the estimation data. Firstly, because Statrec targets the four year reconviction rate[2], sufficient follow-up time should be available for each observation. Secondly, because we had to take incapacitation times into account; the DOI does not yet contain execution dates of sanctions, such as prison terms. The length of the stay in prison is estimated from the actual sentence length minus the (fixed, rule-based) duration of conditional release. Offenders sentenced to longer prison terms (i.e. 2 years or longer) do however still run the risk to have an observation window of less than four years, due to too long prison terms. This showed to be 1,2% of the population and these were withdrawn from the analyses.

Model

The scale was drawn up using logistic regression. This analysis technique is widely used for the prediction of a dichotomous outcome (see also Hosmer and Lemeshow, 2000). The starting point in the construction of the scale was optimising the quality of the prognosis. The predicted probability of recidivism should approach the actual observed recidivism as closely as possible. The simplicity and ease of use of the scale were leading development criteria too. StatRec was intended to be a component of the QuickScan; therefore its implementation should only take several minutes. Six individual characteristics were included in the regression model: gender, age, country of birth, the type of offence of which the person is suspected, the number of previous criminal cases and the 'conviction density'.[3]

Testing the Predictive Validity

In order to test the quality of the model, we made use of split-half cross-validation. One half of the data from 2005 is used to estimate the model, whereas the other half is used to validate the correspondence between the observed and predicted recidivism. The predictive power of a model, the predictive validity, can be evaluated in multiple ways. Vergouwe (2003) distinguishes three aspects to validation: calibration, discrimination and clinical usefulness. *Calibration* relates to the degree of similarity between the observed and predicted chances. The differences should be as small as possible. This aspect is quantified as follows. The window is set to 100. Then the data are first sorted by the estimated probabilities. The first hundred probabilities are averaged as well as their corresponding outcomes and the difference is computed. Then this same difference is calculated for the second to the 101st observation, i.e. the window is shifted one observation. These steps are continued until all observations are covered. Finally, all differences are averaged and the mean calibration error over the 0-1 probability range is obtained. If the probabilities correspond well with the observed outcome, this difference will be

[2] The timing of recidivism is defined from the date of registry of the penal case to the date of committing the new offence.

[3] OGRS largely contains the same characteristics. The 'reconviction density' is similar to the Copas rate (Francis et al., 2005) and indicates how rapidly in succession penal cases occurred, in the active period of the criminal career.

small. The calibration aspect can also be graphically depicted by a calibration plot. In this plot, the line of observed versus predicted is obtained by regressing the outcome on the risk score using kernel regression.

Discrimination is the scale's ability to differentiate between re-offenders and those who do not re-offend on the basis of the predicted score. This is typically measured by the 'area under the ROC curve' (AUC). The AUC indicates which percentage of correctly ranked pairs the instrument will provide overall (Hanley and McNeil, 1982). If the value is 0.5, its performance can be compared to flipping a coin. In the statistical literature an AUC of more than 0.75 is considered 'large' (Shapiro, 1999; Dolan and Doyle, 2000). Hosmer and Lemeshow (2000) consider AUC values starting at 0.70 'acceptable', those from 0.80 on up are considered 'excellent', and values of 0.90 and higher are 'outstanding'. In practice, the AUC is often between 0.65 and 0.80. The AUC value is relatively insensitive to differences in the base rate, i.e. the observed percentage of recidivists in a research population.

Clinical usefulness is related to the number of errors that will be made if the instrument is used to classify individuals into those who will be reconvicted and those who are not. A threshold value (cut-off score) is set during clinical use, typically at 0.5 or at the base rate. Two types of errors are made during this decision-making process: some individuals are wrongly classified as reconvicted ('false positives'), others wrongly as non-reconvicted ('false negatives'). Taken together, these errors constitute a standard for clinical usefulness. The typical measure for clinical usefulness is the error rate or its opposite, the accuracy (ACC). It does however have a serious drawback. If the base rate is relatively high or low, the 'classify all in the largest category' scheme will always yield an impressive accuracy in its own right, while information about individuals is ignored. This is called the no information rate. Copas and Loeber (1990) suggested using the relative improvement over chance (RIOC) in order to correct for this effect. This measure deals with unequal margins of the classification table and chance differences.

StatRec is meant to calculate the probability of recidivism. The scale is not used as a classification instrument so the last dimension might be disregarded. On the other hand, the clinical usefulness does provide information about how well the model predicts.[4]

Testing the Predictive Validity over Time

In order to test the predictive performance of the scale over time we followed these steps:

1. Estimate a model on the 1999 data and generate Statrec-scores using the background characteristics of the 2000-2005 conviction cohorts and establish the height of the performance indices. The complete data sets are used.
2. For each cohort 2000-2005, estimate the model on the estimation half and establish the performance indicators on the split-half validation half. Plot the coefficients of each model over cohorts and thus establish the ´coefficient drift´, i.e. the variation of the coefficients over time.

Step number 1 represents the most direct measurement of change in predictive performance of the original (i.e. not updated) scale. If the results of this step are compared to the performance results of the second step, the gain in performance by updating the scale can be seen. In the coefficient plots, the quantitative effects of the changing coefficient might cancel each other out in the total risk score. Even so, the substantive change may be interesting and indicative on the stability of the scale.

[4] In the annex, additional metric characteristics are given.

These two approaches are schematically depicted in Figure 1.

XX [Insert figure 1 about here]

Testing the Validity over Regions using Court District

The DOI does not contain information on the residential address of the convicts. To provide a geographical proxy, we use the court district. The Netherlands are divided into 19 court district areas. This can be a good proxy, because in the Netherlands, an offender can be prosecuted by the court district of the place in which the offence took place, or by the court district of the residential address or the last known residential address of the offender. In practice, the residential address will be leading. Furthermore, in the majority of cases, the two will coincide. We will compare the same performance criteria over regions.

Testing the Validity in Non-random Subsamples

If the models are adequately specified for each variable, they should generalize well to subsamples that do not reflect the general population in terms of background characteristics. Moreover, if an instrument is widely used, it will in practice also be used in atypical populations. Schmidt and Witte (1988: 131) were the first to do extensively analyse the performance of recidivism models in non-random subsamples. By doing this, we can be certain that the model will perform adequately in samples that do not reflect the distribution of background characteristics of the total population.

In order to test the generalisability of the StatRec model, we use the following subsamples to test the sensitivity of the performance criteria:

- females
- age<=24
- age>=40
- born outside Netherlands
- first offenders
- type of offence
- court disposal type
- releasees from detention

Again, only the validation half of the data is used to estimate the performance in these subsamples.

Estimating the Effect of Additional Dynamic Data

In order to investigate to what extent the predictive quality of the general StatRec model increases when dynamic and situational factors are added to the regression model, we linked the judicial data from the DOI to data originating from criminal files. The additional data are extracted from the Criminal Law Monitor (SRM), a database in which information from a nation-wide random sample of criminal law files were scored in accordance with a fixed protocol (Projectteam SRM, 1997). The judicial information

of more than two thousand randomly selected persons, whose criminal case was disposed of in 1993, 1995 or 1999, was linked to information from the criminal files of the original cases. The additional data contain detailed information about dynamic background characteristics of the offender.

Results

In the next section we will first discuss the change in predictive performance of the old model compared to the performance of annually re-estimated models. Then we deal with the size and sign of coefficients over the period 1999-2005. After that we present the coefficients of the new model based on 2005. Next we elaborate on the validation results of the 2005 model with static information only. Finally, we will present the effect of adding dynamic predictors to the set of static predictors.

Change in Predictive Performance over Time

In Figure 2, the AUC, ACC and '1 minus calibration error' are shown for the StatRec99-score tested in the period 1999-2005. The higher each of these criteria, the better the prediction of the StatRec99 model. The dotted lines represent the Statrec99 model and the other lines represent the re-estimated models. The figure shows that the performance drop over is hardly noticeable in the AUC and the ACC, as their respective lines coincide. The only indicator showing a slight decrease is '1 minus calibration error'. This means that, the older the scale, the larger the discrepancy between the predicted probabilities and the categorized observed recidivism proportions. This can also be seen later in Figure 4a which is a calibration plot of the Statrec99 scores on 2005 data.

XX [Insert figure 2 about here]

Figure 3 shows the trends in the actual regression coefficients if the model is re-estimated each year. It shows that some coefficients change over time. This holds especially for the country of birth. Apparently this characteristic is losing its predictive power with respect to recidivism over time. This probably is caused by the fact that effects correlated to ethnicities cannot be distinguished by country of birth because inhabitants of foreign origin are increasingly born within the Netherlands. The changing absolute values of the coefficients however do not seem to have large consequences as can be seen in the previous predictive performance analyses. This can also be seen in a scatter plot of the old and new StatRec-score (Figure 4b). This figure shows that there is a close correspondence between the 1999 score and the 2005 score. The Statrec99 score provides a slight underprediction of the actual recidivism over the complete range.

XX [Insert figure 3 about here]

XX [Insert figure 4a and 4b about here]

The final StatRec 2005 Model

Table 1 provides the actual parameters of the model. Jointly, these factors provided maximum predictive power. Using different alternatives for logistic regression models did not provide any improved

performance (see Tollenaar and Van der Heijden, 2013). All parameters are statistically significant, which is not surprising given the large number of observations. The relations with recidivism shown by the table are well known. Female offenders and older offenders have lower odds of renewed contact with the judicial authorities than male and younger suspects. The odds of reconviction for women keeping all other predictors constant, is approximately two thirds of the odds for men. The odds of recidivating is generally larger for persons born outside Holland than for persons born in the Netherlands. Offenders from non-Western countries are an exception to this rule. The risk is highest among suspects born in the Netherlands Antilles or Aruba. The number of previous penal cases is a powerful predictor as well. The more extensive the criminal past, the larger the odds that people will revert to criminal behaviour. This is especially apparent in the dummy effects for 11-20 and 21 or more previous convictions. The conviction density is also a powerful predictor on top op that effect. John Copas (Copas & Marshal, 1998) was the first to apply a transformation of this ratio as a predictor in a recidivism prediction model. The more rapid the succession of earlier penal cases, the larger the chance that someone will re-offend and be reconvicted. Finally, the type of offence also makes a difference. The risk of recidivism is largest among suspects of property offences with violence (robbers and extortionists) and the smallest among persons suspect of sex crimes.

XX [Insert table 1 about here]

Statrec Validation

General Adult Offender Population - Figure 5 reflects the results of the comparison of the predicted and the observed recidivism in the random sample for validation purposes, which is the second half of the population of adult suspects from 2005. The calculated score hardly differs from the recidivism that actually occurred in all categories of the predicted risk. As a group, the 1999 suspects have acted as predicted by the model. The number of repeat offenders is predicted fairly accurately with respect to both the high and the low categories of risk scores[5].

XX [Insert figure 5 about here]

The AUC-value of the scale is 0.78. The RIOC is 48,6%. The discriminatory power of the StatRec scale is therefore ‘large’, nearly ‘excellent’ and in any case ‘acceptable’. This is better than many other static risk scales for general recidivism such as the SIR-R1 (Nafek and Motiuk, 2002) that had an AUC of 0.75 and a RIOC of 24%. Hanson, Helmus and Thornton (2010) estimated an AUC of 0.71 with their Static-2002 scale on a population of sexual offenders.

Geographic Validation - In figure 6, the four performance criteria are plotted for all 19 court districts. The figure shows no big variation across regions amongst all four criteria. Even the minimum performance of each of the performance criteria is ‘good’.

XX [Insert figure 6 about here]

[5] The fit of a logistic model can be tested by differencing the observed and the predicted values, as reflected in the figure, in a X^2-*statistic* (Hosmer and Lemeshow, 1980). Any test statistic is, however, not very useful in large sample sizes. Instead, the relevant Effect Size (ES) can be calculated. The ES of the Hosmer-Lemeshow goodness-of-fit test with ten risk categories is 0.03. An Effect Size of 0.1 is ‘small’ (Cohen, 1988).

Non-random subsample validation - Table 2 shows the base rate, AUC, ACC, 1-calerr and RIOC on the various subsamples of the general offender population. It indicates that the scale also performs adequately in subpopulations with lower and higher base rates and different case-mix.[6] In most selections, the AUC value is larger than or equal to 0.75. Only with respect to offenders with 11 or more penal cases and first offenders StatRec's predictive quality scores somewhat lower, because there is little variation in these groups in the number of previous contacts with the judicial authorities.

XX [Insert table 2 about here]

Comparison to OGRS3 in England and Wales - A comparison of the StatRec to the different versions of the OGRS might fall short because of the following considerations. The England and Wales' scales include juvenile offenders and up to the OGRS3, cautions were excluded while similar cases are included in the Dutch data. Nevertheless, a general idea of the relative predictive performance can be established. On a general offender population, the OGRS 3 has an AUC of 80%, compared with 78% for OGRS 2. For prisoners only, it has an AUC of 84% (Howard, Francis, Soothill and Humphreys, 2009). This is not very different from performance of the StatRec. For the general offender population we found an AUC of 0,78 and the StatRec prediction for a prison term group is very similar to the OGRS' performance on prisoners, namely an AUC of 0.83.

Additional Predictive Power of Dynamic Predictors

After linking the criminal data from the DOI to the data from the criminal files, a subset of dynamic predictors were added to the regression equation. In order to investigate to what extent the predictive quality of the general StatRec model increases by adding dynamic predictors, we first calculated the StatRec scores for these 'SRM suspects' classified on the basis of the judicial documentation, and used as a predictor in a logistic regression. This provided a solid prediction of the observed recidivism in the group. The AUC value as estimated on 10-fold cross validation was 0.8 (see Table 2).

Next, dynamic and situational factors are added to the regression equation. Table 3 shows the results of adding the socio-demographic variables, derived from criminal records, to the model.[7] The addition of dynamic factors plus the ethnic background of the suspect increased the AUC to 0.82. Three characteristics were statistically significant. The accommodation, ethnic background and residence status of the offender made a unique contribution on top of the StatRec score to predict recidivism. The other variables, namely civil status, addiction problems of the suspect and day activity program, did not improve the quality of prognosis in this random sample. However, keeping StatRec scores constant, the odds of recidivism are significantly smaller among foreign nationals and tourists. This is not surprising, as especially the most of the latter are probably no longer present in the Netherlands after the index case.

In short, the addition of substantive dynamic factors seems to improve the prediction of recidivism only marginally.

XX [Insert table 3 about here]

[6] Apart from the SRM group, all offender groups are selections from the random sample for validation purposes.

[7] Due to the missing values, this part of the analysis relates to 1,241 suspects. A missing value means that no remark about the relevant characteristic was found in the file. The missings were mainly caused by civil status, day programme and accommodation situation. Omitting these characteristics does not change the outcome of the logical regression with respect to the remaining characteristics.

Discussion

StatRec, a scale for assessing the static risk of recidivism, calculates the probability that a suspect of a crime will again come into contact with the judicial authorities within four years. The prediction is based on data recorded in the judicial documentation. Overall, it provides a relatively sound prediction of general recidivism among adults. Its performance is similar to the OGRS-scale used in England, Wales and Scotland. The performance is very stable over time; the temporal validation showed that there is only a slight decrease in the calibration of predicted probabilities. The other performance indicators where constant over time. This implies that there is no direct need for regular updating. The performance was very similar amongst different regions, suggesting regional independency.

StatRec generalised acceptably to populations that have a different case-mix than the general population, with the possible exceptions of groups consisting of first offenders and groups of frequent offenders. As variables containing information about previous offending are the most predictive of future offending and these are lacking in the first offenders group, this is not surprising. It might also indicate that part of the group of first offender has undetected previous offending and is thus their recidivism is systematically underestimated. On the other extreme, frequent offenders are difficult to predict because after a certain amount of penal cases the probability of recidivism hardly increases. The approach of splitting the population by characteristics is also limited by the covariate information available. In practice, the instrument might be applied to a population that differs on unmeasured factors with a unique relation to recidivism, causing a drop in performance.

Adding dynamic factors scored from criminal files improves StatRec's predictive power only to a small extent. Of course, the outcome might have been different if other 'changeable' factors were used. The choice of dynamic factors, however, seems to fit well with the broad nature of the group under study. Based on the outcome we may conclude that the predictability of known repeated criminal behaviour in general, i.e. for non-specific offender groups, seems to be well represented by a small number of easy to score, static variables. This does *not* mean that, clinically speaking, dynamic factors are irrelevant. For treatment purposes they are important. For prediction purposes they appear to add little or nothing to the static factors included in StatRec.

Implementation of the scale takes little time and does not require special expertise. The metric characteristics of the instruments are good. The instrument scores satisfactorily on calibration and discriminatory power. StatRec is not suitable for clinical use, because it does not provide information that could be important for screening and treatment. It is only useful for making predictions. StatRec estimates the base rate, the normal probability of recidivism in the group of persons with a similar static background profile as the suspect involved. StatRec gives no information on the way the prediction in a relevant case will turn out if a certain intervention is, or is not, implemented. This is why the scale was included in a broader instrument, the QuickScan. The QuickScan also includes dynamic as well as situational factors, and relates to the chances that treatment of those will be successful. Future versions of Statrec will also combine sanction execution data with criminal proceedings data. Doing so, more recent data can be used to estimate the scale.

Estimated risks do not determine whether someone will actually re-offend. Ultimately it is not the criminal data, but the individual circumstances and situational factors that cause someone to revert to criminal behaviour. Information on dynamic factors need not always be available, in order to make the best possible prediction. Knowledge of individual and situational factors remains necessary to be able to interpret the results of the prediction and to be able to oversee any treatment possibilities. In other words, *screening* of suspects is not possible without investigation of the dynamic factors. Estimating the statistical risk of recidivism sometimes is.

The isolated StatRec-scale cannot replace the risk assessment instruments used in the Netherlands, because it is not suitable for clinical use. It can however play a role in the further development of dynamic risk instruments. The relations identified between the variables in the model and the chance of new penal cases were found to be very stable and robust. As a result of its general nature and its broad deployability, StatRec provides, as it were, the 'basic predictability' of criminal recidivism in a group. In doing so, it provides a lower limit for the predictive validity of more specialised systems for individual risk assessment. The predictive error of a SCRA instrument should be lower than based on the empirical connections that also apply to other groups. It is not self-evident that SCRA instruments provide better predictions than actuarial methods. For instance the England and Wales' OGRS was applied to random samples of detainees with psychological disorders in studies conducted by Gray et al. (2004) and Coid et al. (2007). OGRS contained fewer prediction errors than HCR-20 and PCL, which are questionnaires used by psychologists and psychiatrists and which have been specifically designed for this target group. Developers of specialised risk assessment instrument using Dutch data can run StatRec parallel to their own instrument to assess its performance. If the predictive performance of the SCRA instrument is as good as or better than the predictive performance of StatRec, this constitutes an indication that the developers are on the right track as regards the prediction of the risk of recidivism inherent in the target group. But if the SCRA is outperformed by a general, easy to use and theory-neutral prognostic device, one will have to go back to the drawing board.

References

Aegisdottir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, Nichols CN, Lampropoulos GK, Walker BS, Cohen G, et al. (2006) The Meta-analysis of Clinical Judgment Project: Fifty-six Years of Accumulated Research on Clinical versus Statistical Prediction. *Counseling Psychologist* 34(3): 341-82.

Boer DP, Hart SD, Kropp PR and Webster CD (1997) *Manual for the Sexual Violence Risk-20: Professional Guidelines for Assessing Risk of Sexual Violence*. Vancouver BC: Institute against Family Violence.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale NJ: Lawrence Erlbaum Associates.

Coid J, Yang M, Ullrich S, Zhang T, Roberts A, Roberts C, Rogers R and Farrington D (2007) *Predicting and Understanding Risk of Re-offending: The Prisoner Cohort Study*. London: Ministry of Justice, Research Summary 6.

Cook TD and Campbell DT (1979) *Quasi-experimentation: Design and Analysis for Field Settings*. Chicago IL: Rand McNally.

Copas JB and Loeber R (1990) Relative Improvement over Chance (RIOC) for 2×2 Tables. *British Journal of Mathematical and Statistical Psychology* 43(2): 293-307.

Copas J and Marshall P (1998) The Offender Group Reconviction Scale. *Applied Statistics* 47: 159-71.

De Ruiter C and de Jong E (2006) *Handleiding QuickScan Reclassering Nederland*. Utrecht: Trimbos-instituut.

De Vogel V, de Ruiter C, van Beek D and Mead G (2004) Predictive Validity of the SVR-20 and Static-99 in a Dutch Sample of Treated Sex Offenders. *Law and Human Behaviour* 28(3): 235-51.

Dolan M and Doyle M (2000) Violence Risk Prediction: Clinical and Actuarial Measures and the Role of the Psychopathy Checklist. *British Journal of Psychiatry* 177: 303-11.

Francis, B, Soothill, K, & Humphreys, L (2007). Development of a reoffending measure using the

Police National Computer database. Lancaster: Centre for Applied Statistics, Lancaster University.

Howard, P, Francis, B, Soothill, K, & Humphreys, L. (2009). OGRS 3: The revised offender group reconviction scale. Lancaster: Centre for Applied Statistics, Lancaster University.

Gray NS, Snowden RJ, MacCulloch S, Phillips H, Taylor J and MacCulloch MJ (2004) Relative Efficacy of Criminological, Clinical, and Personality Measures of Future Risk of Offending in Mentally Disordered Offenders: A Comparative Study of HCR-20, PCL:SV, and OGRS. *Journal of Consulting and Clinical Psychology* 73(3): 523-30.

Hanley JA and McNeil BJ (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143: 29-36.

Hanson RK and Thornton D (2003) *Notes on the Development of Static-2002*. Ottowa: Department of the Sollicitor General.

Hanson RK, Helmus L and Thornton D (2010) Predicting Recidivism among Sexual Offenders: A Multi-site Study of Static-2002. *Law and Human Behavior* 34, 198-211.

Hare RD., Clark D, Grann M and Thornton D (2000). Psychopathy and the Predictive Validity of the PCL-R: An International Perspective. *Behavioral Sciences and the Law* 18(5), 623-45.

Harris GT and Rice ME (1997) Risk Appraisal and Management of Violent Behavior. *Psychiatric Services* 48: 1166-76.

Hosmer D and Lemeshow S (2000) *Applied Logistic Regression*. New York: John Wiley, Sons Inc.

Maden A, Rogers P, Watt A, Lewis G, Amos T, Gournay K and Skapinakis P (2005) *Assessing the Utility of Offenders Group Reconviction Scale-2 in Predicting the Risk of Reconviction within 2 and 4 Years of Discharge from English and Welsh Medium Secure Units (MRD 12/58)*. London: Academic Unit of Psychiatry.

Mossman D (1994) Assessing Predictions of Violence: Being Accurate about Accuracy. *Journal of Consulting Psychology* 62(2): 783-92.

Nafek M and Motiuk LL (2002) The Statistical Information on Recidivism - Revised 1 (SIR-R1) Scale: A Psychometric Examination. Research Report R-126. Ottawa, ON: Correctional Service of Canada.

Philipse M (2005) *Psychopathy in the Treatment of Forensic Psychiatric Patients: Assessment, Prevalence, Predictive Validity, and Clinical Implications*. Amsterdam: Amsterdam University Press.

Projectteam SRM (1997) *WODC-Strafrechtmonitor: Verslag van de ontwikkeling van een systeem voor periodiek dossieronderzoek naar de achtergronden van de strafrechtspleging in Nederland*. Den Haag: WODC.

Shapiro DE (1999) The Interpretation of Diagnostics Tests. *Statistical Methods in Medical Research* 8: 113-34.

Schmidt P and Witte AD (1988) *Predicting Recidivism using Survival Models.* New York: Springer-Verlag.

Taylor R (1999) *Predicting Reconvictions for Sexual and Violent Offences using the Revised Offender Group Reconviction Scale*. London: Home Office/RDS.

Trout JD and Bishop MA (2002) 50 years of Successful Predictive Modelling Should Be Enough: Lessons for Philosophy of Science. *Philosophy of Science* 69: 197-208.

Tollenaar, N, & Van der Heijden, PGM (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.

Vergouwe Y (2003) *Validation of Clinical Prediction Models: Theory and Applications in Testicular Germ Cell Cancer*. Rotterdam: EMC, Erasmus University 143.

Webster CD, Douglas KS, Eaves SD and Hart SD (1997) Assessing Risk of Violence to Others. In: Webster CD and Jackson MA (eds) *Impulsivity: Theory, Assessment, and Treatment*. New York: Guilford, 251-77.

Wartna B, Tollenaar N and Bogaerts S (2009) Statrec: Inschatting van het recidiverisico van verdachten van een misdrijf. *Tijdschrift voor Criminologie* 53(3): 277-92.

Figure 2: Temporal validation of Statrec99 versus re-estimated Statrec-scales

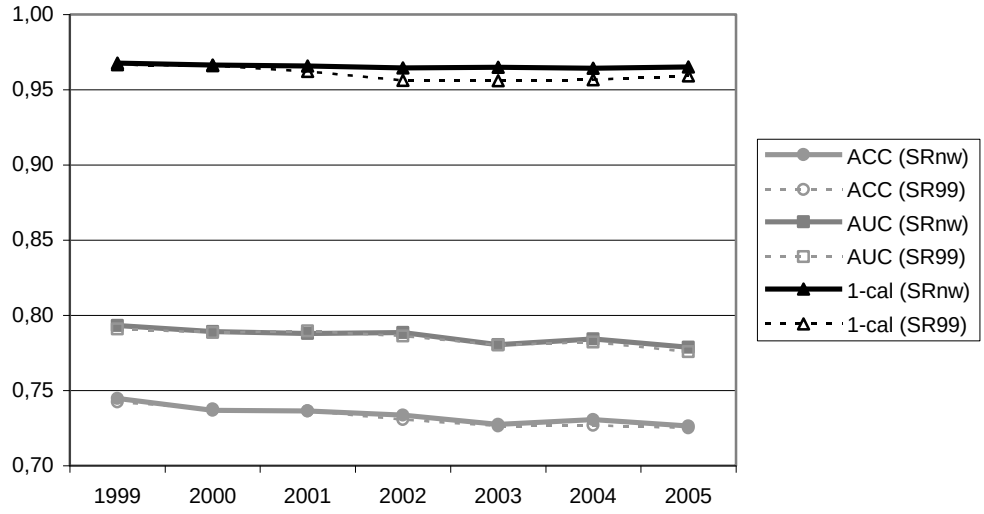

Figure 1: Schematic representation of temporal validation testing procedure

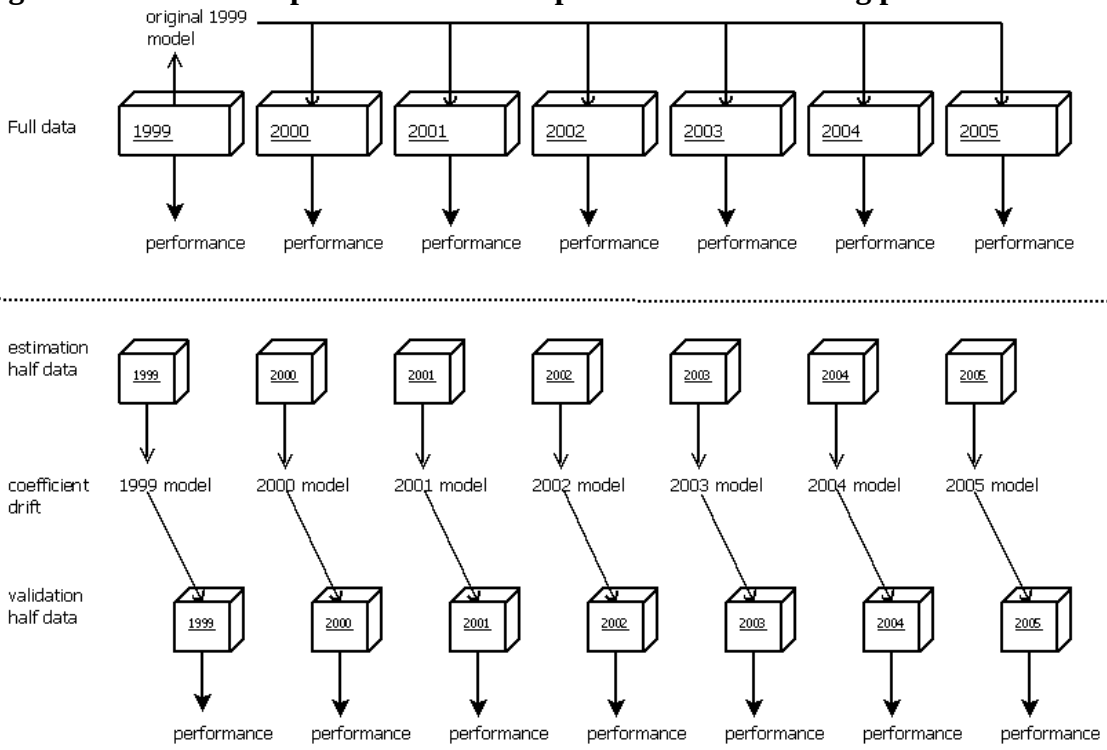

Figure 3: Coefficient drift of re-estimated Statrec-scales

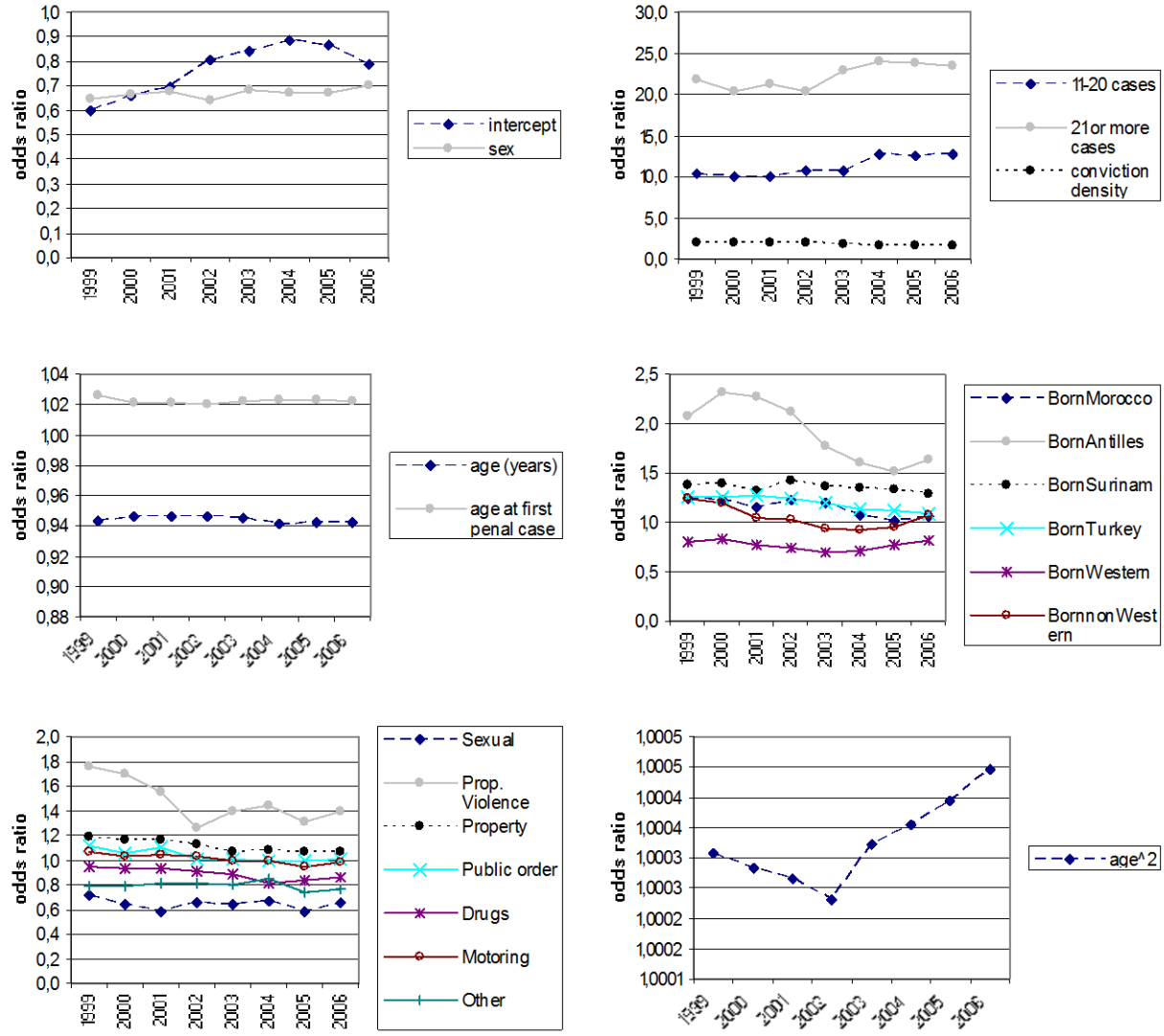

Table 1: Parameters of the 2005 StatRec model (N=159.298)

	Coefficient	Std. Error	Odds ratio
Predictor			
Constant	-0.147	0.026	0.863
Gender: male(0)/female(1)	-0.386	0.018	0.679
Age in years	-0.060	0.001	0.942
Age in years squared	0.0004	0.00004	10.004
Age at first conviction	0.024	0.002	1.024
Conviction density	0.453	0.045	1.573
Most serious offence			
Violence (reference)	0		1.00
Sexual	-0.525	0.078	0.591
Property with violence	0.270	0.053	1.310
Property without violence	0.072	0.020	1.075
Public order	-0.005	0.023	0.995
Drug offence	-0.194	0.027	0.823
Motoring offence	-0.029	0.019	0.972
Other offence	-0.307	0.023	0.735
Country of birth			
Netherlands (reference)	0		1.00
Morocco	0.028	0.034	1.029
Neth. Antilles/Aruba	0.412	0.033	1.510
Surinam	0.290	0.028	1.337
Turkey	0.109	0.034	1.115
Other Western countries	-0.263	0.023	0.769
Other non-Western countries	-0.049	0.024	0.952
log number of previous convictions	0.968	0.027	2.633
Dummy for 11-20 previous convictions	2.580	0.065	13.197
Dummy for 21 or more previous convictions	3.250	0.085	25.790

Figure 4a: calibration of Statrec99 on 2005 data

Figure 4b: plot Statrec99 vs. Statrec05 scores

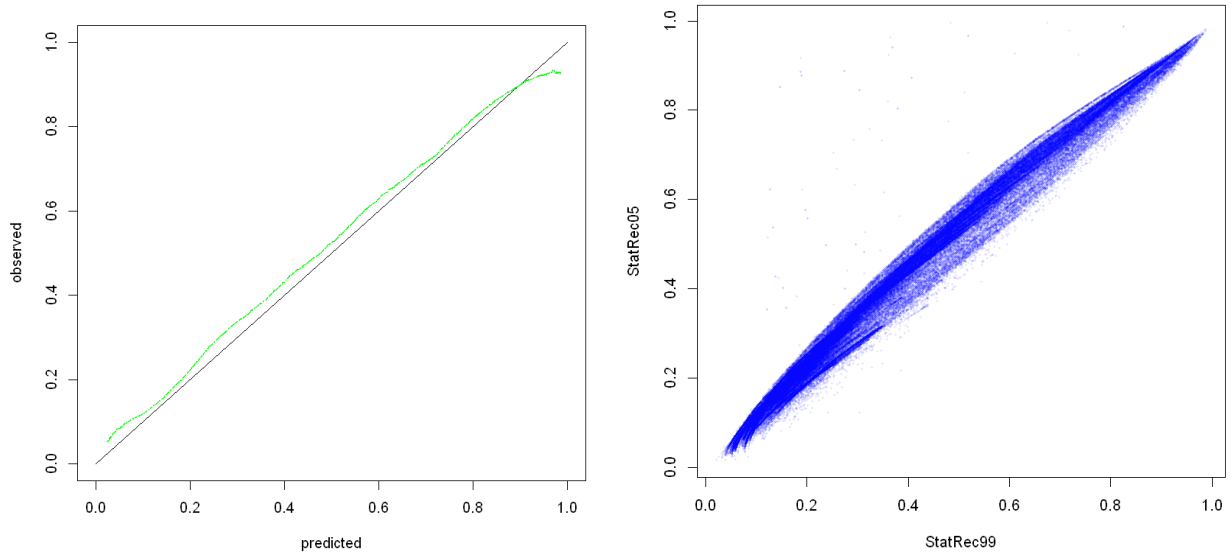

Figure 5: Calibration plot for StatRec: similarity between predicted and observed recidivism in the validation sample

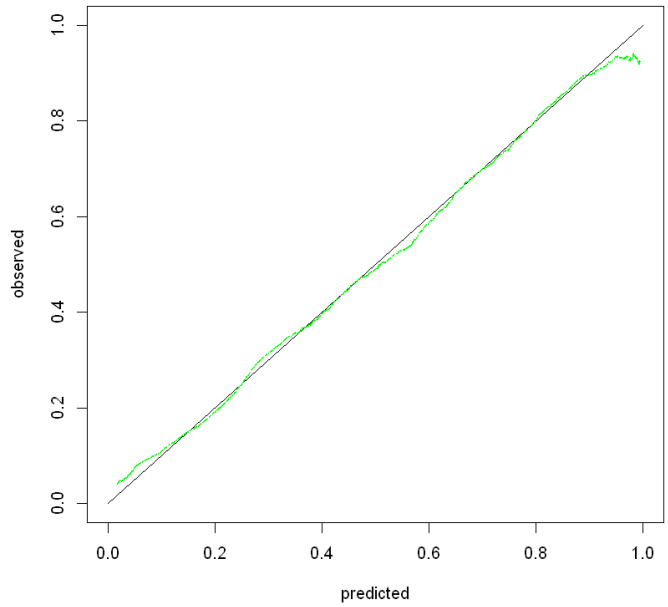

Figure 6: Strip plot of AUC, ACC, 1-cal and RIOC for the 19 court districts

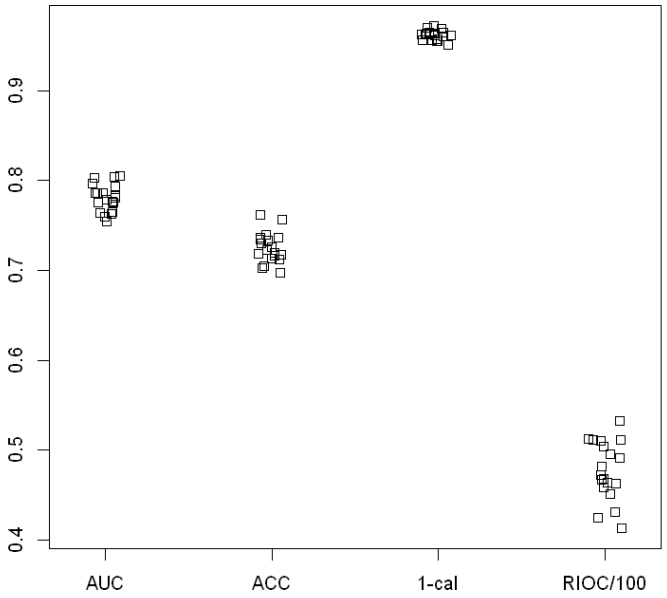

Table 2: Predictive validity of StatRec05 in various subpopulations[1]

Offender group	*N*	*Base rate*	*AUC*	*ACC*	*1-calerr*	*RIOC*
offenders	81,351	0,39	0,78	0,73	0,97	48.5
females	12,414	0.24	0.75	0.80	0.97	57.9
age<=24	18,740	0.47	0.75	0.69	0.95	44.2
age>=40	27,065	0.30	0.78	0.76	0.97	52.9
Born outside Netherlands	23,931	0.39	0.79	0.73	0.96	50.7
First offenders	13,527	0.33	0.68	0.68	0.96	34.9
More than 10 penal cases	9,567	0.65	0.65	0.63	0.95	14.7
Offense types						
Violence	11,442	0.44	0.77	0.70	0.96	42.4
Sexual	533	0.27	0.78	0.80	0.97	60.3
Property with violence	1,145	0.67	0.77	0.73	0.96	43.7
Property without violence	18,217	0.44	0.81	0.75	0.96	53.8
Public order	8,708	0.45	0.75	0.69	0.96	40.1
Drug offence	5,665	0.40	0.78	0.73	0.96	48.4
Motoring offence	24,710	0.35	0.74	0.71	0.96	43.2
Misc, offence	10,931	0.30	0.77	0.76	0.96	53.0
Disposal types						
measure	239	0.40	0.76	0.62	0.71	70.8
prison term	8,894	0.65	0.83	0.78	0.95	54.9
learning order	595	0.43	0.71	0.65	0.94	29.4
penal labour	13,977	0.48	0.75	0.69	0.96	38.9
suspended prison sentence	2,119	0.51	0.76	0.70	0.95	44.3
monetary sanction	48,730	0.33	0.74	0.72	0.97	43.7
other	800	0.14	0.81	0.89	0.87	60.8

no sanction	4,448	0.29	0.78	0.77	0.95	47.3
SRM suspects	1,239	0.49	0.80	0.72	0.92	50.5

Note: subpopulations of the general population consist of subsamples of the validation sample. The SRM suspects and ex-detainees are the complete samples.

Table 3: Logistic regression model of 4-year reconviction rate with STATREC-score and dynamic characteristics as predictors (N=1.239)

Predictor	Odds Ratio	95% C.I.			Predictor	Odds Ratio	95% C.I.		
STATREC score ($x\beta$)	2.22***	1.65	-	2.98	*Day programme*				
Civil status					Employed (ref.)	1			
Unmarried (ref.)	1				Student	0.62	0.41		0.93
Married	0.70	0.48	-	1.01	Occasional work	1.43	0.65	-	3.12
Divorced	1.29	0.71	-	2.34	Mainly care duties	0.34	0.16	-	0.71
			-	2.60					
Cohabitating	1.19	0.55			Nothing/unemployed	1.37	0.85	-	2.19
			-	0.94					
Widow/widower	0.44	0.21			Pension/old-age pension	1.22	0.09	-	17.19
Other	0.85	0.14	-	5.20	Invalidity benefit	1.11	0.49	-	2.53
Accomodation					Other	1.11	0.44	-	2.38
With parents	1.24	0.53	-	2.93	*Addiction*				
Alone + mother	1.13	0.40	-	3.17	Only hard drugs	1.32	0.70	-	2.51
Alone + father	2.03	0.13	-	32.26	Including hard drugs	2.22	0.28		17.87
Foster family	0.16	0.10	-	0.25	Only soft drugs	0.91	0.42	-	1.97
			-	5.25	Alcohol/soft drugs/			-	
Family	1.37	0.35			medication	1.84	0.24		14.39
Independent/alone	0.92	0.54	-	1.57	Only alcohol	1.55	0.68	-	3.53
Alone with child	0.51	0.33	-	0.79	Only gambling	0.46	0.27	-	0.80
Home	1.48	0.11	-	20.17	Medication	5.79	0.00	-	$11 \cdot 10^6$
With partner	1.13	0.64	-	1.99	No addiction (ref.)	1			
With partner and child (ref.)	1				*Residence status*				
			-	1.25					
Asylum seekers' centre	0.71	0.40			Dutch citizen (ref.)	1			
Without a permanent or									0.51
temporary address	2.11	0.35	-	12.78	Foreign national	0.37*	0.2		
Other	1.06	0.39	-	2.89	Asylum seeker	2.16	0.07	-	68.21
Ethnic background					Illegal alien	0.32	0.22	-	0.46
Dutch (ref.)	1				Tourist	0.03**	0.03	-	0.04
Surinamese	2.11	0.35	-	12.78	Not available	0.63	0.39	-	1.05
Antillean	1.06	0.39	-	2.89	*Constant*	0.77	0.25	-	2.35
Aruban	1.74	0.38	-	7.95					
Turkish	2.56*	0.28	-	23.71					
Moroccan	1.89	0.44	-	8.00					
Other	2.28	0.33	-	15.89					

Log likelihood = -629.47; Pseudo R2 = 0.267; AUC=0.829; LR chi^2 (43)=458.27; P(χ^2(df=43))<0.0001. Note: * >0,05; ** < 0,01; *** <0,001