Psychological Assessment

Predictive Validity of the Short-Term Assessment of Risk and Treatability for Violent Behavior in Outpatient Forensic Psychiatric Patients

Nadine A. C. Troquete, Rob H. S. van den Brink, Harry Beintema, Tamara Mulder, Titus W. D. P. van Os, Robert A. Schoevers, and Durk Wiersma

Online First Publication, December 15, 2014. http://dx.doi.org/10.1037/a0038270

CITATION

Troquete, N. A. C., van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2014, December 15). Predictive Validity of the Short-Term Assessment of Risk and Treatability for Violent Behavior in Outpatient Forensic Psychiatric Patients. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/a0038270

Psychological Assessment

© 2014 American Psychological Association
1040-3590/14/$12.00 http://dx.doi.org/10.1037/a0038270

Predictive Validity of the Short-Term Assessment of Risk and Treatability for Violent Behavior in Outpatient Forensic Psychiatric Patients

Nadine A. C. Troquete and Rob H. S. van den Brink
University Center for Psychiatry, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands

Harry Beintema
Mental Health Organisation Lentis, and Forensic Psychiatric Clinic dr S. van Mesdag, Groningen, the Netherlands

Tamara Mulder
Mental Health Organisation Drenthe, Assen, the Netherlands

Titus W. D. P. van Os
Mental Health Organisation Friesland, Leeuwarden, the Netherlands

Robert A. Schoevers and Durk Wiersma
University Center for Psychiatry, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands

It remains unclear whether prediction of violence based on historical factors can be improved by adding dynamic risks, protective strengths, selection of person-specific key strengths or critical vulnerabilities, and structured professional judgment (SPJ). We examine this in outpatient forensic psychiatry with the Short-Term Assessment of Risk and Treatability (START) at 3 and 6 months follow-up. An incident occurred during 33 (13%) out of 252 3-month and 44 (21%) out of 211 6-month follow-up periods (n = 188 unique clients). Pearson correlations for all predictor variables were in the expected directions. Prediction of recidivism based on historical factor ratings (odds ratio [*OR*] = 1.10) could not be improved through the addition of dynamic risk, protective strength, or key or critical factor scores (all *OR*s *ns*). The addition of the SPJ improved the model to modest accuracy (area under the curve [AUC] = .64) but made no independent significant contribution (*OR* = 1.55, p = .21) for the 3-month follow-up. For the 6-month follow-up, SPJ scores also increased predictive accuracy to modest (AUC = .67) and made a significant independent contribution to the prediction of the outcome (*OR* = 1.98, p = .04). Multicollinearity limits were unviolated. Limitations apply, however, results are similar to those from clinical, researcher rated samples and are discussed in the light of setting specific characteristics. Although it is too early to advocate implementing risk assessment instruments in clinical practice, we can conclude that clinicians in a heterogeneous outpatient forensic psychiatric setting can achieve similar results with the START as clinicians and research staff in more homogeneous inpatient settings.

Keywords: outpatient forensic psychiatry, Short-Term Assessment of Risk and Treatability (START), dynamic risk factors, predictive validity, violence risk assessment

Assessment of violence risk based on static historical factors, like young age at first offense or prior unauthorized leave, has proven to be predictive for violence occurring over a period of several years in forensic populations (de Vogel & de Ruiter, 2005; Doyle, Carter, Shaw, & Dolan, 2012; Hastings, Krishnan, Tangney, & Stuewig, 2011; Mudde, Nijman, van der Hulst, & van den Bout, 2011;). However, such historical, lifetime factors do not provide indications for treatment options to reduce violence risk.

Nadine A. C. Troquete and Rob H. S. van den Brink, University Center for Psychiatry, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands; Harry Beintema, Mental Health Organisation Lentis, and Forensic Psychiatric Clinic dr. S. van Mesdag, Groningen, the Netherlands; Tamara Mulder, Mental Health Organisation Drenthe, Assen, the Netherlands; Titus W. D. P. van Os, Mental Health Organisation Friesland, Leeuwarden, the Netherlands; Robert A. Schoevers and Durk Wiersma, University Center for Psychiatry, University of Groningen and University Medical Center Groningen.

This study was funded by a grant from ZonMw, The Netherlands Organisation for Health Research and Development (100 003 023). Further funding was provided by the participating Mental Health Organisations of Drenthe, Friesland, and Groningen (Lentis), and the University Center for Psychiatry of the University Medical Center Groningen. Harry Beintema, Tamara Mulder, and Titus W. D. P. van Os were the department heads of the clinics at which the RACE study was conducted, but they were not directly involved in the data collection. Findings from the randomized controlled trial were previously published by the current authors in *The British Journal of Psychiatry*, *202* (pp. 365–371) and are included in Nadine A. C. Troquete's doctoral dissertation. We express our great appreciation to all case managers, other staff, and clients of the Mental Health Organisations of Drenthe, Friesland, and Lentis, who contributed to this study.

Correspondence concerning this article should be addressed to Nadine A. C. Troquete, Rob Giel Research Center, University Center for Psychiatry, University Medical Center Groningen, CC72, PO Box 30.001 9700 RB Groningen, the Netherlands. E-mail: NadineTroquete@gmail.com

Management of risk for violence requires ongoing evaluation of current risk factors through the use of instruments that assesses risk for shorter periods of time, months rather than years, and consisting of factors responsive to treatment (Douglas & Skeem, 2005). Examples of such current, dynamic risk factors are external triggers, social skills, and emotional and mental state.

Inclusion of dynamic risk factors and short-term risk prediction are especially important in outpatient forensic psychiatry. For clients in inpatient settings, various restrictions apply: Treatment tends to be mandatory, contact with people outside the clinic is limited and/or supervised, drug use is actively discouraged, and triggers for violent behavior are kept to a minimum. Admission usually lasts for several years, providing ample opportunity for observation and intervention by treatment staff. For clients living in the community, the situation is quite different. Fewer restrictions apply, triggers are omnipresent, and contact with treatment staff is shorter and often less frequent and less comprehensive. Therefore, particularly in outpatient forensic psychiatry, there is a need for the identification of emerging risks through short-term risk assessment and prediction.

Moreover, Ward and coauthors (Ward & Brown, 2004; Ward, Mann, & Gannon, 2007; Ward & Stewart, 2003) argue that a sole focus on risks leads to a one-sided picture of the client and limits interventions to restrictive measures. This is contrary to psychiatric treatment aims focused at increasing a client's strengths, which in turn, are also thought to prevent recidivism (McGowan, Horn, & Mellott, 2011; Nicholls, Brink, Desmarais, Webster, & Martin, 2006; Ward & Brown, 2004; Ward et al., 2007; Ward & Stewart, 2003; Webster, Nicholls, Martin, Desmarais, & Brink, 2006). The Short-Term Assessment of Risk and Treatability (START; Nicholls et al., 2006; Webster et al., 2006) was developed to provide short-term dynamic risk assessment of both vulnerabilities and strengths of the client. The START combines several elements of various approaches to risk assessment, which are (a) a focus on dynamic risk factors, such as social skills, attitude and coping, for the assessment of short-term risk for violence; (b) assessment of these factors as both vulnerabilities, increasing risk for future violence, and strengths, reducing risk for future violence; (c) the selection of critical vulnerabilities and key strengths considered crucial for a specific client; and (d) a final risk estimate made by case managers weighing the identified vulnerabilities and strengths and applying clinical judgment, which is in line with the structured professional judgment (SPJ) approach (Doyle & Dolan, 2002). Although earlier instruments, for example, the Historical, Clinical, Risk Management–20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997), the Dynamic Appraisal of Situational Aggression (DASA; Ogloff & Daffern, 2006), the Spousal Assault Risk Assessment (SARA; Kropp & Hart, 2000) and the Structured Assessment of Violence Risk in Youth (SAVRY; Borum, Bartel, & Forth, 2003) encompass one or several of these elements, the START is the first instrument to include all of these in a single instrument.

Studies with the START are ongoing and publications are emerging from research groups in various countries (Australia: Chu, Thomas, Ogloff, & Daffern, 2011, 2013; Canada: Braithwaite, Charette, Crocker, & Reyes, 2010; Desmarais, Nicholls, Wilson, & Brink, 2012; Desmarais, van Dorn, Telford, Petrila, & Coffey, 2012; Nicholls et al., 2006; Wilson, Desmarais, Nicholls, & Brink, 2010; Norway: Nonstad et al., 2010; and the United Kingdom: Gray, Taylor, & Snowden, 2011). However, these studies suffer from several limitations. First, they are limited to small samples. Most commonly the sample size is approximately ≤50. The exception is a study by Desmarais, Nicholls, et al. (2012), who reported on 120 cases. However, they do not mention to what extent the various risk-assessment scores predict violent outcome. The second limitation of earlier studies is that they are restricted to inpatient settings. Aside from differences with outpatient settings noted in preceding paragraphs, the most common diagnosis in these clinical samples is psychosis (commonly 80% or more of sample, though see Gray, Benson, et al. (2011) for a sample with only 66%). A third limitation is that risk assessments are completed by research assistants rather than clinicians. Moreover, these assessments are based on case files instead of clinical contact. Useful as such an approach can be, it is not in line with the way the developers have described the STARTs intended purpose, namely as an "instrument intended to guide assessment and management of diverse populations of mentally and personality disordered persons and intended to act as a *clinical* indicator of treatment progress" (p. 323, emphasis added; Nicholls et al., 2006). Additionally, areas under the curve (AUC) for final prediction models rather than odds ratios (*OR*s) for the individual predictors tend to be reported, making it difficult to determine the incremental contributions of the specific elements introduced by the START (though, see Braithwaite et al., 2010; Chu et al., 2011; Desmarais, Nicholls et al., 2012; Desmarais, van Dorn, et al., 2012; Wilson et al., 2010 for clinical research assistant file-based results and practitioner assessments). Unfortunately, the only study to report on a large community sample with START assessments completed in daily practice by case managers (Nicholls, Petersen, Brink, & Webster, 2011) does not report on the predictive ability of the START scores. The proposed added benefits of the inclusion of both vulnerabilities and strengths in violence prediction and the selection of client specific key and critical factors, therefore, remain uncertain. Moreover, the question of whether the SPJ-based final risk estimate improves the prediction of violence remains largely uninvestigated (Lewis & Doyle, 2009).

To further complicate the interpretation of the findings from these earlier studies, important differences exist between countries in the demographic characteristics of clients treated in forensic psychiatry (Salize & Dressing, 2004; Salize, Dressing, & Peitz, 2002). Most studies concerning the START so far address populations with over 80% of clients having a psychotic disorder (though, see Gray, Benson, et al., 2011; Nicholls et al., 2011 for samples with 66% and 54%, respectively). In contrast, clients in outpatient forensic psychiatry in The Netherlands have a broad range of psychiatric disorders and only a minority of about 7% have a psychotic disorder (Troquete et al., 2013). These differences might also influence the START assessments ability to predict future violence for forensic psychiatric clients.

To address these concerns, the present study investigates whether (a) dynamic risk factor scores improve the prediction of future violent and criminal behavior above static, historical risk factor scores; (b) the addition of protective factor scores improves the prediction above and beyond one based on only scores of risk factors; (c) prediction should take all factor scores into account or should focus on the scores of the items considered most important for the individual client's risk of future violence; and (d) a final clinical judgment score improves the prediction of future violent and criminal behavior compared with simple numerical summing of risk-factor scores, as suggested by the SPJ approach. The START manual (Webster, Martin, Brink, Nicholls, & Desmarais, 2009) suggests that any prediction with the START scores will only be valid for the following 3 months. In line with this suggestion, we examine the hypotheses for a follow-up period of 3 months. Additionally, we investigate the predictive ability of the various START scores for a longer term of up to 6 months.

Method

Design and Setting

Data for the current study were collected as part of a larger study into Risk Assessment and Care Evaluation (RACE; trial number 1042, see www.trialregister.nl) in outpatient forensic psychiatry. RACE is a multisite clustered randomized controlled trial (RCT) conducted in the northern part of The Netherlands in three outpatient forensic psychiatric services. These services provide treatment for clients with psychiatric needs who have or are at risk of having contact with the criminal justice system (Wubs & Wijnen, 2005). At the time when the study was conducted, September 2007 until September 2010, formal and structured risk assessment occurred only on an infrequent, ad hoc basis.

The RACE study investigated the preventive effect of routine use of risk assessment as part of treatment plan discussions between case managers and their clients on new violent and criminal incidents. Additionally, this approach is hypothesized to result in better or more suitable treatment for clients, foster the development of a good therapeutic relationship between the client and the case manager, and have positive effects on client psychiatric and social well-being. For details of the trial and results of the RCT, see Troquete et al. (2013).

Subjects

Although all case managers were eligible for participation, we excluded those who were either expected to leave their post within 6 months or had no eligible clients ($n = 4$) and randomized the remaining 58 to either intervention ($n = 29$) or control ($n = 29$) group. Case managers in the intervention group were instructed to use the START (Webster et al., 2006) as part of the RACE protocol for all evaluations of client treatment plans. In The Netherlands, yearly treatment plan evaluation is compulsory. Case managers in the control group provided care as usual to their clients, which generally did not involve standardized risk assessment. The current article is concerned with the prediction of violent and criminal behavior with the START scores, therefore this report is limited to data collected in the intervention group.

The intervention group consisted of 29 case managers, most of whom were psychologists (29%), occupational therapists (25%), psychiatric nurses (18%), or specialists providing only forensic psychiatric home care (17%). These case managers had primary responsibility for the care planning of 558 eligible clients. However, 44% of these eligible clients were out of care before case managers completed the necessary baseline assessment of the client, so, ultimately, 310 (56%) clients were included in the intervention group. Risk assessments with the START completed for these clients are analyzed in the current article.

Instruments

Background information on clients (e.g., treatment order at start of treatment, psychiatric diagnoses, prior offenses or incidents) was provided by case managers who consulted their clients' case histories with the help of a protocol developed for this purpose. Although different guidelines exist for the interpretation of interrater correlation coefficients (ICCs) one set commonly used is that by Landis and Koch (1977). Reliability of an instrument's test scores is indicated as follows: .21–.40 is *fair*; .41–.60 is *moderate*; .61–.80 is *substantial*, and over .81 is *excellent* or *almost perfect*.

Short-Term Assessment of Risk and Treatability (START). Case managers in the intervention group of the RACE-study received the official training in the use of the official Dutch translation of the START (Nicholls et al., 2006; Webster et al., 2006) by the translators of the instrument ('t Lam, Lancel, & Hildebrand, 2009). The START is a structured professional judgment instrument consisting of 20 items scored both as vulnerabilities and strengths for a given client. There is room for the addition of two client-specific items. Each item can be scored as 0 (*absent*), 1 (*possibly present*), or 2 (*present*). After this initial scoring of all items, those of particular importance for the client are identified and marked as respectively critical vulnerabilities and key strengths. Taking all information into account, a final risk estimate is given for seven client outcomes: violence against others, self-harm, suicide, unauthorized leave, substance abuse, self-neglect, and victimization. Clients are scored as being either at 1 (*low*), 2 (*medium*), or 3 (*high*) risk for these outcomes. The present study only considers the final risk estimate for violence against others.

Previous investigations of START's psychometric properties have examined various follow-up periods, ranging from 30 days (Braithwaite et al., 2010) to 12 months (Desmarais, Nicholls, et al., 2012; Wilson et al., 2010) for various outcomes, such as severe violence (Nonstad et al., 2010), violence against others, self-harm, suicide attempts, unauthorized leave, substance abuse, self-neglect, victimization (Braithwaite et al., 2010), physical aggression against objects or people, verbal aggression (Gray, Benson, et al., 2011; Nicholls et al., 2006), and any inpatient aggression (Chu et al., 2011; Chu, Thomas, et al., 2013; Wilson et al., 2010). Excellent interrater reliability (ICCs between .81 and .95; Desmarais, Nicholls, et al., 2012; Nicholls et al., 2006; Wilson et al., 2010) and good internal consistency (Cronbach's α = .85–.87; Nicholls et al., 2006; Nonstad et al., 2010) have been reported for the START scores. Vulnerability scores were found to be modest (AUC = .66) to excellent (AUC = .83) predictors of violent outcome, whereas strength scores ranged from modest (AUC = .65) to acceptable (AUC = .77; Braithwaite et al., 2010; Chu et al., 2011; Chu, Thomas, et al., 2013; Desmarais, Nicholls, et al., 2012; Nonstad et al., 2010; Wilson et al., 2010). Reports on the predictive validity of the final risk estimate for violence against others range from no better than chance (AUC = .52) to excellent (AUC = .82; Braithwaite et al., 2010; Chu et al., 2011; Wilson et al., 2010). Braithwaite et al. (2010) are the only ones to report on predictive accuracy for the other six risk estimates, which was, for the most part, no better than chance (.42–.55), except for substance abuse, which was acceptable at .78.

To determine interrater reliability in our sample, case managers identified 30 clients well known by a second clinician in the intervention arm of the study. This second clinician was then asked to fill out a START form for the client at the same time as, but without conferring with, the primary case manager. The interrater reliability of the various test scores was assessed by a two-way, random, absolute agreement analysis resulting in ICCs. Results were as follows: sum of vulnerability scores (ICC = .64, $p < .01$); sum of strength scores (ICC = .49, $p < .01$); mean of scores on critical vulnerability items (ICC = .32, $p = .05$); mean of scores on key strength items (ICC = .30, $p = .07$); and the SPJ score of violence against others (ICC = .58, $p < .01$).

Historical, Clinical, and Risk Management–20 (HCR–20). The START manual (Webster et al., 2009) contains the instruction to also score the 10 historical items (H10) of the HCR–20 (Webster et al., 1997) as a baseline for risk assessment. Therefore, case managers received instructions, full descriptions and scoring guidelines from the HCR–20 manual on how to score the historical items and did so as part of the baseline assessment of the client, independently of the first risk assessment with START. The HCR-20 is a scheme case managers are familiar with from clinical practice, and particularly the H10 items consist of information commonly available in clients' case files. Formal training of case managers in the completion of the H10 was beyond the scope and means of the study. Case managers were provided with the full item descriptions, and at least the first three protocols were completed under supervision of research staff. Research staff checked all H10 items for consistency with case file information provided by case managers at the same time. If ambiguities in coding were found, case managers were asked for clarification. Individual items are scored as either 0 (*definitely absent or does not apply*), 1 (*possibly* or *partially present*), or 2 (*definitely* or *clearly present*). Consequently, the sum score for the 10 historical items ranges from 0 to 20. The reliability and predictive validity of the scores of the HCR–20 for violent outcome have been reported in numerous studies (e.g., Douglas, Blanchard, Guy, Reeves, & Weir, 2010). In forensic psychiatric samples substantial to excellent values for the interrater reliability of the HCR–20 scores tend to be found (ICCs > .70; Douglas et al., 2010). In general, H10 items have been reported to be modest to acceptable predictors of violent outcome (AUC range: .60–.79; Arbach-Lucioni, Andrés-Pueyo, Pomarol-Clotet, & Gomar-Soñes, 2011; Douglas et al., 2010; Doyle et al., 2012; Mudde et al., 2011), although some studies have reported values over .80 (e.g., Telles, Folino, & Taborda, 2012). Assessment of interrater reliability of H10 scores in our sample was beyond the scope of the study.

Outcome

Outcome consists of all violent or criminal behavior as reported by the case manager in the client's case file. Violent behavior includes intentional behavior with the potential to physically harm a person or animal and seriously threatening or intimidating aggression. Criminal behavior additionally covers exhibitionism, possession of child pornography, stalking, drug dealing, driving without a license or while under the influence of alcohol or drugs, possession of an illegal weapon, vandalism, and theft. Not included is the use of illegal drugs because this is not considered a crime under Dutch law. This definition of violent outcome is somewhat wider than commonly used for risk-assessment instruments like the START and the HCR-20. However, interpersonal violence, violence against animals, and criminal behavior are all used in outpatient forensic psychiatry as indicators of the success, or lack thereof, of a client's treatment. They have a signaling function indicating a further need for intervention. Therefore we considered these types of incidents as an appropriate part of the outcome in the setting studied. However, to facilitate comparisons with earlier studies, we also include models for the prediction of the outcome as defined by the START manual. Case managers recorded incidents that could potentially satisfy either definition on a standard form and included them in the client's files. New arrests, charges, or convictions were recorded in a similar manner. At follow-up, research assistants, blind to client randomization status, collected the forms and additionally checked all case files for any potential incident. If they came across any indication that a potential incident had occurred and no form was completed, they did so at the time. Research assistants were instructed to err on the side of caution during this process. Three outpatient forensic psychiatric experts, also blind to client randomization status, then determined, through consensus, whether individual reports should be considered a violent or criminal incident according to the preceding definitions.

For the analyses reported here, we recoded the outcome as either 0 (*absence*) or 1 (*presence*) of one or more incidents during the client's follow-up period. As planned, we did so separately for the first 3 months (1–3) and for the first 6 months (1–6) following the completion of the START by the case manager.

Procedure

According to protocol, all clients treated by case managers in the intervention group were to receive the full RACE intervention at least once every 6 months. The current article reports on the first step of the intervention, which consisted of the completion of the START by the case manager prior to a formalized treatment-plan discussion with the client (Step 2). The treatment plan discussion focused on the items of the START identified as key strengths or critical vulnerabilities of the client. For further details of the procedure, see Troquete et al. (2013). Case managers informed their clients, in word and writing, about the study and explained that data would be collected anonymously and by independent researchers to evaluate the new method of care planning. The Dutch Medical Ethical Committee for Mental Health Care approved the study. Because of the nature of the RACE intervention—START assessment by both case manager and client and consequent shared decision making to formulate a treatment plan—neither case manager nor client could be blind to this process.

Analyses

Hierarchical logistic regression was conducted with the occurrence of incidents during follow-up (*absent* or *present*) as outcome variable. Sum scores (possible range: 0–40) for both the vulnerability and strength scales of the START were calculated in accordance with the instructions in the manual, correcting for up to four missing items per scale (Webster et al., 2009). Depending on the model tested, the independent variables were sum of the historical factor ratings (H10) of the HCR–20; sum of the vulnerability scores; sum of the strength scores; the mean score of the vulnerabilities identified as critical items; the mean score of the strengths identified as key items; and the case managers' rating on the final risk estimate for violence against others.

AUCs were determined with receiver operating characteristics (ROC) analyses for all models. AUCs reflect predictive accuracy and those over .90 are considered as *outstanding*, those between .80 and .89 as *excellent*, between .70 and .79 as *acceptable*, between .60 and .69 as *modest*, with those around .50 considered *no better than chance* (Hosmer & Lemeshow, 2000).

The sum of the scores on the H10 items made up the first block of the model. The second step consisted of two parallel analyses in which either the sum of the vulnerability scores or the mean of the critical vulnerability scores was entered as the second block of the model. Both models were then elaborated with a third block, which added either the sum of the strength ratings or the mean of the key strength scores to the model. In the last block, the score for the final risk estimate for violence against others was added. To ensure proper testing of the hypotheses, nonsignificant predictors from previous steps were retained in the models to test the incremental predictive value of predictors entered in the subsequent step. This procedure was conducted separately for cases with 3-month and 6-month follow-up information available after START assessment. First, we composed the models for the more broadly defined outcome which was more suitable for our setting, second we repeated the analyses for the outcome as defined by the START manual. All analyses were carried out with PASW statistics 20 (SPSS, 2012).

Results

The 29 case managers were slightly more often female (55%), on average 43 years old ($SD = 11$; range: 23–59), without a university degree (66%); that is, they were psychiatric nurses or occupational therapists rather than psychologists or psychiatrists, and had a mean of 7 years ($SD = 6$ years; range: 0–20 years) of experience working in forensic psychiatric care. Size of caseload varied widely (2–40 clients), with each case manager being principally responsible for an average 17 clients ($SD = 11$).

Most of the 310 clients in the intervention group were male (94%), aged 40 years on average ($SD = 11$), with personality disorders (69%)—mostly personality disorder not otherwise specified (33%) or Cluster B (26%)—substance-related disorders (38%), impulse control disorders (27%), mood disorders (21%), and paraphilia (20%). Only 7% had a psychotic disorder, and 7% had no diagnosis on Axis I. Mean length of treatment before inclusion was 26 months ($Mdn = 16$, $SD = 24$; range: 1–116 months). Clients had mostly committed violent (56%), property (37%), or sexual (32%) offenses, but a fair proportion (15%) had also been involved in substance-related offenses. The majority of clients were treated either voluntarily (55%) or were on probation (28%). Similar findings with respect to age, gender, and diagnosis were reported by Bouman, van Nieuwenhuizen, Schene, and de Ruiter (2008), which is an outpatient forensic psychiatric sample from another part of The Netherlands.

All clients ($N = 310$) were supposed to receive at least one, but preferably multiple, interventions, including risk assessments with the START; however, only 201 (65%) did. The 109 clients without an intervention had been significantly longer in treatment before inclusion ($M = 30$ v. 23 months, $t = 2.28$, $p = .02$), and were less likely to have a personality disorder (62% v. 73%, $\chi^2 = 3.84$, $p = .05$). Additionally, they tended to be more likely to be under probation (36% v. 26%, $\chi^2 = 3.61$, $p = .06$), less likely to have committed a sex offense with a victim aged 16 or younger (15% v. 24%, $\chi^2 = 3.38$, $p = .07$), and less likely to have a mood disorder (15% v. 24%, $\chi^2 = 3.70$, $p = .06$). There were no other significant differences (all ps $> .10$).

In total, case managers completed 326 STARTs for clients in the intervention group (range 1–6 per client). Low frequencies for third, fourth, fifth, and sixth assessments led us to exclude these assessments from the current analyses, resulting in the selection of 200 first and 85 second START assessments. Psychometric properties for the START assessments reported in following paragraphs are based on the 200 initial assessments. For 32 assessments, no appropriate follow-up period could be determined, and for one case the second START was completed during the follow-up period of the first START assessment. Therefore, these 33 assessments had to be excluded from the logistic regression analyses, resulting in 252 risk assessments ($n = 188$ unique clients) with at least a 3-month follow-up period available. For 211 risk assessments ($n = 163$ unique clients) a 6-month follow-up period was available.

During the first 3 months following START assessment, 13% ($n = 33$) of clients had at least one incident, as defined for the RACE-study, which increased to 21% ($n = 44$) over the next 3 months. See Table 1 for frequencies of specific types of incidents. When the stricter definition of the START manual was used, these numbers dropped to 11% ($n = 27$) and 17% ($n = 36$), respectively. The low absolute numbers for the specific types of incidents prevented us from conducting meaningful statistical analyses for the individual outcome categories (sexual assault, physical violence, threatening aggression, etc.).

There was good dispersion on the START in the sense that for all items the full range of options (0–2) was used. Table 2 presents the psychometric properties of individual items for the 200 first START assessments. The mean score for the sum of strengths was 24.3 ($SD = 7.3$; range 6–40) and the mean score for the sum of vulnerabilities was 14.0 ($SD = 6.7$; range 0–34). The average mean score of the selected key strengths was 1.7 ($SD = 0.4$; range 0–2), indicating that most key strengths identified were considered

Table 1
Proportion of Follow-up Periods in Which an Incident Was Observed

	Months 1–3[a]		Months 1–6[b]	
Incident	*n*	%	*n*	%
Sexual assault, victim ≤16	3	1.2	3	1.4
Sexual assault, victim >16 or unspecified	2	0.8	2	0.9
Physical violence	9	3.6	16	7.6
Threatening aggression[c]	17	6.7	21	10.0
Stalking	1	0.4	1	0.5
Property offence	7	2.8	9	4.3
Substance-related offense	2	0.8	5	2.4
Any violent or criminal[d]	33	13.1	44	20.9

Note. No arson related incidents occurred.
[a] $n = 252$. [b] $n = 211$. [c] Includes threatening or intimidating verbal and nonverbal aggression. [d] Numbers do not add up because of multiple incidents of individual clients and individual incidents fitting multiple categories.

Table 2
Psychometric Characteristics of the START in Outpatient Forensic Psychiatry (n = 200; First STARTs) in Percentages

START item	*M*	*SD*	Key/critical item	Minimally present	Moderately present	Maximally present
1. Social skills						
Strength	1.13	0.60	13.5	12.5	62.0	25.5
Vulnerability	0.92	0.58	12.0	21.0	66.0	13.0
2. Relationships						
Strength	1.09	0.62	18.5	15.0	61.5	23.5
Vulnerability	0.91	0.64	26.5	25.1	58.8	16.1
3. Occupational						
Strength	1.26	0.79	34.7	21.1	31.7	47.2
Vulnerability	0.62	0.71	14.1	51.3	35.7	13.1
4. Recreational						
Strength	1.21	0.74	20.7	18.7	41.9	39.4
Vulnerability	0.70	0.73	12.1	45.5	38.9	15.7
5. Self-care						
Strength	1.57	0.56	16.0	3.5	36.5	60.0
Vulnerability	0.35	0.54	5.5	68.5	28.5	3.0
6. Mental state						
Strength	1.21	0.65	5.0	12.6	53.8	33.7
Vulnerability	0.71	0.64	14.5	39.5	50.5	10.0
7. Emotional state						
Strength	0.97	0.52	8.0	15.0	73.0	12.0
Vulnerability	1.12	0.59	32.5	12.0	64.5	23.5
8. Substance use						
Strength	1.38	0.78	22.3	18.3	25.4	56.3
Vulnerability	0.60	0.76	20.8	57.1	26.0	16.8
9. Impulse control						
Strength	1.01	0.61	9.0	18.5	62.5	19.0
Vulnerability	0.99	0.66	31.7	22.1	56.3	21.6
10. External triggers						
Strength	1.14	0.70	4.6	18.3	49.2	32.5
Vulnerability	0.74	0.73	13.1	42.7	40.7	16.6
11. Social support						
Strength	1.20	0.67	29.6	14.6	50.8	34.7
Vulnerability	0.78	0.67	16.7	35.4	51.0	13.6
12. Material resources						
Strength	1.30	0.67	15.5	11.6	47.2	41.2
Vulnerability	0.59	0.68	10.6	52.3	36.7	11.1
13. Attitudes						
Strength	1.16	0.64	9.5	13.5	57.5	29.0
Vulnerability	0.68	0.60	13.0	39.0	54.0	7.0
14. Medication adherence						
Strength	1.31	0.87	13.6	27.1	15.1	57.8
Vulnerability	0.26	0.55	8.5	79.9	14.6	5.5
15. Rule adherence						
Strength	1.52	0.63	11.0	7.5	33.5	59.0
Vulnerability	0.36	0.56	4.0	68.5	27.5	4.0
16. Conduct						
Strength	1.33	0.60	7.0	6.5	53.8	39.7
Vulnerability	0.57	0.59	4.5	47.7	47.2	5.0
17. Insight						
Strength	1.13	0.62	16.0	13.5	60.0	26.5
Vulnerability	0.80	0.63	13.5	32.0	56.0	12.0
18. Planning						
Strength	1.07	0.66	7.1	18.2	56.6	25.3
Vulnerability	0.70	0.66	7.1	41.4	47.5	11.1
19. Coping						
Strength	0.89	0.59	6.1	23.7	63.6	12.6
Vulnerability	1.17	0.60	38.2	11.1	60.8	28.1
20. Treatability						
Strength	1.48	0.65	23.1	8.5	34.7	56.8
Vulnerability	0.51	0.63	3.5	56.3	36.7	7.0

Note. START = Short-Team Assessment of Risk and Treatability; SD = standard deviation.

to be *definitely present*. For the critical vulnerabilities, the average mean score was 1.3 (SD = 0.5; range 0–2), indicating that case managers mostly considered the identified vulnerabilities as being *possibly present*. The mean score for the risk estimate of violence against others was 1.4 (Mdn = 1.0, SD = 0.6; range 1–3), reflecting that 60% of clients were judged to be at minimal risk, 36% at moderate risk, and 4% at high risk of being involved in a new violent incident.

The clients who had an incident during the first 3-month follow-up did not differ significantly from the clients without an incident on scores for strengths, vulnerabilities, key strengths, critical vulnerabilities, or the final risk estimate for violence against others (all ps $>$ .10). However, clients who had an incident during the 6 months following risk assessment did have a significantly higher score on the risk estimate (M = 1.6 vs. 1.4; t = 2.33, p = .02) and showed a trend for a lower score on the strengths scale (M = 23.3 vs. 25.3; t = 1.73, p = .08) and a higher score on the vulnerabilities scale (M = 15.2 vs. 13.4; t = 1.66, p = .10). There were no significant differences in scores of key strengths or critical vulnerabilities.

All Pearson correlations for the variables were in the expected directions (see Table 3). There was a significant negative correlation (–.71) between the sum score of the vulnerabilities and the sum score of the strengths. Earlier reports (e.g., Braithwaite et al., 2010) have suggested that multicollinearity between vulnerability and strength scales might exist. If this is the case in the current sample, then the predictive abilities of the models would be undermined because the effects of the different predictors could not be separated; that is, they would explain the same variance and the best predictor could not be identified. Therefore, the various models reported below were checked for indications of multicollinearity by examining the variance inflation factor (VIF) and tolerance values. No violations of limits were found (VIF range: 1.00–2.00; tolerance between 0.45 and 1.00), indicating that the vulnerability and strength scales, though significantly correlated, made independent, but nonsignificant, contributions to the models.

Results for the hierarchical logistic regression analyses are shown in Table 4 for the first 3-month follow-up and in Table 5 for the 6 months following START assessment. These sets of analyses used the broader definition of the outcome as the dependent variable. The best model to predict future incidents over a short-term (3-month) period was found to consist of scores on historical, strength, and vulnerability items and on the final risk estimate. There was little difference between the model including full strength and vulnerability scales and the model with the key and critical items (full scales: Nagelkerke R^2 = .04; AUC = .62, p = .03; key and critical items: Nagelkerke R^2 = .05; AUC = .64, p = .01). None of the variables in these models had an independent significant contribution.

Table 3
Pearson Correlations Between Predictor Variables (n = 285 STARTS)

Variable	1	2	3	4	5	6
1. H10	—					
2. Sum vulnerabilities	.31**	—				
3. Sum strengths	−.29**	−.71**	—			
4. Mean critical	.11	.63**	−.49**	—		
5. Mean key	−.24**	−.32**	.44**	−.08	—	
6. Final risk estimate for violence	.12*	.44**	−.33**	.24**	−.18**	—

Note. START = Short-Term Assessment of Risk and Treatability.
* p < .05. ** p < .01.

The models testing the hypotheses for 6-month follow-up also showed that the best prediction of future violence was achieved when scores on historical, strength (either full scale or only key items), and vulnerability items (either full scale or only key items), and the score on the final risk estimate were included (for model with full scales: Nagelkerke R^2 = .07; AUC = .65, p < .01; for model with critical and key items: R^2 = .08; AUC = .67, p < .01). In these final two models, both the scores on the historical items and on the final risk estimate made significant independent contributions to the prediction of future violent behavior.

Both sets of analyses were repeated with the stricter outcome, as defined by the START manual, as the dependent variable. Tables 6 and 7 show that these results do not change the conclusions obtained with the broader definition of the outcome.

We completed several sensitivity analyses to examine the robustness of our results. First, given the range of completed STARTs per individual case manager (1 to 27), the experience of the case manager with the START could have influenced our findings. Therefore, we repeated these analyses excluding the assessments from inexperienced (less than 5 STARTs completed) case managers (n = 5). For the 3-month follow-up 241 START assessments remained. For the 6-month follow-up, there were 202 START assessments. Second, our findings could have been influenced by the inclusion of duplicate ratings completed for the same client. As such, this is a violation of the assumption of independent observations, which in general we would not advocate. However, given the low frequencies of the outcome, we felt that including duplicate ratings would improve the robustness of the models and thus outweigh violating the assumption. To make certain that this was not problematic in this case, we repeated the analyses including only the first START assessment for each unique client (n = 188 for 3-month follow-up, and n = 163 for 6-month follow-up). Findings from both sets of analyses were in line with reported results, therefore a certain lack of independence in the H10 and START ratings seems not to have influenced our findings (tables available from first author).

Discussion

Previous studies investigating the predictive validity of the START for violent outcomes have suffered from limitations in sample size, lack of diversity in psychiatric diagnoses, restrictions to inpatient settings, assessments based on case files, rating by research assistants, or a combination of these limitations. The current study was conducted in outpatient forensic psychiatry with clients who lived in the community, had a diversity of psychiatric problems, and ample opportunity to engage in new transgressions because fewer restrictions apply. Additionally, data were collected as part of a cluster randomized clinical trial in which case managers scored clients on the START as part of routine treatment plan evaluations. This provided an opportunity to examine the predictive qualities of the START assessments when completed in daily clinical practice.

Table 4
Hierarchical Logistic Regression for Violent and Criminal Outcome in the First 3 Months Following Risk Assessment With the START (n = 252)

Step	Variable	*OR*	95% CI	*p*	R^2	AUC	95% CI	*p*
1	Sum: Historical	1.08	[0.98, 1.19]	.11				
					.02	.59	[.48, .69]	.11
2a	Sum: Historical	1.07	[0.97, 0.19]	.19				
	Sum: Vulnerabilities	1.02	[0.96, 1.08]	.48				
					.02	.59	[.49, .70]	.09
2b	Sum: Historical	1.08	[0.98, 1.19]	.12				
	M: Critical vulnerabilities	1.03	[0.48, 2.19]	.95				
					.02	.58	[.48, .69]	.12
3a	Sum: Historical	1.07	[0.96, 1.18]	.21				
	Sum: Vulnerabilities	1.00	[0.93, 1.09]	.88				
	Sum: Strengths	0.98	[0.91, 1.05]	.57				
					.03	.59	[.49, .70]	.08
3b	Sum: Historical	1.07	[0.97, 1.18]	.20				
	M: Critical vulnerabilities	0.97	[0.45, 2.11]	.94				
	M: Key strengths	0.60	[0.25, 1.42]	.24				
					.03	.59	[.50, .70]	.07
4a	Sum: Historical	1.08	[0.98, 1.20]	.13				
	Sum: Vulnerabilities	1.00	[0.92, 1.09]	.95				
	Sum: Strengths	1.00	[0.92, 1.08]	.94				
	Risk estimate for violence	1.55	[0.77, 3.10]	.22				
					.04	.62	[.52, .72]	.03
4b	Sum: Historical	1.08	[0.97, 1.20]	.16				
	M: Critical vulnerabilities	0.78	[0.35, 1.76]	.55				
	M: Key strengths	0.62	[0.26, 1.50]	.29				
	Risk estimate for violence	1.55	[0.79, 3.04]	.21				
					.05	.64	[.54, .73]	.01

Note. START = Short-Term Assessment of Risk and Treatability; OR = odds ratio; CI = confidence interval; R^2 = Nagelkerke R^2; AUC = area under curve.

For both the outcome as defined by the START manual as well as a somewhat broader definition of the outcome, which suited our setting better, we completed the following analyses: First, we examined whether the prediction of future violent behavior based on scores on static, historical risk factors, could be improved by the inclusion of scores on dynamic risk factors. This was not the case, regardless of whether all dynamic risk factor ratings were taken into account or only those that were identified as critical for the client. We then examined whether the further addition of dynamic strength scores could improve a prediction model already consisting of ratings on historical and dynamic risks. Neither inclusion of scores for the full strength scale nor inclusion of scores for only the key strengths improved the models significantly, regardless of the follow-up period examined. These various additions did not improve model accuracy and none of the individual predictors reached statistical significance, except for the scores on historical risk factors when predicting for a longer (6-month) follow-up period. This latter finding seems to reflect the general ability of historical risk-factor ratings to predict recidivism over a longer period, as established extensively in other studies (Douglas et al., 2010).

The last addition to both models consisted of the scores on the final risk estimate for violence risk against others. For the shorter follow-up period (3 months), the model was only minimally improved, with none of the individual predictors reaching statistical significance, although the model as a whole was significant and modestly accurate (AUC = .64, p = .01) in predicting recidivism. For the longer follow-up period of 6 months, 8% of total variance could be explained with modest accuracy (AUC = .67, $p < .01$) by a model including scores on historical, critical, and key items and the final risk estimate score. A similar model, but including all vulnerability and strength scores, explained 7% of variance with modest accuracy (AUC = .65, $p < .01$). In both models, the ratings on the historical items of the HCR–20 (*OR* = 1.10 for both models) and on the SPJ for violence against others (*OR* = 1.98 for the model with full scales, and *OR* = 1.83 for the model with only critical and key items) made significant independent contributions to the prediction of future violence. None of the scores on the vulnerabilities, strengths, critical or key items provided significant, independent contributions to the prediction model. Therefore, the first two research questions regarding whether prediction of future violence based on only historical risk-factor scores can be improved by inclusion of ratings on dynamic risk and protective factors must be answered in the negative.

Third, we were interested in the role of the assessment of key strengths and critical vulnerabilities in the prediction of recidivism. Specifically, could a model be defined by including only the scores on those key and critical items, or, alternatively, were full vulnerability and strength total scales necessary to achieve satisfactory prediction? On the basis of the logistic regression analyses, the answer seems to be that the choice is arbitrary because models with scores on full scales did not outperform those with only ratings of key and critical items or vice versa.

The final research question concerned the contribution of the assessment of risk for violence against others. For short-term (up to 3 months) prediction of violence, inclusion of the final risk

Table 5
Hierarchical Logistic Regression for Violent and Criminal Outcome in the First 6 Months Following Risk Assessment With the START ($n = 211$)

Step	Variable	*OR*	95% CI	*p*	R^2	AUC	95% CI	*p*
1	Sum: Historical	1.11	[1.01, 1.21]	.03	.04	.61	[.52, .70]	.03
2a	Sum: Historical	1.09	[0.99, 1.20]	.07	.04	.62	[.52, .71]	.02
	Sum: Vulnerabilities	1.03	[0.98, 1.08]	.30				
2b	Sum: Historical	1.11	[1.01, 1.21]	.03	.04	.61	[.52, .71]	.02
	M: Critical vulnerabilities	0.92	[0.46, 1.85]	.82				
3a	Sum: Historical	1.09	[0.99, 1.20]	.09	.05	.62	[.53, .72]	.01
	Sum: Vulnerabilities	1.02	[0.95, 1.09]	.66				
	Sum: Strengths	0.98	[0.92, 1.05]	.54				
3b	Sum: Historical	1.10	[1.00, 1.21]	.05	.04	.61	[.52, .70]	.03
	M: Critical vulnerabilities	0.88	[0.44, 1.78]	.72				
	M: Key strengths	0.67	[0.44, 1.78]	.34				
4a	Sum: Historical	1.10	[1.00, 1.22]	.05	.07	.65	[.56, .74]	<.01
	Sum: Vulnerabilities	1.00	[0.93, 1.08]	.93				
	Sum: Strengths	1.00	[0.93, 1.07]	.89				
	Risk estimate for violence	1.83	[0.96, 3.50]	.07				
4b	Sum: Historical	1.10	[1.00, 1.22]	.05	.08	.67	[.58, .76]	<.01
	M: Critical vulnerabilities	0.66	[0.31, 1.40]	.28				
	M: Key strengths	0.73	[0.31, 1.70]	.46				
	Risk estimate for violence	1.98	[1.05, 3.74]	.04				

Note. START = Short-Term Assessment of Risk and Treatability; OR = odds ratio; CI = confidence interval; R^2 = Nagelkerke R^2; AUC = area under curve.

rating did not improve the prediction. However, for the longer term of up to 6 months, the final risk estimate of the case manager did provide an independent, significant contribution to the prediction of future violence. Clients estimated to be at high risk were twice as likely as those considered to be at medium risk and four times as likely as those estimated to be at low risk, to recidivate preceding 6-month follow-up. This substantiates the structured professional judgment approach of risk assessment (Doyle & Dolan, 2002). As such, this finding is not new, neither for the START nor for risk-assessment instruments in general. For instance, for the START, Desmarais, Nicholls, et al. (2012) showed that the scores on the final risk estimates not only predicted future violence above chance level, but also significantly improved predictions made with only the H10 and the START total scores. Including a SPJ seems to improve predictions in general, as both a recent systematic review (Tully, Chou, & Browne, 2013) and examinations of specific instruments like the HCR–20 (Douglas et al., 2010) and the DASA (Chu, Daffern, & Ogloff, 2013) have shown. However, most of these findings are based on research assistants coding from case files. As such, it is encouraging that our own study, in which case managers completed risk assessments and gave the SPJ in clinical practice, supports these earlier findings.

That we did not find proof for incremental predictive validity for including either the dynamic risk or strength scores is contrary to expectations. Previous research has found the opposite, for risk assessment instruments in general and the START in particular. For instance, both Wilson et al. (2010) and Desmarais, Nicholls et al. (2012) found incremental predictive validity for both assessments of the STARTs vulnerability and strength factors over and above that achieved with the H10 total scores. It has to be noted, though, that both these studies relied on retrospective assessments of case files completed by research assistants. This in contrast to our own approach, in which clinicians in daily practice completed both measures. This potentially explains the differences in findings regarding the predictive validity of the START assessments for violent outcome. With respect to findings for other risk-assessment instruments, it is less clear to which factors our different findings should be contributed. For instance, McGowan et al. (2011) found that including the scores on protective factors of the SAVRY improved the prediction of violence. Similarly, the Clinical and Risk Management total scales of the HCR–20, both dynamic in nature, have shown incremental predictive validity over that of the Historical subscale, although it has to be noted that most of these studies also used a retrospective file-based assessment (O'Shea, Mitchell, Picchioni, & Dickens, 2013).

The global finding of a lack of predictive power of the assessment of vulnerability, strength, key and critical items can be explained in several ways. First, although these items, particularly the critical and key items, provide handholds for treatment as the START manual suggests (Webster et al., 2009), their assessment might not be sufficient for the prediction of recidivism. Or the exact reverse is true, and scores on these items are actually extremely good at predicting imminent transgressions, which, in turn, leads case managers to intervene and prevent new incidents. This is specifically relevant for the dynamic factors addressed by the START. Being attentive to both signs of increasing risks and to opportunities for strengthening the

Table 6
Hierarchical Logistic Regression for Violent and Criminal Outcome as Defined in the START Manual in the First 3 Months Following Risk Assessment With the START (n = 252)

Step	Variable	*OR*	95% CI	*p*	R^2	AUC	95% CI	*p*
1	Sum: Historical	1.07	[0.96, 1.19]	.21				
					.01	.58	[.47, .69]	.18
2a	Sum: Historical	1.07	[0.96, 1.19]	.24				
	Sum: Vulnerabilities	1.01	[0.94, 1.07]	.89				
					.01	.58	[.47, .69]	.17
2b	Sum: Historical	1.07	[0.96, 1.19]	.21				
	M: Critical vulnerabilities	0.91	[0.40, 2.06]	.82				
					.01	.58	[.48, .69]	.15
3a	Sum: Historical	1.07	[0.96, 1.20]	.23				
	Sum: Vulnerabilities	1.01	[0.93, 1.11]	.75				
	Sum: Strengths	1.01	[0.93, 1.10]	.77				
					.01	.58	[.47, .69]	.19
3b	Sum: Historical	1.06	[0.95, 1.18]	.30				
	M: Critical vulnerabilities	0.86	[0.38, 1.98]	.73				
	M: Key strengths	0.64	[0.25, 1.64]	.35				
					.02	.60	[.50, .70]	.09
4a	Sum; Historical	1.07	[0.96, 1.19]	.24				
	Sum: Vulnerabilities	1.01	[0.92, 1.10]	.90				
	Sum: Strengths	1.01	[0.93, 1.10]	.77				
	Risk estimate for violence	1.25	[0.59, 2.66]	.57				
					.02	.58	[.47, .69]	.17
4b	Sum: Historical	1.06	[0.95, 1.18]	.33				
	M: Critical vulnerabilities	0.84	[0.36, 2.00]	.70				
	M: Key strengths	0.67	[0.26, 1.75]	.41				
	Risk estimate for violence	1.19	[0.57, 2.47]	.64				
					.02	.60	[.50, .70	].09

Note. START = Short-Term Assessment of Risk and Treatability; OR = odds ratio; CI = confidence interval; R^2 = Nagelkerke R^2; AUC = area under curve.

client's resilience are essential parts of a case managers' work. For research on the prediction of future violence, however, this is a serious complication. Successful intervention, for example, through treatment or increased supervision, will reduce the likelihood of the predicted event to occur. The result is successful clinical prediction and prevention, which, however, cannot be shown through statistical prediction models.

Our AUC findings, particularly for the 6-month follow-up, are mostly in line with previous research by Braithwaite et al. (2010), who also found modest values. Studies by Nonstad et al. (2010), Chu et al. (2011, Chu, Thomas, et al. (2013), and Gray, Benson, et al. (2011) have reported higher AUCs (range = .68–.83). The *OR*s found in the current study for the scores on vulnerability and strength, strength scales, and risk estimate for violence against others are similar to earlier reports by Braithwaite et al. (2010); Chu et al. (2011); and Desmarais, Nicholls, et al. (2012). Wilson et al. (2010) are the only ones to report significant *OR*s for the prediction of aggressive behavior with either the vulnerability or strength total scales. However, their results could be due to a Type I error because their analyses are based on 30 patients rated four times each, for which it seems they did not control in their analyses. Furthermore, when combined in the same model, scores on neither scale contributed independently to their prediction model, making multicollinearity a likely problem, though they did not report on the issue. We explicitly tested for multicollinearity and found no violations of established limits, even though the sum scores of strengths and vulnerability items were highly correlated (–.71) in our sample.

A further explanation for the differences between our findings and those of previous studies lies in our study design in which clinicians completed the START as part of daily practice relying first and foremost on their own contacts with the client. This is in contrast to most other studies where researchers or research assistants were trained to achieve a certain level of agreement and then used case files as a basis for their assessments. As de Vogel and de Ruiter (2004) have argued, there are fundamental differences between researchers and clinicians as assessors of a client's violence risk. Clinicians will try to maintain a positive therapeutic relationship and have a personal stake in the treatment of the client or could be sensitive to pressures about client placement. Therefore, clinicians might focus more on signs of reduced risk than researchers do. Additionally, researchers may not be aware of signs of reduced risk because these regularly go unreported in the case files researchers necessarily rely on for their assessments. It remains undecided as to whether better predictions of future violence in outpatient forensic psychiatry can be achieved through ratings by more impartial researchers or by more engaged clinicians.

An intriguing difference between our findings and previous research is that the pilot of the RACE study (van den Brink, Hooijschuur, van Os, Savenije, & Wiersma, 2010), conducted in the same outpatient forensic psychiatric setting, did find an incremental predictive value for scores on dynamic risk factors, which the current, larger study was unable to replicate. It should be noted, however, that dynamic risk factors were not assessed with the START but with various other instruments (for details, see van den Brink et al., 2010). Regardless of this issue, the different findings

Table 7
Hierarchical Logistic Regression for Violent and Criminal Outcome as Defined in the START Manual in the First 6 Months Following Risk Assessment With the START (n = 211)

Step	Variable	*OR*	95% CI	*p*	R^2	AUC	95% CI	*p*
1	Sum: Historical	1.08	[0.98, 1.19]	.13	.02	.58	[.48, .69]	.11
2a	Sum: Historical	1.07	[0.97, 1.19]	.17				
	Sum: Vulnerabilities	1.01	[0.95, 1.07]	.76	.02	.59	[.48, .69]	.11
2b	Sum: Historical	1.08	[0.98, 1.19]	.12				
	M: Critical vulnerabilities	0.73	[0.35, 1.53]	.41	.02	.60	[.51, .70]	.06
3a	Sum: Historical	1.07	[0.97, 1.19]	.18				
	Sum: Vulnerabilities	1.01	[0.94, 1.09]	.81				
	Sum: Strengths	1.00	[0.93, 1.07]	.99	.02	.59	[.48, .69]	.11
3b	Sum: Historical	1.08	[0.98, 1.19]	.14				
	M: Critical vulnerabilities	0.71	[0.34, 1.49]	.36				
	M: Key strengths	0.93	[0.37, 2.33]	.88	.02	.60	[.51, .70]	.05
4a	Sum: Historical	1.07	[0.97, 1.19]	.18				
	Sum: Vulnerabilities	0.99	[0.91, 1.07]	.77				
	Sum: Strengths	1.00	[0.93, 1.08]	.96				
	Risk estimate for violence	1.81	[0.91, 3.60]	.09	.04	.62	[.52, .72]	.02
4b	Sum: Historical	1.07	[0.97, 1.19]	.17				
	M: Critical vulnerabilities	0.59	[0.27, 1.30]	.19				
	M: Key strengths	1.08	[0.42, 2.75]	.88				
	Risk estimate for violence	1.93	[0.98, 3.79]	.06	.05	.65	[.55, .74]	.01

Note. START = Short-Term Assessment of Risk and Treatability; OR = odds ratio; CI = confidence interval; R^2 = Nagelkerke R^2; AUC = area under curve.

from both studies could be explained by differences in study population because the current study included all outpatient clients, but the pilot was a subsample consisting only of clients receiving forensic psychiatric home care. Individual case manager's contact with clients receiving home care differs from that with clients visiting the outpatient clinic in nature (at home vs. outpatient clinic) and duration (for several years vs. months). Thus, case managers providing home care arguably resemble case managers in inpatient settings more than their direct colleagues in outpatient settings in that they have a more comprehensive view of their client's life, making it easier for them to assess factors addressed in risk assessment. Thus if this subsample of case managers has a more informed view than their direct colleagues, it might explain why the pilot study, in which they were the only participating case managers, did find incremental predictive validity for violent outcome with scores on dynamic factors and this study did not. Repetition of our analyses with those in our sample who received psychiatric home care seems preferable. However, the sample was so small ($n = 59$), prohibiting useful conclusions. Therefore the possibility remains that adequate violence risk assessment requires a more holistic view of clients than case managers in outpatient forensic psychiatry can acquire within the limits of the treatment they provide.

Overall, our findings contribute to the emerging knowledge about the ability of START assessments to predict future violent and criminal behavior in daily clinical practice. The current article is the first to examine this in an outpatient forensic psychiatric sample rather than in the hospital setting. Additionally, risk assessments for the RACE study were carried out as part of daily practice by case managers, with various professional backgrounds who had several other claims on their time, and might therefore be a more realistic estimate of their usefulness in patient care and management than studies reporting on research assistant, file based assessments (e.g., Chu et al., 2011; Desmarais, Nicholls, et al., 2012; Wilson et al., 2010). This naturalistic approach increases the ecological validity and generalizability of our findings. However, it could also explain the lower ICC values (.49–.64, all $p < .001$) for the interrater reliability of the customary START summary scores compared with Desmarais, Nicholls, et al. (2012; .85–.95, all $p < .001$), Wilson et al. (2010; .81–.90, all $p < .001$) and Nicholls et al. (2006; .87, $p < .001$). For instance, it is common in our outpatient setting that clients will be mostly in contact with their own case manager and only incidentally with other case managers. Consequently, we had some trouble finding clients who were well-known enough to more than one case manager so as to make the completion of the START by both case managers feasible. This is in stark contrast to previous studies that employed research assistants trained for agreement and dedicated to the task of completing the START using the same case file information (discussed more fully in the following paragraphs). Given these differences, it is therefore maybe not that surprising that we find lower levels of interrater reliability. In fact, finding lower levels of interrater reliability for instruments scored in clinical rather than research settings is not new. For instance, Bjørkly, Hartvig, Heggen, Brauer, and Moger (2009) found an ICC of .87 for a test sample of clinicians using the Violence Risk Screening–10 to assess case vignettes. However, when clinicians completed assessments in clinical practice the ICC dropped to .62, which is comparable to our own findings. Similarly, Almvik, Woods, and Rasmussen (2000) have reported values ranging from .44 to 1.00 for the interrater reliability of the Brøset Violence Checklist when completed in clinical practice. Last, Miller, Kimonis, Otto, Kline, and Wasserman (2012) showed that the field reliability of the Static-99, Psychopaty Checklist–Revised, and the Minnesota Sex Offender Screening Tool–Revised was lower than reported in the various manuals. Moreover, in general researchers tend to provide higher risk estimates than do clinicians (de Vogel & de Ruiter, 2004).

As already mentioned, a further explanation for the differences in interrater findings can be found in the training of those scoring the START. Clinicians in the RACE study received the official training by the Dutch translators ('t Lam et al., 2009). This training included the assessment of several case vignettes. Booster sessions were offered to clinicians several months later, but no specific monitoring or promotion of agreement was undertaken. In contrast, the other studies trained the research assistants, who scored the START from case files, to "demonstrate adequate agreement with the trainers" or to "meet the interrater criterion (ICC≥.80)" (Desmarais, Nicholls, et al., 2012; Wilson et al., 2010). Such an extensive training, however, would be unfeasible in clinical practice because of constraints on clinician's time and resources. Therefore, it seems likely that our findings are an accurate reflection of the level of reliability that can be achieved with the START assessments completed by clinicians in clinical practice. Those desiring higher levels of reliability might consider more extensive training of clinicians (to reach a certain level of agreement) or employing research assistants specifically dedicated to the completion of risk assessments.

None of these other studies report on the interrater reliability for scores on key strength or critical vulnerability factors. In our study, ICC for them was found to be .30 and .32, respectively, which would be considered *fair* agreement. However, along with Philipse, Koeter, van der Staak, and van den Brink (2005), we would argue that this reflects clinical practice where case managers with various backgrounds and levels of training focus on different aspects of a client's risk. That is, a psychiatrist is more likely to focus on mental state and medication adherence, whereas a creative therapist might prefer to focus on impulse control and external triggers. In inpatient settings, these various preferences can be addressed through consensus sessions in which the team as a whole scores the START for the client. In outpatient forensic psychiatry, however, most clients are only seen by one clinician, making such consensus sessions unfeasible. So although we asked case managers to identify clients also known by one of their colleagues, their various treatment areas of expertise might have resulted in differences in critical and key items chosen and, hence, lower levels of agreement in their scoring of these items.

Of further note is the relatively low base rate for incidents. Most studies addressing the predictive validity of START assessments report base rates of over 50%. Specifically, defining outcome as "any violence," Nicholls et al. (2006) reported a base rate of 65% and Desmarais, Nicholls, et al. (2012) reported one of 54% for a 1-year follow-up. Using the same definition of the outcome, Braithwaite et al. (2010) reported a base rate of 87% over a period of 30 days. Wilson et al. (2010) explicitly matched their sample of participants to achieve a 50% base rate of "any violence" over a 1-year follow-up. Considerably lower, but using a stricter definition of the outcome ("severe violence"), Nonstad et al. (2010) reported a base rate of 35%. Arguably with a broader definition of the outcome, more comparable to our own, this would have been higher, even in the high security hospital where this latter study was conducted. In line with this, Gray, Benson, et al. (2011) reported that 25% of their sample was at least once "physically violent" toward someone else during a 6-month follow-up, but 52% of the sample was "verbally aggressive." Last, Chu et al. (2011 and Chu, Thomas, et al. (2013) reported base rates between 18% and 29% for "any violence" for 1- and 6-month follow-up, respectively. However, these latter data were retrospectively coded by the author from case files, which might have resulted in underreporting. In contrast, in our sample only 13% had an incident during a 3-month follow-up and 21% had an incident during a 6-month follow-up. This was so even though the outcome was broadly defined and outcome was recorded prospectively by case managers. In fact, this relatively low base rate could be the consequence of our comparatively less problematic group as reflected by their treatment in an outpatient setting and by the provision of adequate treatment. Most studies so far have reported on inpatient samples with high proportions of psychotic disorders (in general, ≥85%) and low proportions of personality disorders (≤20%). Nicholls et al. (2011) is the only other study to also report on a community sample, however still more than half their sample (54%) presented with psychotic problems, whereas only 2% had a personality disorder or traits thereof. This contrasts with the current sample in which only 7% presented with psychotic problems but the majority (72%) were diagnosed with a personality disorder. As such, these differences reflect both the variety of treatment services worldwide and a more recent explosive growth of facilities for forensic psychiatric clients. For instance, Priebe et al. (2008) reported a triplication of the number of forensic psychiatric beds and a duplication of residential care and supported housing facilities for psychiatric patients in general from 1990 to 2006 in nine European countries, including The Netherlands. Further differences in forensic psychiatric populations arise from legal and cultural differences between countries (Salize & Dressing, 2004; Salize et al., 2002). Regardless of these differences, the current study showed that the START can be used with similar results for the prediction of violence in outpatient forensic psychiatry as in the inpatient setting in which it was developed.

However, a note of caution seems appropriate regarding the use of risk-assessment instruments such as the START in clinical practice as a method of violence prediction and prevention. There is ample research linking the occurrence of recidivism with mental illness, substance misuse, client well-being and quality of life. In contrast, there is a decided lack of studies addressing the effectiveness of risk-assessment instruments, particularly dynamic risk assessment instruments, once implemented in practice (though, see Abderhalden et al., 2008; Kling, Yassi, Smailes, Lovato, & Koehoorn, 2011; Troquete et al., 2013; van de Sande et al., 2011 for notable exceptions). The evidence base is not large enough as yet for a systematic review of the subject. Therefore, it is still too early to advocate changing treatment policies to include the systematic use of structured risk assessment. However, in the same vein, it is also still too early to argue against their use. In the present study, for example, our findings for the predictive validity of the START for a 6-month period are in line with results from previous studies that used different designs. For a 3-month period we were unable to establish a successful prediction model. This might be the result of characteristics and limitations specific to our study.

Limitations and Strengths

There are limitations to our study that could have influenced our findings. The current data were collected as part of a larger cluster, randomized controlled trial (Troquete et al., 2013) that studied the preventive effect of risk assessment and shared care planning on the occurrence of new violent incidents. As noted in preceding paragraphs, risk assessment by the case manager who subsequently also provided treatment may have reduced the predictive value of the START assessments because successful treatment may have prevented new incidents from occurring. Furthermore, the case manager also made the case notes on which the assessment of violent outcomes was based. This means that outcome data did not come from an independent source. It is of course possible that case managers underreported or over reported incidents because they participated in the study. Although interesting in this respect, collection of more objective outcomes, such as re-arrests, was beyond the scope of the study. Additionally, it is common in clinical practice that the same case manager completes risk assessment, provides treatment and records incidents. It was our specific aim to test the predictive validity of the START assessments in such a setting. As such, we would consider this a strength rather than a limitation of our study. Given the nature of the intervention, the use of blinded case managers, clients, or both was not possible. This could have resulted in a bias in our findings because clinicians might underreport or over report the outcome of interest. However, as reported earlier in the article, we found no significant differences in case manager- or client-reported rates of violence between intervention and control groups of the study (Troquete et al., 2013). Therefore, it is unlikely that the current findings are negatively influenced by the implementation of treatment and outcome assessment associated with the START assessments. Moreover, it is unlikely that a lack of blinding could have influenced our findings.

An additional problem experienced during the completion of the RCT was that case managers were asked to complete various tasks pertaining to the study that they found burdensome and difficult to combine with their clinical work. Case manager motivation might therefore have influenced the attention with which they completed the START and the conscientiousness with which they recorded incidents in client's case files. Low case manager motivation might have also resulted in the 109 clients who were supposed to receive the intervention (including a START assessment) but did not. This may have influenced the generalizability of our findings because we found some significant differences between those who did receive the intervention and those who did not. Specifically, the latter had been in treatment longer before inclusion in the study and were less likely to have a personality disorder. However, our earlier study also showed that there were no significant differences in outcome between those who did and those who did not receive the intervention (either at baseline or follow-up; Troquete et al., 2013). Specifically, multilevel logistic analyses were completed for different groups of clients. Initial intention-to-treat analyses included all participants in both the control and intervention groups, regardless of the number of interventions received. We found no significant difference between the two groups. Clients in the intervention group were neither more nor less likely to have an incident at follow-up than clients in the control group ($OR = 1.46$, 95% confidence interval [CI] [0.89, 2.44], $p = .15$). We conducted additional analyses in which we took into account the frequency with which the intervention was completed by clients in the intervention group. These "as treated" (including clients who had received at least one intervention, $n = 201$, 65%) and "treatment as planned" (including clients receiving multiple interventions, $n = 72$, 23%) analyses did not change the interpretation of our results (as treated: $OR = 1.34$, 95% CI [0.76, 2.38], $p = .32$; as planned: $OR = 1.89$, 95% CI [0.89, 3.99], $p = .10$). Taken together, this makes it unlikely that the generalizability of our findings was negatively influenced by the failure to provide the intervention to part of the study group.

Finally, our data prevent us from examining changes from one risk assessment to the next as well as the effect case manager characteristics and received treatment might have had on the occurrence of violent incidents. Therefore we could not include these potential moderators in our predictive models. Previous studies concerning the predictive validity of the START have not addressed these issues either. Future studies should do so.

Our study has some fundamental strengths. It is one of the first studies to report on a community, rather than a clinical sample. Additionally, we present data on a larger sample than has been done before and, as mentioned, the START was scored as part of daily treatment practice by clinicians rather than retrospectively by research assistants using case files. This different approach, as well as the lack of monitoring to ensure good agreement between raters, could explain the relatively low results we found for the interrater reliability of the various START scores. Last, this study is the first, to our knowledge, to examine the role of ratings on critical and key items in the prediction of future violence.

Conclusion

The findings of our study are a first contribution to the examination of the predictive validity of the START when implemented in clinical practice. Although the explicit aim of START assessments is to provide a short-term prediction of violence risk, we could not establish their incremental predictive validity over the total score of the historical items of the HCR–20 during a 3-month follow-up. However, it is encouraging to note that the structured professional judgment ratings of the clinician increase the prediction of future violence above and beyond a mere actuarial summation of historical and dynamic risk and protective factor scores for a 6-month follow-up. Even though our findings are mixed, they are in line with earlier studies conducted in higher security settings with more homogenous samples and risk assessments completed by research assistants. Therefore, it seems that clinicians in a more heterogeneous outpatient forensic psychiatric setting can use the START with similar results as the clinicians and research staff in the more homogeneous inpatient setting in which the START was developed. However, there is still too little of an evidence base to advocate the implementation of risk-assessment instruments in general as a regular part of treatment policies.

References

Abderhalden, C., Needham, I., Dassen, T., Halfens, R., Haug, H.-J., & Fischer, J. E. (2008). Structured risk assessment and violence in acute psychiatric wards: Randomised controlled trial. *The British Journal of Psychiatry, 193,* 44–50. http://dx.doi.org/10.1192/bjp.bp.107.045534

Almvik, R., Woods, P., & Rasmussen, K. (2000). The Brøset Violence Checklist—Sensitivity, specificity, and interrater reliability. *Journal of Interpersonal Violence, 15,* 1284–1296. http://dx.doi.org/10.1177/088626000015012003

Arbach-Lucioni, K., Andrés-Pueyo, A., Pomarol-Clotet, E., & Gomar-Soñes, J. (2011). Predicting violence in psychiatric inpatients: A prospective study with the HCR-20 violence risk assessment scheme. *Journal of Forensic Psychiatry & Psychology, 22,* 203–222. http://dx.doi.org/10.1080/14789949.2010.530290

Bjørkly, S., Hartvig, P., Heggen, F. A., Brauer, H., & Moger, T. A. (2009). Development of a brief screen for violence risk (V-RISK-10) in acute and general psychiatry: An introduction with emphasis on findings from a naturalistic test of interrater reliability. *European Psychiatry, 24,* 388–394. http://dx.doi.org/10.1016/j.eurpsy.2009.07.004

Borum, R., Bartel, P., & Forth, A. (2003). *Manual for the Structured Assessment of Violence in Youth (SAVRY), Version 1.1*. Tampa, FL: University of South Florida.

Bouman, Y. H. A., Van Nieuwenhuizen, C., Schene, A. H., & De Ruiter, C. (2008). Quality of life of male outpatients with personality disorders or psychotic disorders: A comparison. *Criminal Behaviour and Mental Health, 18,* 279–291. http://dx.doi.org/10.1002/cbm.703

Braithwaite, E., Charette, Y., Crocker, A. G., & Reyes, A. (2010). The predictive validity of clinical ratings of the Short-Term Assessment of Risk and Treatability (START). *The International Journal of Forensic Mental Health, 9,* 271–281. http://dx.doi.org/10.1080/14999013.2010.534378

Chu, C. M., Daffern, M., & Ogloff, J. R. (2013). Predicting aggression in acute inpatient psychiatric setting using BVC, DASA, and HCR-20 Clinical scale. *Journal of Forensic Psychiatry & Psychology, 24,* 269–285. http://dx.doi.org/10.1080/14789949.2013.773456

Chu, C. M., Thomas, S. D. M., Ogloff, J. R., & Daffern, M. (2011). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) in a secure forensic hospital: Risk factors and strengths. *The International Journal of Forensic Mental Health, 10,* 337–345. http://dx.doi.org/10.1080/14999013.2011.629715

Chu, C. M., Thomas, S. D. M., Ogloff, J. R., & Daffern, M. (2013). The short- to medium-term predictive accuracy of static and dynamic risk assessment measures in a secure forensic hospital. *Assessment, 20,* 230–241.

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24,* 685–700. http://dx.doi.org/10.1037/a0026668

Desmarais, S. L., van Dorn, R. A., Telford, R. P., Petrila, J., & Coffey, T. (2012). Characteristics of START assessments completed in mental health jail diversion programs. *Behavioral Sciences & the Law, 30,* 448–469. http://dx.doi.org/10.1002/bsl.2022

de Vogel, V., & de Ruiter, C. (2004). Differences between clinicians and researchers in assessing risk of violence in forensic psychiatric patients. *Journal of Forensic Psychiatry & Psychology, 15,* 145–164. http://dx.doi.org/10.1080/14788940410001655916

de Vogel, V., & de Ruiter, C. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. *Clinical Psychology & Psychotherapy, 12,* 226–240. http://dx.doi.org/10.1002/cpp.452

Douglas, K. S., Blanchard, A. J. E., Guy, L. S., Reeves, K. A., & Weir, J. (2010). HCR-20 violence risk assessment scheme: Overview and annotated bibliography. Retrieved from http://kdouglas.wordpress.com/

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment—Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11,* 347–383. http://dx.doi.org/10.1037/1076-8971.11.3.347

Doyle, M., Carter, S., Shaw, J., & Dolan, M. (2012). Predicting community violence from patients discharged from acute mental health units in England. *Social Psychiatry and Psychiatric Epidemiology, 47,* 627–637. http://dx.doi.org/10.1007/s00127-011-0366-8

Doyle, M., & Dolan, M. (2002). Violence risk assessment: Combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing, 9,* 649–657. http://dx.doi.org/10.1046/j.1365-2850.2002.00535.x

Gray, N. S., Benson, R., Craig, R., Davies, H., Fitzgerald, S., Huckle, P., . . . Snowden, R. J. (2011). The Short-Term Assessment of Risk and Treatability (START): A Prospective Study of Inpatient Behavior. *The International Journal of Forensic Mental Health, 10,* 305–313. http://dx.doi.org/10.1080/14999013.2011.631692

Gray, N. S., Taylor, J., & Snowden, R. J. (2011). Predicting violence using structured professional judgment in patients with different mental and behavioral disorders. *Psychiatry Research, 187,* 248–253. http://dx.doi.org/10.1016/j.psychres.2010.10.011

Hastings, M. E., Krishnan, S., Tangney, J. P., & Stuewig, J. (2011). Predictive and incremental validity of the Violence Risk Appraisal Guide scores with male and female jail inmates. *Psychological Assessment, 23,* 174–183. http://dx.doi.org/10.1037/a0021290

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley. http://dx.doi.org/10.1002/0471722146

Kling, R. N., Yassi, A., Smailes, E., Lovato, C. Y., & Koehoorn, M. (2011). Evaluation of a violence risk assessment system (the Alert System) for reducing violence in an acute hospital: A before and after study. *International Journal of Nursing Studies, 48,* 534–539. http://dx.doi.org/10.1016/j.ijnurstu.2010.10.006

Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law and Human Behavior, 24,* 101–118. http://dx.doi.org/10.1023/A:1005430904495

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. http://dx.doi.org/10.2307/2529310

Lewis, G., & Doyle, M. (2009). Risk Formulation: What are We Doing and Why? *The International Journal of Forensic Mental Health, 8,* 286–292. http://dx.doi.org/10.1080/14999011003635696

McGowan, M. R., Horn, R. A., & Mellott, R. N. (2011). The predictive validity of the Structured Assessment of Violence Risk in Youth in secondary educational settings. *Psychological Assessment, 23,* 478–486. http://dx.doi.org/10.1037/a0022304

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment, 24,* 944–953. http://dx.doi.org/10.1037/a0028411

Mudde, N., Nijman, H., van der Hulst, W., & van den Bout, J. (2011). Het voorspellen van agressie tijdens de behandeling van forensisch psychiatrische patiënten aan de hand van de HCR-20. [Predicting aggression during the treatment of forensic psychiatric patients by means of the HCR-20]. *Tijdschrift voor Psychiatrie, 53,* 705–713.

Nicholls, T. L., Brink, J., Desmarais, S. L., Webster, C. D., & Martin, M. L. (2006). The Short-Term Assessment of Risk and Treatability (START): A prospective validation study in a forensic psychiatric sample. *Assessment, 13,* 313–327. http://dx.doi.org/10.1177/1073191106290559

Nicholls, T. L., Petersen, K. L., Brink, J., & Webster, C. D. (2011). A Clinical and Risk Profile of Forensic Psychiatric Patients: Treatment Team STARTs in a Canadian Service. *The International Journal of Forensic Mental Health, 10,* 187–199. http://dx.doi.org/10.1080/14999013.2011.600234

Nonstad, K., Nesset, M. B., Kroppan, E., Pedersen, T. W., Nøttestad, J. A., Almvik, R., & Palmstierna, T. (2010). Predictive Validity and Other Psychometric Properties of the Short-Term Assessment of Risk and Treatability (START) in a Norwegian High Secure Hospital. *The International Journal of Forensic Mental Health, 9,* 294–299. http://dx.doi.org/10.1080/14999013.2010.534958

Ogloff, J. R., & Daffern, M. (2006). The dynamic appraisal of situational aggression: An instrument to assess risk for imminent aggression in psychiatric inpatients. *Behavioral Sciences & the Law, 24,* 799–813. http://dx.doi.org/10.1002/bsl.741

O'Shea, L. E., Mitchell, A. E., Picchioni, M. M., & Dickens, G. L. (2013). Moderators of the predictive efficacy of the Historical, Clinical and Risk Management-20 for aggression in psychiatric facilities: Systematic review and meta-analysis. *Aggression and Violent Behavior, 18,* 255–270. http://dx.doi.org/10.1016/j.avb.2012.11.016

Philipse, M., Koeter, M., van der Staak, C., & van den Brink, W. (2005). Reliability and discriminant validity of dynamic reoffending risk indicators in forensic clinical practice. *Criminal Justice and Behavior, 32,* 643–664. http://dx.doi.org/10.1177/0093854805279946

Priebe, S., Frottier, P., Gaddini, A., Kilian, R., Lauber, C., Martínez-Leal, R., . . . Wright, D. (2008). Mental health care institutions in nine European countries, 2002 to 2006. *Psychiatric Services, 59,* 570–573. http://dx.doi.org/10.1176/appi.ps.59.5.570

Salize, H. J., & Dressing, H. (2004). Epidemiology of involuntary placement of mentally ill people across the European Union. *The British Journal of Psychiatry, 184,* 163–168. http://dx.doi.org/10.1192/bjp.184.2.163

Salize, H. J., Dressing, H., & Peitz, M. (2002). *Compulsory admission and involuntary treatment of mentally ill patients—Legislation and practice in EU-member states*. Mannheim, Germany: Central Institute of Mental Health.

SPSS. (2012). PASW Statistics (Version 20.0.01). Chicago, IL: SPSS Inc.

Telles, L. E., Folino, J. O., & Taborda, J. G. (2012). Accuracy of the Historical, Clinical and Risk Management Scales (HCR-20) in predicting violence and other offenses in forensic psychiatric patients in Brazil. *International Journal of Law and Psychiatry, 35,* 427–431. http://dx.doi.org/10.1016/j.ijlp.2012.09.001

't Lam, K., Lancel, M., & Hildebrand, M. (2009). *Handleiding bij de Short-Term Assessment of Risk and Treatability (START): Richtlijnen bij het beoordelen van korte termijn risico's en behandelmogelijkheden* [Manual for the Short-Term Assessment of Risk and Treatability (START): Guidelines for assessment of short-term risks and treatment opportunities (Dutch translation)]. Assen, The Netherlands : GGZ Drenthe.

Troquete, N. A. C., van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient forensic psychiatry: Cluster randomised controlled trial. *The British Journal of Psychiatry, 202,* 365–371. http://dx.doi.org/10.1192/bjp.bp.112.113043

Tully, R. J., Chou, S., & Browne, K. D. (2013). A systematic review on the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders. *Clinical Psychology Review, 33,* 287–316. http://dx.doi.org/10.1016/j.cpr.2012.12.002

van den Brink, R. H. S., Hooijschuur, A., van Os, T. W. D. P., Savenije, W., & Wiersma, D. (2010). Routine violence risk assessment in community forensic mental healthcare. *Behavioral Sciences & the Law, 28,* 396–410.

van de Sande, R., Nijman, H. L. I., Noorthoorn, E. O., Wierdsma, A. I., Hellendoorn, E., van der Staak, C., & Mulder, C. L. (2011). Aggression and seclusion on acute psychiatric wards: Effect of short-term risk assessment. *The British Journal of Psychiatry, 199,* 473–478. http://dx.doi.org/10.1192/bjp.bp.111.095141

Ward, T., & Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime & Law, 10,* 243–257. http://dx.doi.org/10.1080/10683160410001662744

Ward, T., Mann, R. E., & Gannon, T. A. (2007). The good lives model of offender rehabilitation: Clinical implications. *Aggression and Violent Behavior, 12,* 87–107. http://dx.doi.org/10.1016/j.avb.2006.03.004

Ward, T., & Stewart, C. A. (2003). The treatment of sex offenders: Risk management and good lives. *Professional Psychology, Research, and Practice, 34,* 353–360. http://dx.doi.org/10.1037/0735-7028.34.4.353

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. (1997). *HCR-20. Assessing the Risk of Violence. Version 2*. Burnaby, BC, Canada: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Desmarais, S. L. (2009). *Manual for the Short-Term Assessment of Risk and Treatability (START) (Version 1.1)*. Port Coquitlam, BC: Forensic Psychiatric Services Commission and St. Joseph's Healthcare.

Webster, C. D., Nicholls, T. L., Martin, M. L., Desmarais, S. L., & Brink, J. (2006). Short-Term Assessment of Risk and Treatability (START): The case for a new structured professional judgment scheme. *Behavioral Sciences & the Law, 24,* 747–766. http://dx.doi.org/10.1002/bsl.737

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., & Brink, J. (2010). The Role of Client Strengths in Assessments of Violence Risk Using the Short-Term Assessment of Risk and Treatability (START). *The International Journal of Forensic Mental Health, 9,* 282–293. http://dx.doi.org/10.1080/14999013.2010.534694

Wubs, H., & Wijnen, G. H. (2005). *Ambulante forensisch-psychiatrische behandelingen op de AFPN* [Out-patient forensic psychiatric treatment in the out-patient forensic psychiatric clinics in the northern Netherlands]. Vol. 11). Groningen, The Netherlands : University of Groningen.

Received February 8, 2013
Revision received August 4, 2014
Accepted August 23, 2014 ■