Predictive Validity of the Short-Term Assessment of Risk and Treatability (START) for Aggression and Self-Harm in a Secure Mental Health Service: Gender Differences

O'Shea, Laura E. & Dickens, Geoffrey L.

This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Forensic Mental Health 14(2) on 1st June 2015, available online: http://www.tandfonline.com/10.1080/14999013.2015.1033112

Abstract

The START predicts aggressive outcomes, and to some extent self-harm. However, it is not known whether gender moderates its performance. This study used routinely collected data to investigate the predictive ability of the START for aggression and self-harm in secure psychiatric patients. Utility of the START was examined separately for men and women. The START was a stronger predictor of aggression and self-harm in women than men. The specific risk estimates produced large effect sizes for the prediction of aggression and self-harm in women; none of the AUC values reached the threshold for a large effect size in the male sample.

Introduction

Aggression and self-harm are serious problems in forensic mental health settings. Across studies, an average of 48% of inpatients have been aggressive (Bowers et al., 2011) while 42.9% have self-harmed (James, Stewart, & Bowers, 2012). Aggression and self-harm both have obvious and serious consequences for patients and staff. Bowers et al. (2011) found that 2-13% of aggressive incidents resulted in serious injury and 5-28% in moderate to severe injury; James et al. (2012) found that 12-20% of self-harm incidents were classified as severe, resulting in deep cuts, fractures, or internal injuries. Further, witnessing such events in secure psychiatric settings is correlated with illness-related work absence (Nijman, Bowers, Oud, & Jansen, 2005). The prevention and management of aggression and self-harm is therefore a key objective for mental health professionals. Risk assessment tools have been widely adopted by clinicians as a structured method of guiding their formulation and prediction of patients' risk; decisions about management interventions and strategies are commonly informed by these assessments.

Given that risk assessments are frequently used to inform decisions about restrictive management interventions it is crucial to determine their effectiveness in all the groups to which they are applied. This is an important consideration as clinicians have to determine the relevance of evidence derived from validation studies to the individual case at hand when making risk judgments (Buchanan, 2013). However, samples in studies of the predictive validity of risk assessment tools have been primarily male; 91% of samples included in a meta-analysis of risk assessment tools contained over 50% male participants (Singh, Grann, & Fazel, 2011). This has limited the detailed examination of whether, and the extent to which, their performance significantly differs as a function of gender.

Different or additional risk factors may underlie risk of aggression in women. For example, de Vogel and de Ruiter (2005) found that the risk factors most frequently identified

by clinicians in addition to those in the HCR-20 scheme differed between men (financial problems, lack of prospects for the future, and violent fantasies) and women (forming a new intimate relationship, care for children, and prostitution). Further, Yang, Wong, and Coid (2013) identified factors that increase odds of engaging in aggression in women that are not covered by existing risk assessment schemes, such as experience of domestic violence, traumatisation as a result of separation/divorce, and presence of self-harm/suicide attempts. Similarly, previous research has identified gender differences in reasons for engaging in self-harm (Claes, Vandereycken, & Vertommen, 2007); women more commonly endorsed statements describing self-harm as serving an avoidance or punishment function, while men endorsed statements about self-harm as an attention-getter or as a show of personal strength.

The underlying differences in the factors predicting risk behaviours suggests that formal risk assessment schemes might perform differently as a function of gender. From a theoretical gendered perspective, it may be expected that this difference would manifest in the form of superior performance of risk assessment tools among males, due to the predominance of male samples in their development (Singh et al., 2011). It is therefore perhaps counterintuitive that recent research has suggested that structured professional judgment tools have at least equal efficacy in women compared with men. For example, while cautioning against widespread generalisation due to small sample sizes, Singh et al. (2011) found higher diagnostic odds ratios, suggesting better performance, in female samples compared with male samples in a meta-regression analysis of nine risk assessment tools. Further, and perhaps most pertinent to the START which is largely used in secure psychiatric settings, the HCR-20 has been found to predict inpatient aggression and self-harm more accurately in females (O'Shea, Mitchell, Picchioni, & Dickens, 2013; O'Shea, Picchioni, Mason, Sugarman, & Dickens, 2014a; O'Shea, Picchioni, Mason, Sugarman, & Dickens, 2014b). The relative importance of individual items in predicting outcome also differs as a

function of gender; factors pertaining to future risk are most relevant in women and those relating to current clinical presentation are more relevant in men (O'Shea, Picchioni, Mason, Sugarman, & Dickens, 2014a; O'Shea, Picchioni, Mason, Sugarman, & Dickens, 2014b).

The START

The Short-Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Desmarais, 2009) is a commonly used structured professional judgment tool which was developed in the context of two frequent criticisms of such schemes. First, that previous risk assessment tools have focused exclusively on factors associated with increased risk while ignoring protective factors (Hart, 2001). Second, that risk schemes focus on aggression and violence despite the range of clinical issues facing psychiatric patients (Webster et al., 2009). Further, the START consists of entirely dynamic variables, in contrast to previous risk assessment tools which have been composed of primarily static variables (e.g., the VRAG; Harris, Rice, & Quinsey, 1993), or a combination of dynamic and static variables (e.g., the HCR-20; Douglas, Hart, Webster, & Belfrage, 2013). However, the START authors (Webster, Nicholls, Martin, Desmarais, & Brink, 2006) state that historical variables should be considered in any risk assessment. They further state that it is essential to complete the H10 scale of the HCR-20 if the risk is related and restricted to others; additional items should be considered if other risks are implicated. The START aims to assist assessment of risk of a range of outcomes occurring in patients with mental and personality disorders, while considering both risk and protective factors, termed Strengths and Vulnerabilities. The START has received considerable attention due to its relatively unique features. Recent research (e.g., O'Shea, Picchioni, & Dickens, 2014) has established that the START can predict aggressive outcomes, and that the corresponding risk estimates, but not scale totals, can predict self-harm (see O'Shea & Dickens, 2014 for a review and synthesis of

the existing literature). Given that previous research has shown that the performance of similar risk assessment tools, such as the HCR-20, differs based on gender (e.g., O'Shea et al., 2013), it is reasonable to assume that the START may perform similarly. However, gender differences in START performance is yet to be tested.

Contribution of the Current Study

The current study aims to establish whether the predictive efficacy of the START for inpatient aggression and self-harm differs as a function of gender, whilst controlling for significant covariate characteristics. We also aimed to examine the relative importance of the individual items for each of the groups, as this has important implications for the development of risk management strategies. We hypothesised that the START would perform best in women, due to increasing evidence that risk assessment tools perform more accurately among this group in inpatient settings (e.g., O'Shea et al., 2013; O'Shea, Picchioni, et al., 2014b).

Method

Participants

St. Andrew's provides secure inpatient mental health care at four sites in England, for patients admitted under civil and forensic sections of the Mental Health Act. Accommodation is provided in gender-specific medium and low secure wards, with a small number of rehabilitation beds in unlocked units. All patients in the current study were previously reported on by O'Shea, Picchioni, and Dickens (2014). Eligible patients were consecutive admissions between May 2011 and July 2012 who had a START risk assessment completed and remained for at least three months. Patients were excluded if their START assessment had more than five missing Strength items or five missing Vulnerability items in accordance with guidelines in the START manual (Webster et al., 2009).

Procedure

The study design was pseudo-prospective; START assessments were completed by multidisciplinary teams as part of routine clinical practice during admission and risk incidents were recorded in electronic progress notes by qualified clinical staff on a per shift basis. Patients' demographic and clinical characteristics were extracted in a pseudonymised form and linked by a unique identification number with their first START assessment and risk incidents for the subsequent three months. The study was approved as a service evaluation by the Head of Clinical Effectiveness in the study organisation.

Measures

START assessment. The START was designed to be completed by a "number of mental health specialists who work together as a team" (Webster et al., 2009; p. 24). It comprises 20 items, scored twice on a 3-point scale (0 = no/minimal strength/vulnerability, 1 = moderate strength/vulnerability, 2 = high strength/vulnerability); once in terms of risk factors (termed Vulnerabilities) and once regarding protective factors (Strengths). Raters are advised to indicate key and critical items to identify strengths and vulnerabilities that seem especially important for the case at hand. Raters also record whether the patient has a history of behaviours relating to each of the seven risk areas the START aims to address: violence to others, self-harm, suicide, substance abuse, victimisation, self-neglect, and unauthorised absence. Specific risk estimates (SREs; low, moderate, or high risk) regarding the likelihood of each of these outcomes occurring over the subsequent three months are then made by the rating team. For research purposes, the total Strength and Vulnerability scales can be summed and prorated to account for missing items following guidelines in the START manual (Webster et al., 2009).

In the current study setting, all raters were provided with structured theoretical and practical training in START completion. Training involved team discussion and rating of

pseudonymised cases. This was followed by feedback regarding ratings given by teams during previous training sessions and by START experts. Completed START assessments are signed off by three members of the multidisciplinary team from different professions. These assessments are completed every three months for each patient and are routinely audited to ensure compliance. It was not possible to calculate inter-rater reliability as the START assessments were completed, as per the START manual's recommendations, for clinical purposes by the patients' multidisciplinary team.

Demographic and clinical data. Patients' age, gender, self-reported ethnicity, admission/discharge date, security level, legal status, and ICD-10 (World Health Organisation, 1992) psychiatric diagnoses were extracted from clinical records.

Risk outcomes. For each patient, an electronic progress note was entered on every nursing shift by a qualified member of clinical staff. Notes were flagged if any of a range of risk outcomes occurred. As part of a previous study (O'Shea, Picchioni, & Dickens, 2014), incidents flagged as containing the following outcomes were collated: "Aggression – Physical", "Aggression – Verbal", "Absconding", "Self-harm/Suicide", "Self-neglect" "Substance Misuse", and "Vulnerability". Collated notes were then coded by both authors, who were blind to the START assessment at the point of coding, using the START Outcome Scale (SOS; Nicholls et al., 2007). The SOS was adapted from the Overt Aggression Scale (OAS; Yudofsky, Silver, Jackson, Endicott, & Williams, 1986) and contains 12 outcome categories rated on a scale of 0 (outcome absent) to 4 (most severe outcome). Raters were required to judge whether each note met the criteria for a level 1 incident or above. Inter-rater reliability was in the excellent range; Kappa ranged from .83 to 1.00, the lowest being for self-neglect and the highest for self-harm and physical aggression. For the purpose of the current study, we were only interested in the categories of verbal aggression, aggression against property, physical aggression against others, self-harm, suicide ideation and planning,

and suicide behaviours. We treated aggression against property and physical aggression against others as a single outcome (physical aggression), to minimise the number of reported outcomes and due to an overlap between the more serious forms of property aggression, such as throwing objects dangerously, and physical aggression against others. Self-harm, suicide ideation and planning, and suicide behaviours were combined into "self-harm/suicidal behaviour", due to difficulties in separating non-suicidal self-harm and actual suicide attempts (Gray, Taylor, & Snowden, 2011). We further amalgamated physical aggression (including property aggression and aggression against others) and verbal aggression to form an "any aggression" category. Therefore, the final outcomes categories for the purpose of the current study were verbal aggression, physical aggression, any aggression and self-harm/suicidal behaviour.

Data Analysis

Descriptive statistics were used to examine the characteristics of the two samples, distribution of START scores, SREs, key and critical items, and the occurrence of risk outcomes. Independent *t* tests and Pearson's chi squared tests were used to investigate differences in mean scores and risk level between those who had and had not engaged in each outcome within the two samples, differences in mean scores and sample characteristics, differences in the number of key and critical items, and differences in rates of engagement in aggression and self-harm/suicidal behaviour. One way-ANOVAs were used to determine if mean Strength and Vulnerability scores differed between risk levels assigned by the SREs in both the male and female samples.

Positive Predictive Values (PPV) and Negative Predictive Values (NPV) of the SREs were calculated to examine true positive and true negative predictions of outcomes measured using the SOS; this can assist in indicating whether a tool is of greater value for screening out low risk individuals or in identifying higher risk individuals. To do this we assigned those

rated at elevated risk (moderate or high) as a positive test result, and those rated at low risk as a negative test result. PPVs can be interpreted as the percentage of people who are rated as test positive, in this case moderate or high risk, that actually engage in the outcome; conversely, NPVs represent the percentage of individuals rated as test negative (rated low risk) that do not engage.

The differential predictive validity of the START Strength scores, Vulnerability scores, SREs, and all individual Strength and Vulnerability items as a function of gender was examined using the rocreg function in Stata version 12 for Windows. The total Strength score was inverted prior to ROC analysis such that a higher score represented less strength to facilitate comparisons with the predictive efficacy of the Vulnerability scores and SREs. Rocreg performs a regression using Receiver Operating Characteristic (ROC) principles and therefore calculates sensitivity and specificity based on variables of interest, in this case gender, whilst controlling for covariates. The Area Under the Curve (AUC) value obtained from ROC analysis ranges from 0 to 1, with .5 representing performance equivalent to chance. Typically, .75 is considered the threshold for a large effect size (Dolan & Doyle, 2000); however, there is some variation in the literature (Singh, Desmarais, & Van Dorn, 2013). Rice and Harris (2005) report that AUC values of .556, .638 and .714 respectively are equivalent to small (.2), moderate (.5) and large (.8) Cohen's *d* values (Cohen, 1992), which are one of the most commonly reported measures of effect size (Kraemer & Kupfer, 2006; Rice & Harris, 2005). The AUC value can be interpreted as the probability that an individual who has engaged in the outcome in question will have a higher score on the risk assessment than someone who has not engaged. Significance of rocreg coefficients (representing significant differences in performance between men and women) and AUC values were inferred from absence of zero and .5 respectively from 99% confidence intervals (equivalent to $p<0.01$). Odds ratios (ORs) were also calculated to present the increase in odds for each

one point increase on the Vulnerability and inverted Strength scales, and between risk levels assigned by the SREs, for each adverse outcome occurring.

Finally, block entry logistic regression was used to examine whether significantly predictive Strength and Vulnerability scores had incremental validity over one another, and whether significant SREs had incremental validity over both scores. Where both scales were significantly predictive of an outcome, Vulnerability scores were entered in step 1 of the model, followed by Strength scores in step 2; this order was then reversed. The SREs were always entered in the final step of the model as they should be formed based on consideration of the scores. Significant changes in chi-squared values indicate a significant improvement in model fit (Field, 2009); changes in the percentage of correctly classified cases were also presented. Multicollinearity was investigated using the variance inflation factor (VIF) and the related tolerance statistic, which is the reciprocal of the VIF (Field, 2013). Typically, there is a potential problem if the largest VIF exceeds 10, or any of the tolerance statistics are less than 0.2 (Field, 2013; Menard, 1995; Myers, 1990). Except where stated, analyses were conducted using PASW Statistics version 18 for Windows (SPSS Inc., Version 18).

Results

Participants

In total, 214 patients met the inclusion criteria; 14 were excluded due to missing START item ratings leaving a final sample of N=200 (response rate 93.5%). The sample contained 149 (74.5%) males and 51 (25.5%) females; differences in characteristics are presented in Table 1. Women were more likely to be Caucasian than men (n=32, 62.7% vs. n=50, 33.6%), $\chi^2(2, N=200)=13.39$, $p=.001$ and differences in psychopathology $\chi^2(8, N=200)=43.95$, $p<.001$ were due to overrepresentation of personality disorder and neurotic disorder diagnoses in women and an underrepresentation of personality disorder in men and

organic disorders in women. There were no gender differences between age at assessment, legal status, time between admission and START assessment, or security level (see Table 1). Therefore, ethnicity and diagnosis were controlled for in the rocreg and OR analyses.

Incidents

Across the sample as a whole, over two thirds engaged in any aggression (*n*=138, 69.0%), 123 (61.5%) engaged in verbal aggression and 108 (54.0%) engaged in physical aggression; just under a quarter of the sample (*n*=48, 24.0%) engaged in self-harm/suicidal behaviour. Women were significantly more likely to engage in self-harm/suicidal behaviour than men (*n*=24, 47.1% vs. *n*=24, 16.1%), $\chi^2(1, N=200)=19.96$, $p<.001$. There were no significant differences in rates of engagement in any of the aggressive outcomes between men and women. In terms of historical behaviour, men were more likely to have a recorded history of violence than women (*n*=112, 75.2% vs. *n*=21, 41.2%), $\chi^2(1, N=200)=19.71$, $p<.001$; there were no significant differences in recorded history of self-harm or suicide (see Table 1). Therefore, a flagged history of violence was controlled for in rocreg and OR analyses pertaining to the aggressive outcomes.

START scores and SRE Distribution as a Function of Gender and Outcome

Mean Strength scores for the females (*M*=14.6, *SD*=6.3) and males (*M*=16.3, *SD*=6.7) were not significantly different, $t(198)=1.55$, $p=.124$. Mean Vulnerability scores were significantly higher for women compared with men (28.7, *SD*=6.21 vs. 24.4, *SD*=6.31), $t(198)=-4.23$, $p<.001$. For men, Strength scores were significantly smaller, and Vulnerability scores significantly larger in those who had engaged in all three aggressive outcomes compared with those who had not engaged; there were no significant differences in scores between those who did and did not engage in self-harm/suicidal behaviour. Strength and Vulnerability scores only differed significantly as a function of any aggression and verbal aggression among the female sample (see Table 2).

Risk levels assigned by the SRE for violence did not differ between men and women. However, women were significantly less likely to be rated as low risk and significantly more likely to be rated as high risk for engaging in self-harm, $\chi^2(2, N=152)=21.02$, $p<.001$. The SRE for suicide differed as a function of gender, with women more likely to be rated as moderate risk, $\chi^2(2, N=133)=17.05$, $p<.001$ (see Table 3). Among the female sample, mean Strength and Vulnerability scores did not differ between those classified as low, moderate, or high risk by any of the three SREs. Scores significantly differed as a function of the SRE for violence in the male sample, such that Strength scores were higher, $F(2, 132)=12.59$, $p<.001$, and Vulnerability scores lower, $F(2, 132)=12.38$, $p<.001$ in the group classified as low risk, compared with the moderate or high risk group, which did not significantly differ from each other.

There was no difference in engagement in verbal aggression as a function of risk level assigned by the SRE for violence in the male sample. However, for all other outcomes, and for both males and females, those rated as low risk of engaging in violence, self-harm and suicide were less likely to engage in their associated outcome than those rated moderate or high risk. This difference was most pronounced between engagement in self-harm/suicidal behaviour between those rated as low risk of suicide ($n=8$, 10.5%) compared with those rated at moderate ($n=10$, 47.6%) or high risk ($n=4$, 57.1%) in the male sample, $\chi^2(2, N=104)=19.40$, $p<.001$.

Key and Critical Items

In both the male, $t(148)=-8.64$, $p<.001$, and female samples, $t(50)=-8.18$, $p<.001$, the mean number of critical items identified was higher than the mean number of key items. A significantly higher number of critical, $t(198)=1.99$, $p=.048$, and key items, $t(197)=3.46$, $p<.001$, were identified for men than for women. The most commonly identified key and critical items also differed as a function of gender, with the exception of V6 (mental state)

which was one of the most common critical items for both groups; V9 (impulse control) in the male sample and V7 (emotional state) in the female sample were the other frequently identified critical items. In terms of key items, S11 (social support) and S5 (self-care) were most common in males, whilst S3 (occupational) and S20 (treatability) were identified more frequently in females.

Predictive Validity

The PPV of the SRE for violence was 83.2% for any aggression, 74.3% for verbal aggression and 70.3% for physical aggression; NPVs were 45.7%, 50.6% and 64.2%, respectively. The PPV of the self-harm SRE for the self-harm/suicidal behaviour outcome was 55.2% and the NPV was 87.2%. The PPV for the suicide SRE exceeded that of the self-harm SRE for the same outcome (59.4%), but the NPV was not as large (81.2%). Rocreg analyses revealed that the Vulnerability scale was a significantly stronger predictor of verbal aggression and self-harm/suicidal behaviour in females, compared with males. There were no other significant differences in performance as a function of gender; however, the Strength scale was a significant predictor of physical aggression and any aggression for males, but not females. Similarly, the SRE for self-harm was a significant predictor of self-harm/suicidal behaviour in women, but not in men. With the exception of verbal aggression, the SREs exceeded the predictive ability of the scale scores in the female sample; in the male sample, the SREs only exceeded both scale scores for physical aggression. In all cases, except for physical aggression and any aggression as predicted by the Strength scale, the AUC value in the female sample exceeded that of the male sample, although this was not significant in most cases. The prediction of physical aggression, any aggression and self-harm/suicidal behaviour by their corresponding SRE in the female sample all produced AUC values that exceeded the threshold for a large effect size (Dolan & Doyle, 2000) (see Table 4).

ORs were largely consistent with the rocreg analyses. Increases in Vulnerability and inverted Strength scores resulted in small, but significant, increases in the odds of engaging in verbal aggression for both men and women, and for physical and any aggression for men only. Increases in Vulnerability and inverted Strength scores did not increase odds of engaging in self-harm/suicidal behaviour. Those rated at elevated (moderate or high) risk of engaging in violence were more likely to have engaged in all three aggressive outcomes. ORs were highest for physical aggression in the female sample; those rated as moderate and high risk were 52 and 36 times more likely to have engaged in physical aggression compared with those rated low risk. Being rated as high risk did not increase odds of engaging in aggressive outcomes compared with those rated as moderate risk with the exception of physical aggression in males, where those rated as high risk were six times more likely to engage than those rated as moderate risk. Having an elevated self-harm risk rating increased odds of engaging in self-harm/suicidal behaviour in both samples; this was non-significant for the suicide risk estimate.

Item-outcome Analysis

The most potent predictive items differed between men and women. For verbal aggression, S10 (external triggers) and S11 (social support) were the strongest predictors for women, producing large AUC values of .77 and .76, respectively; for men S9 (impulse control; AUC=.65) and S15 (rule adherence; AUC=.66) were the best predictors. For physical aggression, S14 (medication adherence; AUC=.69) was the only significant predictor among women; S9 was among the strongest predictors in the male group (AUC=.70) along with V16 (conduct; AUC=.68). S6 (mental state) was one of the strongest predictors of any aggression in men (AUC=.71) and women (AUC=.74); S9, once again, was one of the best predictors in men (AUC=.69) whilst S10 (AUC=.78) was more important for women. Finally, for self-harm/suicidal behaviour, only S9 (AUC=.67) was significantly predictive for the male

sample; V19 (coping; AUC=.75) and V15 (rule adherence; AUC=.70) were the strongest predictors of this outcome for women. None of the AUC values obtained from the male sample reached the threshold for a large effect size (see online data supplement for full results of item-outcome analyses).

Incremental Validity Analyses

Self-harm/suicidal behaviour and verbal aggression in males, and physical aggression in females were excluded from the analyses as one or less of the START components were significantly predictive of these outcome-group combinations. Examination of VIFs and tolerance statistics revealed that none of the VIFs exceeded 10 and none of the tolerance values were less than 0.2 indicating that there was not a problem with multicollinearity. For the prediction of verbal aggression in women, neither the Strength, $\Delta \chi^2(1) = 0.50$, nor Vulnerability scales,$\Delta \chi^2(1) = 0.47$, had incremental validity over one another and actually led to a reduction in the percentage of cases correctly identified (-4.0% and -2.0% for the inclusion of Strengths and Vulnerabilities, respectively). The SRE for violence had incremental validity over the Strengths scores for the prediction of physical aggression, $\Delta \chi^2(1) = 10.08$, p=.002, and any aggression in men, $\Delta \chi^2(1) = 4.71$, p=.030,, and resulted in a 5.2% and 1.4% increase in correctly classified cases respectively. The SRE for violence also had incremental validity over the Vulnerability scale for prediction of any aggression in women, $\Delta \chi^2(1) = 9.83$, p=.002, and the percentage of correctly classified cases increased by 8.6. Finally, for women, both the self-harm,$\Delta \chi^2(1) = 6.01$, p=.014, and suicide, $\Delta \chi^2(1) = 8.86$, p=.003, SREs had incremental validity for the composite self-harm/suicidal behaviour outcome and increased the percentage of cases correctly classified by 14.5 and 10.6, respectively (see Table 5).

Discussion

The current study has provided the first evidence that the START is a better predictor of aggression and of self-harm for women than it is for men. This difference was only significant in two comparisons; however, in nearly all cases AUC values and ORs were larger among the female sample. Importantly, this was the case when significant potential confounders were controlled for including diagnosis, ethnicity, and previous relevant behavioural history. For women, START Vulnerability scores predicted any aggression with a moderate-large effect size, and Strength and Vulnerability scores predicted verbal aggression with AUC values just short of the threshold for a large effect size; the SREs for violence and self-harm predicted their respective outcomes with a large effect size. Women identified as being at elevated risk of violence were 36-52 times more likely to engage in physical aggression than those rated as low risk and those rated at elevated risk of self-harm or suicide were 3-7 times more likely to engage in the corresponding outcome. For men, AUC values for Strength scale and SREs for the prediction of violence only produced moderate AUC values; additionally, the Vulnerability scale was not predictive of any outcome and the composite self-harm/suicidal behaviour outcome was not predicted by any of the START components among males. However, the ORs obtained from the violence and self-harm SREs were significant in males; those at elevated risk for violence were 3-14 times more likely than those rated low risk to engage in physical aggression and those rated at elevated risk of engaging in self-harm were 7-13 times more likely to engage in self-harm/suicidal behaviour than those rated as low risk.

It was not possible to conduct incremental validity analyses for all outcome-group combinations; however, for all examined outcomes, the SREs had incremental validity over the Strength or Vulnerability scale. For men, the addition of the SRE resulted in a 1.4% – 5.2% increase in the percentage of cases classified correctly; this ranged from 8.6% to 14.5% in women. Whilst the percentage increases were much larger for women than men, indicating

that the SREs have more unique predictive ability in women than men, they are not directly comparable as they examine different outcomes and incremental validity over different START scales. It was only possible to examine incremental validity of the Strength and Vulnerability scale over one another for verbal aggression in women; neither scale had incremental validity. However, a previous study in the same sample (O'Shea, Picchioni, & Dickens, 2014) found that the Strength scale had incremental validity over the Vulnerability scale for all examined outcomes, although increases in the percentage of cases correctly classified were very small; the Vulnerability scale did not have incremental validity over the Strength scale for any outcomes. The fact that mean Strength and Vulnerability scores for women did not differ across the risk levels assigned by the SREs, combined with evidence that the SREs have greater incremental validity over scale scores in this group compared with men suggests that clinicians are considering factors additional to scores on the START items when making SREs for women. These additional factors appear to be improving their risk estimates for this group given that the SREs showed greater discrimination between those who did and did not engage in the various outcomes than the scale scores.

These findings are important since they provide further information that should modify implications for practice made about the START in our previous study in the same sample (O'Shea, Picchioni, & Dickens, 2014). There, we found that the START was a moderately good predictor of a range of aggressive outcomes and that the SRE as a predictor of self-harm had strong predictive validity. It is now clear that this conclusion should be subtly altered, and a distinction should be made based on gender. First, the START SREs are strong predictors of aggressive and self-harm outcomes for women; second, that the START is a moderate predictor of aggressive outcomes in men, but the START scores do not predict self-harm/suicidal behaviour. As a result, practitioners may have a degree of added confidence in their START assessment rating if the subject in question is female. This is

consistent with research with the HCR-20 which found superior predictive efficacy and a greater number of relevant items for aggression and self-harm in women compared with men (O'Shea, Picchioni, et al., 2014a; O'Shea, Picchioni, et al., 2014b).

PPVs for aggressive outcomes were moderate to large, but NPVs were around chance levels. This suggests that clinicians can be reasonably confident that those rated at elevated risk of engaging in aggression will do so and implement management strategies accordingly; however, some individuals identified as low risk are engaging in aggressive behaviours, suggesting that there may be additional risk factors that are not covered by the START. One possibility is further patient-related factors, such as recent risk behaviour. However, both staff and environmental factors have been reported as possible influences on patients' engagement in aggressive behaviours (Hallett, Huber, & Dickens, 2014) and should also be considered. In contrast, NPVs for self-harm were high, suggesting that the START may be useful as a screening tool for this outcome, such that those who are identified as low risk are not likely to engage, but the PPVs for this outcome suggest that those rated as moderate or high risk may benefit from further assessment.

The current study adds to a growing body of evidence that risk assessment instruments provide more accurate predictions of inpatient aggression and self-harm for women than for men (O'Shea et al., 2013; O'Shea, Picchioni, et al., 2014a; O'Shea, Picchioni, et al., 2014b). Interpretation of the results of studies of the predictive accuracy of the START, and of the HCR-20, should always consider the proportion of females in the sample since they are likely to inflate the effect sizes detected. There is a lack of theoretical explanation for the repeated empirical finding that risk assessment tools perform better in women than men. When attempting to quantify the relevance of group-derived data to the case at hand, clinicians must determine the degree to which it is reasonable to present the individual as if it was a case from the validation sample (Buchanan, 2013). Part of this calculation will involve

determining which differences in sample characteristics affect estimates of risk and performance of risk assessment tools. Our results suggest that gender is one such factor that should be considered, but perhaps not in the way expected. A number of researchers (e.g., de Vogel &de Vries Robbe, 2013) have suggested that current risk assessment methods perform less well in female populations than they do in males and suggest that female-specific factors may be beneficial (see Nicholls, Ogloff, & Douglas, 2004 for a review of the "gendered perspective"). However, our results suggest that risk assessment in males represents a greater development need. Since only 12% of the population of forensic mental health services are women (Rutherford & Duggan, 2007) developments in risk assessment for inpatient outcomes should focus on the factors that can best predict aggression and self-harm in men. It should not automatically be assumed that risk assessment instruments developed and validated in primarily male samples will not be relevant to females; although, of course, it may be possible that further development of female-specific tools or guidance can further refine their predictive accuracy for women.

Closer examination of individual START items revealed that those with the best predictive potency differed between men and women. S9 (impulse control) seemed a particularly relevant item for males, being among the most predictive items for all outcomes; DBT strategies such as behavioural analysis, distress tolerance skills, and emotion regulation have been suggested as possible treatment targets for impulsivity (McMain & Courbasson, 2001) and may prove useful in reducing aggression and self-harm in males. S10 (External triggers) was identified as another particularly important item for both men and women, although it was a more potent predictor in women. This is consistent with the fact that forming a new intimate relationship, care for children, and prostitution were listed as the most frequent other considerations by clinicians completing the HCR-20 (de Vogel & de Ruiter, 2005), as these can all be considered as external factors. Targeting this item may involve both

limiting exposure to environmental stabilisers, such as drugs (Swanson et al., 2002), or excessive sensory stimulation (Flannery, 2007), and improving patients' ability to recognise, avoid, or cope with triggers (Webster et al., 2009). Mental state and medication adherence were among the potent predictors of aggression in women and are closely related; it is likely that improvements in medication adherence would be linked to improvements in mental state. Borum, Swartz, Swanson, and Wiseman (2001) outline a strategy for improving treatment adherence based on engaging the patient and actively involving them in the treatment process, assessing/planning for potential barriers to compliance, and effective monitoring, which may prove useful in reducing aggression in women. However, there were a large number of non-predictive items, particularly for self-harm, suggesting some refinement of the START may be possible. Further, none of the items produced a large effect size for any of the outcomes among the male sample, suggesting that there may be more important factors for this group that are not captured by the START.

With the exception of self-harm in the female sample, Strength items were more potent predictors than the Vulnerability items, suggesting that interventions aimed at bolstering strengths may be more effective than those aimed at reducing vulnerabilities. This may run contrary to clinicians' perceptions; the fact that the mean number of critical items identified was higher than the mean number of key items identified in both samples suggests that clinicians consider Vulnerabilities to be more important and are giving them more weight. This is congruent with our previous analysis which found that Strength items had incremental validity over Vulnerability items (O'Shea, Picchioni, & Dickens, 2014). While the statistical effect may be quite small we consider that the pre-eminence of protective factors provides a potentially powerful message for clinicians about the need to approach assessment from a position of appreciation of patient's positive attributes.

Interestingly, there was no correspondence between the key Strengths or critical Vulnerabilities identified in either sample, and the most predictive items for the relevant group. This suggests that, at least at a group level, the items which clinicians are identifying as most important to the case in hand are not those that demonstrate the greatest predictive ability. It is likely that clinicians are giving extra weight to items identified as key or critical when forming their SREs. If this is the case, the current analyses suggest that although the SREs have reasonable predictive efficacy among the female samples, the items may not be being considered in the optimal manner. It is reasonable that the items identified as the most potent predictors at the group level may not be the most relevant items in all individual cases, particularly when individuals have low scores on these items. However, it would be beneficial to determine if guidance and training highlighting the most potent predictors impacts which items are considered key or critical, and the accuracy of the SREs.

Limitations

The fact that the START assessments were completed by the patients' multidisciplinary team as part of routine clinical practice may have underestimated the predictive ability of the START. The people completing the risk assessment would be the same people tasked with preventing and managing aggression and self-harm; consequently fewer incidents may occur and it may appear as if the START has falsely predicted positive outcomes, where in fact incidents have just been successfully prevented. However, this methodology is a more accurate representation of what occurs in practice than when the assessment is undertaken by researchers, as the START is intended to be completed by a multidisciplinary team. Further, the use of the START is mandated in the study setting and, therefore, clinical teams would be conducting risk assessments and implementing management strategies accordingly irrespective of whether risk assessment was completed by the research team. Future research should investigate if treatment and interventions, such as

de-escalation, restraint, and seclusion, moderate the relationship between START scores and adverse outcomes to determine the effect of informed risk management on predictive accuracy. This should be examined separately by gender to investigate as an alternative explanation to the current findings; that aggression and self-harm/suicidal behaviour has been managed more effectively in males than females, reducing the perceived predictive ability of the START for this group. A further limitation due to the reliance on routinely collected data is that we were missing a large amount of data pertaining to ethnicity, and, for males, there were high rates of missing SREs for self-harm and suicide. It is possible that the rate of missing SREs for men was much higher than women for these outcomes due to women being perceived as at increased risk of self-harm relative to males (Nijman & Campo, 2002). Therefore, it is possible that clinicians did not deem it necessary to routinely assess risk of this outcome among males.

There is currently a lack of evidence for the validity and reliability of scoring of the SOS; few studies have reported on its inter-rater reliability and those that have found a lower intraclass correlation coefficient than has been observed for other measures. For example, Wilson, Desmarais, Nicholls, and Brink (2010) reported ICC values of .85 and .90 respectively for the START Strength and Vulnerability scales, but only found a mean ICC of .68 for the SOS; similarly Nicholls, Brink, Desmarais, Webster, and Martin (2006) reported an ICC of .70 for the SOS. However, it is likely that the SOS is a valid measure of aggression and self-harm as it closely parallels the Overt Aggression Scale (Yudofsky et al., 1986) for these outcomes, which is one of the most commonly used instruments for measuring aggressive outcomes among psychiatric inpatients (Wilson et al., 2010). Further, we have found excellent inter-rater reliability for coding progress notes using the SOS as part of a previous study (O'Shea, Picchioni, & Dickens, 2014).

The fact that self-harm and suicide were investigated as a composite outcome, rather than separately as intended by the START authors (Webster et al., 2009) may have affected the predictive efficacy of the corresponding SREs. However, the definition of suicide in the START manual includes self-injurious behaviours, as well as suicide and suicide attempts, which are defined as "all behaviours that involve deliberate infliction of direct physical harm to one's body with zero intent to die as a consequence of this behaviour" (ibid p.13). Therefore, if clinicians are forming SREs for suicide using the definition provided, then the composite outcome used is consistent with this definition. If, however, clinicians are using a stricter definition of suicidal behaviour based on intent, then it is likely that the predictive efficacy of this SRE will be underestimated. Finally, the female sample was much smaller than the male sample due to the comparably fewer number of women in secure psychiatric care (Rutherford & Duggan, 2007); therefore these results require replication in larger samples. However, this sample is both the largest overall, and the largest female sample to date in the published START literature (see O'Shea & Dickens, 2014 for a review of the START literature). These results should also be replicated in correctional and community settings before findings are generalised outside of psychiatric inpatient settings.

Conclusions

Results of the current study are consistent with findings from previous research on the HCR-20 (O'Shea et al., 2013; O'Shea, Picchioni, et al., 2014a; O'Shea, Picchioni, et al., 2014b); whilst the START was able to predict aggressive outcomes in both groups, the START was a stronger predictor of inpatient aggression and self-harm in women compared with men. Item-outcome analyses suggested that impulse control is a particularly important risk factor for males and future research should investigate whether targeted efforts to reduce impulsivity minimises aggression and self-harm in this group. Targeting medication

adherence, mental state and exposure to external triggers may be relatively more important in women and should also be investigated.

References

Borum, R., Swartz, M., Swanson, J., & Wiseman, S. (2001). Compliance with remediation attempts. In K. S. Douglas, C. D. Webster, S. D. Hart, D. Eaves & J. R. P. Ogloff (Eds.), *HCR-20: Violence risk management companion guide*. Burnaby, BC, Canada Mental Health, Law, and Policy Institute, Simon Fraser University, and Department of Mental Health Law & Policy, University of South Florida.

Bowers, L., Stewart, D., Papadopoulos, C., Dack, C., Ross, J., Khanom, H., & Jeffery, D. (2011). Inpatient violence and aggression: a literature review: Report from the Conflict and Containment Reduction Research Programme; Institute of Psychiatry, Kings College London. Retrieved from: http://www.kcl.ac.uk/iop/depts/hspr/research/ciemh/mhn/projects/litreview/LitRevAgg.pdf

Buchanan, A. (2013). Violence risk assessment in clinical settings: Being sure about being sure. *Behavioral Sciences & the Law, 31*(1), 74-80. doi: 10.1002/bsl.2045

Claes, L., Vandereycken, W., & Vertommen, H. (2007). Self-injury in female versus male psychiatric patients: A comparison of characteristics, psychopathology and aggression regulation. *Personality and Individual Differences, 42*(4), 611-621. http://dx.doi.org/10.1016/j.paid.2006.07.021

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155. http://dx.doi.org/10.1037/0033-2909.112.1.155

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*(3), 226-240. doi: 10.1002/cpp.452

de Vogel, V., & de Vries Robbe, M. (2013). *Working with Women. Towards a more gender sensitive violence risk assessment.* In C. Logan, & L. Johnstone (Eds.). Managing

Clinical Risk: A guide to effective practice. London: Routledge Dolan, M., & Doyle, M. (2000). Violence risk prediction. *The British Journal of Psychiatry, 177*(4), 303-311. doi: 10.1192/bjp.177.4.303

Douglas, K., Hart, S., Webster, C., & Belfrage, H. (2013). HCR-20: Assessing risk for violence (Version 3). *Burnaby, BC: Mental Health, Law, and Policy Institute, Simon Fraser University*.

Field, A. (2009). *Discovering statistics using SPSS*: Sage publications.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*: Sage.

Flannery, R. B. (2007). Precipitants to psychiatric patient assaults: review of findings, 2004-2006, with implications for EMS and other health care providers. *International journal of emergency mental health, 9*(1), 5.

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

Hallett, N., Huber, J. W., & Dickens, G. L. (2014). Violence prevention in inpatient psychiatric settings: Systematic review of studies about the perceptions of care staff and patients. *Aggression and Violent Behavior, 19*(5), 502-514. doi:10.1016/j.avb.2014.07.009.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders the development of a statistical prediction instrument. *Criminal Justice and Behavior, 20*(4), 315-335. doi: 10.1177/0093854893020004001

Hart, S. D. (2001). Assessing and managing violence risk. In K. S. Douglas, C. D. Webster, S. D. Hart, D. Eaves & J. R. P. Ogloff (Eds.). *HCR-20 violence risk management companion guide* (pp. 13-26): Burnaby, BC, Canada: Mental Health, Law, and Policy

Institute, Simon Fraser University, and Department of Mental Health Law & Policy, University of South Florida.

James, K., Stewart, D., & Bowers, L. (2012). Self-harm and attempted suicide within inpatient psychiatric services: A review of the Literature. *International Journal of Mental Health Nursing, 21*(4), 301-309. doi: 10.1111/j.1447-0349.2011.00794.x

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological psychiatry, 59*(11), 990-996. doi:10.1016/j.biopsych.2005.09.014

McMain, S. F., & Courbasson, C. M. A. (2001). Impulse Control. In K. S. Douglas, C. D. Webster, S. D. Hart, D. Eaves & J. R. P. Ogloff (Eds.), *HCR-20: Violence risk management companion guide*. Burnaby, BC, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University, and Department of Mental Health Law & Policy, University of South Florida.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Myers, R. (1990). *Classical and modern regression with applications (2nd ed.)*. Boston, MA: Duxbury.

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*(3), 313-327. doi: 10.1177/1073191106290559

Nicholls, T. L., Gagnon, N., Crocker, A. G., Brink, J., Desmarais, S. L., & Webster, C. (2007). *START Outcomes Scale (SOS)*. Vancouver: BC Mental Health & Addiction Services.

Nicholls, T. L., Ogloff, J. R. P., & Douglas, K. S. (2004). Assessing Risk for Violence among Male and Female Civil Psychiatric Patients: The HCR-20, PCL:SV, and VSC. *Behavioral Sciences and the Law, 22*(1), 127-158. doi: 10.1002/bsl.579

Nijman, H., Bowers, L., Oud, N., & Jansen, G. (2005). Psychiatric nurses' experiences with inpatient aggression. *Aggressive Behavior, 31*(3), 217-227. doi: 10.1002/ab.20038

Nijman, H., & Campo, J. M. L. G. À. (2002). Situational determinants of inpatient self-harm. *Suicide and Life-Threatening Behavior, 32*(2), 167-175. doi: 10.1521/suli.32.2.167.24401

O'Shea, L. E., & Dickens, G. L. (2014). Short-Term Assessment of Risk and Treatability (START): Systematic review and meta-analysis. *Psychological Assessment, 26*(3), 990-1002. http://dx.doi.org/10.1037/a0036794

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. doi:10.1016/j.avb.2012.11.016

O'Shea, L. E., Picchioni, M. M., & Dickens, G. L. (2015). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) for multiple adverse outcomes in a secure psychiatric inpatient setting. *Assessment*.

O'Shea, L. E., Picchioni, M. M., Mason, F. L., Sugarman, P. A., & Dickens, G. L. (2014a). Differential predictive validity of HCR-20 for inpatient aggression. *Psychiatry Research, 220*, 669-678. doi:10.1016/j.psychres.2014.07.080

O'Shea, L. E., Picchioni, M. M., Mason, F. L., Sugarman, P. A., & Dickens, G. L. (2014b). Predictive validity of the HCR-20 for inpatient self-harm. *Comprehensive psychiatry, 55*(8), 1937-1949. doi:10.1016/j.comppsych.2014.07.010

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

Rutherford, M., & Duggan, S. (2007). Forensic mental health services: facts and figures on current provision. Retrieved from: *http://www.centreformentalhealth.org.uk/pdfs/scmh_forensic_factfile_2007.pdf.*

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review. *Behavioral Sciences & the Law, 31*(1), 55-73. doi: 10.1002/bsl.2053

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499-513. doi:10.1016/j.cpr.2010.11.009

Swanson, J. W., Swartz, M. S., Essock, S. M., Osher, F. C., Wagner, H. R., Goodman, L. A., . . . Meador, K. G. (2002). The social-environmental context of violent behavior in persons treated for severe mental illness. *American Journal of Public Health, 92*(9), 1523-1531. doi: 10.2105/AJPH.92.9.1523

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). Coquitlam, Canada: British Columbia Mental Health & Addiction Services.*

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 and the Law, 24*(6), 747-766. doi: 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*(4), 282-293. doi: 10.1080/14999013.2010.534694

World Health Organisation. (1992). *The ICD-10 classification of mental and behavioural disorders*. Geneva: WHO.

Yang, M., Wong, S. C. P., & Coid, J. W. (2013). Violence, mental health and violence risk factors among community women: an epidemiological study based on two national household surveys in the UK. *BMC Public Health, 13*, 1020. doi:10.1186/1471-2458-13-1020

Yudofsky, S. C., Silver, J. M., Jackson, W., Endicott, J., & Williams, D. (1986). The Overt Aggression Scale for the objective rating of verbal and physical aggression. *The American Journal of Psychiatry, 143*(1), 35-39.

Table 1: Differences in sample characteristics and base rates of behaviour as a function of gender

	Men (*n*=149)	Women (*n*=51)	Test
Diagnosis			$\chi^2(8, N=200)=43.95, p<.001$
(F00–F09) Organic	24 (16.1%)	1 (2.0%)	
(F10–F19) Substance use	4 (2.7%)		
(F20–F29) Schizophrenia	33 (22.1%)	6 (11.8%)	
(F30–F39) Mood	4 (2.7%)		
(F40–F48) Neurotic		3 (5.9%)	
(F60–F69) Personality disorder	3 (2.0%)	10 (19.6%)	
(F70-F79) Mental retardation	7 (4.7%)		
(F80–F89) Developmental	14 (9.4%)	2 (3.9%)	
Multiple diagnoses	60 (40.3%)	29 (56.9%)	
Ethnicity			$\chi^2(2, N=200)=13.39, p=.001$
Caucasian	50 (33.6%)	32 (62.7%)	
Non-Caucasian	11 (7.4%)	2 (3.9%)	
Unknown	88 (59.1%)	17 (33.3%)	
Mean Age (SD)	35.15 (15.8)	31.86 (13.0)	$t(198)=-1.33, p=.184$
Security Level			$\chi^2(1, N=200)=1.42, p=.233$
Low	93 (62.4%)	27 (52.9%)	
Medium	56 (37.6%)	24 (47.1%)	
Mean time (days) admission-assessment (SD)	198.04 (153.9)	154.75 (131.4)	$t(198)=-1.80, p=.074$
Legal Status			$\chi^2(2, N=200)=4.87, p=.088$
Forensic	74 (49.7%)	17 (33.3%)	
Civil	68 (45.6%)	29 (56.9%)	
Informal	7 (4.7%)	5 (9.8%)	
Behaviour			
Any Aggression	107 (71.8%)	31 (60.8%)	$\chi^2(1, N=200)=2.16, p=.142$
Physical Aggression	85 (57.0%)	23 (45.1%)	$\chi^2(1, N=200)=2.18, p=.139$
Verbal Aggression	96 (64.4%)	27 (52.9%)	$\chi^2(1, N=200)=2.12, p=.146$
Self-harm/suicidal behaviour	24 (16.1%)	24 (47.1%)	$\chi^2(1, N=200)=19.96, p<.001$
History of Violence	112 (75.2%)	21 (41.2%)	$\chi^2(1, N=200)=19.71, p<.001$
History of Self-harm	62 (41.6%)	18 (35.3%)	$\chi^2(1, N=200)=0.63, p=.427$
History of Suicide[a]	14 (27.5%)	41 (27.5%)	$\chi^2(1, N=200)=0.00, p=.993$

[a]Defined as suicide, suicide attempts, or self-injurious behaviour (Webster et al., 2009)

Table 2: Mean Strength and Vulnerability scores as a function of engagement in risk outcomes

Outcome	**Mean Strength Score (SD)**		**Mean Vulnerability Score (SD)**	
	Men	**Women**	**Men**	**Women**
Any Aggression				
Yes	14.9 (5.9)	13.2 (5.7)	25.5 (5.6)	30.3 (5.4)
No	19.8 (7.2)	16.8 (6.7)	21.6 (7.1)	26.3 (6.8)
Test	$t(147)=4.35$, $p<.001$	$t(49)=2.06$, $p=.045$	$t(62)=-3.16$, $p=.002$	$t(49)=-2.35$, $p=.023$
Physical Aggression				
Yes	14.3 (5.6)	13.8 (5.4)	26.2 (5.3)	30.3 (4.9)
No	18.9 (7.1)	15.3 (7.0)	22.0 (6.8)	27.5 (7.0)
Test	$t(118)=4.28$, $p<.001$	$t(49)=0.85$, $p=.398$	$t(115)=-4.02$, $p<.001$	$t(49)=-1.63$, $p=.110$
Verbal Aggression				
Yes	14.9 (5.9)	12.6 (5.5)	25.4 (5.5)	30.7 (5.4)
No	18.8 (7.3)	16.9 (6.5)	22.6 (7.3)	26.4 (6.4)
Test	$t(90)=3.36$, $p=.001$	$t(49)=2.60$, $p=.012$	$t(85)=-2.51$, $p=.014$	$t(49)=-2.60$, $p=.012$
Self-harm/suicidal behaviour				
Yes	14.6 (7.1)	14.0 (6.4)	24.4 (5.6)	29.2 (6.3)
No	16.6 (6.6)	15.2 (6.3)	24.4 (6.5)	28.3 (6.2)
Test	$t(147)=1.32$, $p=.189$	$t(49)=0.69$, $p=.492$	$t(147)=0.03$, $p=.975$	$t(49)=-0.53$, $p=.598$

Table 3: Risk levels assigned by the START specific risk estimates

	Low	Moderate	High	Missing	Test
Violence					$\chi^2(2, N=182)=4.15, p=.126$
Men	66 (44.3%)	45 (30.2%)	24 (16.1%)	14 (9.4%)	
Women	15 (29.4%)	20 (39.2%)	12 (23.5%)	4 (7.8%)	
Self-harm					$\chi^2(2, N=152)=21.02, p<.001$
Men	76 (51.0%)	21 (14.1%)	7 (4.7%)	45 (30.2%)	
Women	18 (35.3%)	16 (31.4%)	14 (27.5%)	3 (5.9%)	
Suicide					$\chi^2(2, N=133)=17.05, p<.001$
Men	75 (50.3%)	9 (6.0%)	2 (1.3%)	63 (42.3%)	
Women	26 (51.0%)	16 (31.4%)	5 (9.8%)	4 (7.8%)	

Table 4: Rocreg analyses for the prediction of aggression and self-harm/suicidal behaviour as a function of gender

	Rocreg		Male			Female		
	Coefficient	99%CI	AUC	99%CI	OR	AUC	99%CI	OR
Verbal								
S Total	.28	[-.40, 1.09]	.66**	[.54, 76]	1.11***	.72**	[.55, 88]	1.13*
V Total	.77**	[.05, 1.65]	.57	[.44, .68]	1.10**	.75**	[.55, .88]	1.14*
SRE	.43	[-.62, 1.07]	.59	[.47, .72]	[a]2.24	.71	[.43, .86]	[a]12.78**
					[b]3.55*			[b]13.35**
					[c]1.48			[c]1.21
Physical								
S Total	-.13	[-.84, .68]	.65**	[.54, .75]	1.14***	.62	[.41, .79]	1.02
V Total	.66	[-.03, 1.50]	.59	[.46, .70]	1.16***	.73	[.50, .86]	1.06
SRE	.70	[-.26, 1.75]	.68**	[.57, .80]	[a]3.03*	.85**	[.61, .98]	[a]52.26**
					[b]14.43***			[b]35.63**
					[c]5.80*			[c]0.61
Any								
S Total	-.02	[-.68, .68]	.68**	[.56, .78]	1.15***	.68	[.47, .86]	1.08
V Total	.62	[-.07, 1.50]	.59	[.46, .71]	1.13***	.74**	[.53, .89]	1.10
SRE	.64	[-.22, 1.49]	.68**	[.56, .81]	[a]2.79*	.83**	[.63, .98]	[a]20.61**
					[b]9.13**			[b]18.25**
					[c]3.50			[c]0.92
Self-harm/suicidal behaviour								
S Total	.10	[-.79, 1.12]	.56	[.36, .72]	1.05	.59	[.42, .76]	1.02
V Total	.86**	[.11, 1.84]	.46	[.31, .62]	1.01	.68**	[.52, .83]	1.01
SRE Self-harm	.43	[-.56, 1.82]	.68	[.46, .84]	[a]7.10***	.78**	[.54, .94]	[a]3.07
					[b]13.01**			[b]7.18*
					[c]2.98			[c]2.80
SRE Suicide	.73	[-.73, 2.24]	.51	[.19, .76]	[a]4.46	.74	[.46, .92]	[a]2.60
					[b]5.31			[b]4.18
					[c]0.41			[c]1.40

Note. START Strength scores have been inverted for the purpose of this analysis such that higher scores represented less strength; CI, confidence interval; AUC, area under the curve; OR, odds ratio; SRE, specific risk estimate

[a]Moderate-low

[b]High-Low

[c]High-mod
*p<.05, ** p<.01, ***p<.001

Table 5: Logistic regression analyses of incremental validity

		β(SE)	Wald	Δ cases correctly classified	Model Fit
Verbal - Women					
Step 1					
Vulnerability		.13* (.06)	4.65		$\chi^2(4) = 6.96$
Strength		.12* (.06)	4.63		$\chi^2(4) = 7.00$
Step 2					
Vulnerability - Strength	Vulnerability	.07 (.10)	0.47	-4.0%	$\chi^2(5) = 7.47$
	Strength	.07 (.09)	0.49		$\Delta\chi^2(1) = 0.50$
Strength -Vulnerability	Strength	.07 (.09)	0.49	-2.0%	$\chi^2(5) = 7.47$
	Vulnerability	.07 (.10)	0.47		$\Delta\chi^2(1) = 0.47$
Physical - Men					
Step 1					
Strength		.13*** (.03)	16.32		$\chi^2(4) = 29.53$***
Step 2					
Strength-SRE	Strength	.10** (.04)	8.46	5.2%	$\chi^2(5) = 39.61$***
	SRE	.99** (.33)	8.95		$\Delta\chi^2(1) = 10.08$**
Any – Men					
Step 1					
Strength		.13*** (.04)	13.51		$\chi^2(4) = 24.02$***
Step 2					
Strength-SRE	Strength	.10** (.04)	7.59	1.4%	$\chi^2(5) = 28.73$***
	SRE	1.65** (.61)	7.35		$\Delta\chi^2(1) = 4.71$*
Any – Women					
Step 1					
Vulnerability		.06 (.06)	1.03		$\chi^2(4) = 5.14$
Step 2					
Vulnerability-SRE	Vulnerability	.00 (.07)	0.00	8.6%	$\chi^2(5) = 14.98$*
	SRE	1.68** (.62)	7.28		$\Delta\chi^2(1) = 9.83$**
Self-harm/suicidal behaviour - Women					
Step 1					
Vulnerability[a]		.01 (.05)	0.05		$\chi^2(3) = 2.85$
Vulnerability[b]		.01 (.05)	0.02		$\chi^2(3) = 2.44$
Step 2					
Vulnerability - Self-harm SRE	Vulnerability	-.01 (.06)	0.03	14.5%	$\chi^2(4) = 8.86$
	SRE	1.00* (.43)	5.40		$\Delta\chi^2(1) = 6.01$*
Vulnerability – Suicide SRE	Vulnerability	.01 (.06)	.03	10.6%	$\chi^2(4) = 11.30$*
	SRE	1.46** (.54)	7.17		$\Delta\chi^2(1) = 8.86$**

Note. START Strength scores have been inverted for the purpose of this analysis such that higher scores represented less strength
[a]Vulnerability scale when the self-harm SRE is entered in step 2

[b]Vulnerablity scale when the suicide SRE is entered in step 2