Predictive validity of the Short-Term Assessment of Risk and Treatability (START) for multiple adverse outcomes: the effect of diagnosis

Rebecca Marriot, Laura E. O'Shea, Marco M. Picchioni and Geoffrey L. Dickens

Predictive Validity of the Short-Term Assessment of Risk and Treatability (START) for Multiple Adverse Outcomes: The Effect of Diagnosis

Rebecca Marriot1

Laura E. O'Shea2

Marco M. Picchioni2

Geoffrey L. Dickens3*

1Department of Forensic and Neurodevelopmental Sciences,

 Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK;

2 St Andrew's Academic Centre, King's College London Institute of Psychiatry,

Psychology and Neuroscience, Northampton, UK;

3 School of Social and Health Sciences, Abertay University, Dundee, UK

*Corresponding author: Division of Mental Health Nursing and Counselling, Abertay University, Bell Street, Dundee. DD1 1HG.

t: 07914157365 E: g.dickens@abertay.ac.uk

Abstract

Background: The Short-Term Assessment of Risk and Treatability (START) aims to aid clinicians in making evidence-based and clinically-informed Specific Risk Estimates (SREs) for their patients for seven risk outcomes based on risk ("Vulnerability") and protective ("Strength") factors. Predictive validity for violence and self-harm outcomes has been established in samples of young Caucasian males with schizophrenia, but performance across more pathologically diverse patient samples is unknown. We hypothesised that the predictive validity of START-derived predictors would differ by diagnosis, with greatest predictive efficacy for those patients most like validation samples.

Method: A pseudo-prospective cohort design was used. Routinely collected, anonymised baseline START assessments and subsequent 3-month risk event data of $N = 527$ adult, psychiatric patients resident in a UK secure facility were examined. The sample was subdivided into diagnostic groups for analysis in which potential confounding covariates were controlled for.

Results: There is considerable complexity about which elements of the START predict which outcomes and for which diagnostic groups. The predictive validity of START Strength and Vulnerability scale scores varied with diagnosis, performing best for patients with psychosis or personality disorder. The relevant SREs predicted physical aggression in all diagnostic groups, and verbal aggression, self-harm and self-neglect in most diagnostic groups. Victimisation was not well-addressed by the tool.

Conclusion: The START may help clinicians assess the future risk of aggressive, self-harm and self-neglect behavior over a three-month follow-up period across a range of diagnostic groups. However, the complex evidence means it cannot be wholeheartedly endorsed for use across all diagnostic groups, for all outcomes, and by the recommended rating method. We suggest the START be viewed as a compendium of tools, each of which requires justification for continued use both in terms of predictive validity and clinical utility.

Keywords: Risk assessment, risk management, START, violence, aggression, self-neglect, self-harm, suicide, victimisation

1. Introduction

Structured risk assessment for violence is common and accepted practice in mental health and criminal justice settings. Actuarial tools comprise schedules of empirically-derived risk factors whose presence raters are required to determine, and subsequently subject to an algorithmic scoring system to determine the probability of an individual engaging in future violence (Hart et al., 2007). Structured professional judgement (SPJ) tools have extended this approach by combining the requirement to consider empirically-derived risk factors with a degree of latitude for clinical judgement about individual cases (Guy et al., 2012).

The growth of use of these tools has led to a number of developments, particularly in the case of SPJ instruments. First, the focus of most violence risk assessment tools has been on factors that increase risk. However, some authorities contend that protective factors, "variables that reduce the effect of risk factors or influence the outcome independently" (Braithwaite et al., 2010, p. 272), have been insufficiently addressed (O'Shea and Dickens, 2016a; Rogers, 2000; Stouthamer-Loeber et al., 2002; Webster et al., 2006). From this perspective, a focus on so-called risk factors constitutes an institutionalised focus on the patient's weaknesses or limitations which could lead to an over-estimation of risk and to subsequent overly-restrictive risk management interventions. Working with a patient to identify and bolster their protective factors, or strengths, may help to develop the therapeutic alliance and facilitate the implementation of more effective risk management strategies (de Ruiter and Nicholls, 2011; van den Brink et al., 2015; Wilson et al., 2010).

Second, structured risk assessment tools have focused on relatively static, historical risk factors such as history of violence. While offering important indicators of future risk behaviour, they are inherently insensitive to change and thus have limited utility in the identification of treatment targets. In contrast, dynamic factors, for example the severity and nature of active symptoms of major mental illness, can change with time, are associated with

changes in risk behaviour (Hanson and Harris, 2000) and can add significant incremental validity to risk assessment (Doyle and Dolan, 2006). Therefore, they can usefully contribute to violence risk assessment (Chu et al., 2013; Grevatt et al., 2004; McNiel et al., 2003; Wilson et al., 2013), and management strategies (Whittington et al., 2014).

Additionally, most risk assessment tools used in mental health settings have traditionally been concerned with violence or suicide, and do not aim to inform clinicians about other important risk outcomes such as self-neglect (Gunstone, 2003), or absconding (Muir-Cochrane and Mosel, 2008). Some therefore contend that risk assessment should address a wider range of adverse outcomes to which patients might be at risk (Webster et al., 2009). Finally, there has been increased interest in the prediction of risk behaviour over shorter time-periods than achieved by established instruments like the HCR-20 (6 months; Webster et al., 1999). In inpatient settings, particularly, it may be advantageous to conduct more regular risk assessments and make management adjustments accordingly.

1.1 The Short-Term Assessment of Risk and Treatability (START)

The authors of the START (Webster, et al., 2004; Webster et al., 2009) attempted to address all of the above issues. The START requires raters to consider 20 items both as risk factors ("Vulnerabilities") and protective factors ("Strengths"). Items were selected for their dynamic nature and thus suitability for identifying treatment targets. Raters are required to consider these factors to inform and augment their clinical judgement to make a Specific Risk Estimate (SRE) in seven risk domains: violence to others, self-harm, suicide, substance misuse, victimisation, unauthorised leave and self-neglect. Finally, the START aims to assist assessment for the three-month period ahead, half as long as that recommended for iteration of the widely used HCR-20 (Webster et al., 1997).

A systematic review and meta-analysis (O'Shea and Dickens, 2014) revealed that the START was internally consistent and has convergent reliability with other risk measures.

Predictive validity of the tool for aggression in studies rated as low risk of bias (Braithwaite et al., 2010; Desmarais et al., 2012; WIlson et al., 2013) ranges from .65-.84 (Strength scale), .66-.82 (Vulnerabilities scale), and .52-.89 (Violence Risk Estimate). Studies of other risk outcomes are rarer; pooled effect sizes from all available studies have produced small effect sizes for rediction of self harm, self neglect and victimisation (O'Shea & Dickens, 2014). The START is valued by mental health workers who find it easy to use and it has acceptable inter-rater reliability and good predictive validity for violence and self-harm (Doyle et al., 2008). Subsequent research has demonstrated that respective SREs are predictive of self-harm/suicidality and victimisation (O'Shea et al., 2016), and there is recent evidence that the START may, to an extent, be predictive of substance misuse and unauthorised leave (O'Shea and Dickens, 2015a). Despite this, most of the START literature has been conducted in relatively small samples of young Caucasian males with schizophrenia. Two of the current authors have recently shown that the START has better predictive validity for women than men for aggression and self-harm outcomes (O'Shea and Dickens, 2015b).

1.2 Aim of the current study

One remaining gap concerns the START's performance across different diagnostic groups. From a theoretical perspective, the predictive efficacy of risk assessment tools is expected to be maximal in populations similar to validation samples (Buchanan, 2013). Empirically, there is evidence that the predictive efficacy of the HCR-20 performs across diagnoses in a manner broadly consistent with this (Gray et al., 2011; O'Shea et al., 2014). Therefore, it would be expected that the START would predict risk behaviour best among samples with schizophrenia or personality disorder diagnoses (Nicholls et al., 2006). We have therefore conducted a study to test the predictive validity of the START Strength and Vulnerability scores and SREs as a function of diagnosis, whilst controlling for potential covariates such as gender, age and ethnicity.

2. Method

2.1 Participants and Setting

St Andrew's provides specialist secure psychiatric inpatient care at four hospitals in England. START assessment is routinely conducted by clinical staff. Inclusion criteria were: inpatients resident between May 2011 and January 2014 aged 18 years or older, with one or more completed START assessments and a subsequent 3-month inpatient stay. Exclusion criteria were: missing START-item data in excess of pro-rating guidelines (Webster et al., 2009). Further, since we aimed to examine predictive validity by diagnostic group, exclusions were made to maximise within diagnostic-group homogeneity, namely: i) patients with complex, multiple diagnostic comorbidity, i.e., mental and behavioral disorder diagnoses under three or more category headings of the World Health Organizations' (2004) International Classification of Diseases version 10 (ICD-10); ii) those with an ICD-10 diagnosis of intellectual disability (F70-79); and iii) and those assigned to a diagnostic grouping with fewer than 20 patients. Required sample size to detect significant predictive ability for each risk outcome was calculated using MedCalc for Windows (MedCalc software, Ostend, Belgium) and was based on an expected large effect size ($\alpha=0.05$ and $\beta=0.20$). Sample sizes in the range of $n = 39$ for physical aggression to $n = 115$ for self-neglect were required. For substance misuse and absconding, $n = 734$ and $n = 367$ cases respectively were needed to detect a significant effect. These outcomes were not analyzed due to insufficient sample size and because low base rates for these outcomes revealed in a previous study (1.5% and 3.0%; O'Shea et al., 2016), suggested that they would not occur in all groups..

2.2 Study design

We employed a pseudo-prospective cohort design.

2.3 Procedure

The study was approved by the St Andrew's Head of Clinical Audit. All data were anonymized before analysis and we were advised that approval was not required from an NHS Research Ethics Committee. Details of the first START assessment completed, clinical and demographic data, and all recorded narrative entries in the patient record describing risk incidents related to START outcomes in the three months after the START assessment were obtained.

2.4 Measures

2.4.1 Demographic and diagnostic data

Data included: age, gender, self-reported ethnicity, date of admission, discharge, security level, legal status, and ICD-10 (World Health Organization, 1992) psychiatric diagnosis as assigned by the patient's consultant psychiatrist. Patients were grouped according to major ICD-10 category headings, i.e., organic disorders (F00-09); schizophrenia, schizotypal and delusional disorder (F20-29), disorders of adult personality and behaviour (F60-69); disorders of psychological development (F80-89). Patients with diagnoses under two major ICD-10 categories were assigned to 'dual morbidity' groups ('psychosis and personality disorder'; 'psychosis and developmental disorder'). To reduce diagnostic complexity, substance abuse (ICD-10 F10-19) was recorded as a separate dichotomous variable for each patient.

2.4.2 Risk assessment data

The START (Webster et al., 2009) comprises 20 empirically-derived items which are considered to be dynamic risk factors for seven adverse outcomes: violence, self-harm, suicide, substance misuse, victimization, self-neglect and unauthorized leave. Each is rated independently in terms of the patient's relevant i) strengths, and ii) vulnerabilities. A rating of 0 represents no relevant strengths/vulnerabilities; 1 represents some relevant strengths/vulnerabilities; and 2 represents definite strengths/vulnerabilities. Raters assign a

specific risk estimate (SRE; defined as 'low', 'medium'. or 'high') for each of the seven risk domains over the coming three months; this should be based on the 40 resulting item ratings, the historical occurrence of risk behaviors, any other risk assessments (e.g., HCR-20; Webster et al, 1997), and the imminence and severity of the risk. Guidance about SRE formulation in the START manual is brief but emphasizes that raters should consider all factors, and should not base SREs solely on strength and vulnerability scores. No 'cut-off' scores are provided.

START assessments were completed routinely by members of the Multi-Disciplinary Team (MDT) for each patient. Local policy mandates that the completed START assessment be authorized by three team members. Raters attend a structured one-day training course including theoretical and practical aspects of the assessment based on the START manual, rating of test cases, and feedback from experienced START raters. This training exceeds any training recommended by the START authors.

2.4.3 Risk outcome data

Local policy requires a narrative entry in each patient's electronic record during each nursing shift (up to three daily) by a qualified member of staff. Prior to electronic verification the author is invited to add one or more 'flags' to denote the presence and nature of any risk outcomes; each note can be associated with multiple flags. Anonymized versions of all narrative entries for each patient flagged as an incident involving "Verbal Aggression", "Physical Aggression", "Absconding/ Unauthorized leave", "Self-harm/Suicide", "Self-neglect", "Substance Misuse" or "Vulnerability/ Victimization" were collated. They were read and coded by authors 1,2, and 4 using the START Outcomes Scale (SOS; Nicholls et al., 2007); raters were blinded to the risk assessment at the time of data coding. The SOS comprises 12 categories of risk behaviour (verbal aggression, physical aggression against others, aggression against property, sexual aggression, self-harm, suicidal behaviour, suicidal ideation and planning, unauthorized leave, being victimized, self-neglect, substance misuse,

and stalking). Each is rated on a five-point scale (0 = none or no information to 4 = most severe); each point anchored by a brief description of severity of the risk engagement. Incidents rated at '1' or above were included and patients grouped for each outcome into dichotomous categories representing their engagement/non- engagement during follow-up. The SOS has good inter-rater reliability (Wilson et al., 2010). Since START assessment does not aim to predict stalking or sexual aggression specifically, these outcomes were only included if the event was flagged for physical or verbal aggression and described an incident that could be coded as such. SOS categories were grouped to fit with the START risk domains: physical aggression included aggression against property and physical aggression against others. Due to the difficulty of ascertaining intention, self-harm/suicide was treated as one outcome including self-harm, suicidal ideation and planning, and suicide behaviors (Gray et al., 2011).

2.5 Data analysis

SOS inter-rater reliability was measured using a sample of 20 patients whose data had been coded by authors LO and GD for a previous study (O'Shea et al., 2016). Total scores on the Strength and Vulnerability scales were pro-rated for cases with up to four missing items in line with guidance in the START manual. Descriptive statistics were computed for participant demographics, risk outcome base rates, and the distribution of START Strength and Vulnerability scores and SREs. Pearson's Chi-squared tests were calculated to ascertain significant differences between diagnostic groups on demographic, clinical, legal, and risk outcome variables. Independent variables which differed significantly (indicated by standardized residual $Z = \pm 1.96$ equivalent to $P = 0.05$) were controlled for during rocreg analysis.

Receiver Operating Characteristics (ROC) analysis is the preferred method of evaluating the predictive validity of risk assessment tools (Dolan and Doyle, 2000). Area

Under the Curve (AUC) values derived from ROC analysis were used to examine the predictive validity of the START Strength scores, Vulnerability scores, and SREs respectively. ROC analysis is relatively independent of base rates in comparison to other statistical methods (Mossman, 1994) and is not dependent on how the risk assessment is coded (Pepe et al., 2004), thus enabling comparison of AUCs derived from different tools. The AUC value represents the probability that a randomly chosen patient who has been involved in an incident will have a higher START score or an elevated SRE compared with a randomly selected patient who has not (Rice and Harris, 1995). AUC values can range from 0.0 (perfect incorrect prediction) through 0.5 (chance level of prediction like flipping a coin) to 1.0 (perfect correct prediction). Rice and Harris (2005) reported that AUC values of 0.56, 0.64 and 0.71 are equivalent to small (0.2), moderate (0.5) and large (0.8) Cohen's d values (Cohen, 1992), respectively. The rocreg function in Stata Version 12.0 for Windows (TX: Statacorp LLC) was used to calculate AUC values for the Strength and Vulnerability scores and SREs, and to examine differences in predictive validity between diagnostic groups while controlling for the significant demographic, clinical and legal covariates. Significance is inferred by the absence of zero from 99% confidence intervals (equivalent to $p<0.01$). We have previously demonstrated through examination of the variance inflation factor in a subset of the current data that multicollinearity did not present a problem (O'Shea & Dickens, 2015a).

Since AUC values are known to vary with the range of scores examined, and odds ratios (ORs) are unaffected by this issue (Hanson, 2008), ORs were calculated in SPSS using a binomial logistic regression. ORs represent the increase in odds of an outcome for every one point increase on the Strength and Vulnerability scores; and, for the SREs, the odds of an outcome for those categorized as high or moderate risk, compared to low risk, and high risk

compared to moderate risk. Odds ratios can be considered small (1.5), moderate (2.5), large (4), or very large (10) (Rosenthal, 1996).

In all cases, the violence SRE was used to predict both verbal and physical aggression outcomes, and both the self-harm and suicide SREs were used to predict the composite self-harm/suicide outcome. Since SREs were sometimes missing (see Supplementary Table S2), AUCs were derived from the ratings only of those whose SREs were recorded. The Strength score in all AUC analyses was inverted, so that a higher score corresponded with a lower Strength score, to enable comparison with the predictive efficacy of the Vulnerability scores and SREs. Except where stated, analyses were conducted using SPSS Version 22.0.

3. Results

3.1 Inter-rater reliability

Inter-rater reliability for the coding of each type of risk incident from progress notes with the SOS was satisfactory to excellent (Cohen's Kappa range 0.64 for self-neglect to 1.00 for verbal and physical aggression).

3.2 Participant demographics

The sampling frame comprised $N = 875$ patients. Application of inclusion and exclusion criteria left a cohort of $N = 527$ (60.2%). Reasons for exclusion were: assigned to a diagnostic group $n < 20 = 104$; complex case > 3 ICD-10 major classifications $= 95$; ICD-10 intellectual disability diagnosis $= 94$; missing excessive START ratings $= 48$; no diagnosis recorded $= 7$). Analysis of demographic and clinical characteristics between diagnostic groups revealed a number of significant differences which were controlled for in subsequent rocreg analyses (see Table 1).

>> Insert Table 1 about here<<

3.3 Risk outcome base rates

The most common risk outcome was verbal aggression, and the least common self-neglect (see Table 2). There were significant differences between the proportions in each diagnostic group who engaged in physical aggression and self harm/ suicidality (see Table 2).

>>Insert Table 2 about here<<

3.4 Predictive validity

3.4.1 Whole sample

No SRE was recorded for 10.6-31.9% (Mdn=15.6%) of START assessments. AUC values for the whole sample ranged from 0.54 to 0.81 (see Table 3). The START Strength and Vulnerability scores had good predictive validity, with increased Vulnerability scores being associated with greater probability of all risk outcomes, and greater Strength scores being associated with a lower probability of all risk outcomes. A large effect size (>0.75) was present for the prediction of self-harm by the self-harm SRE while other results were generally in the moderate to large range. By way of contrast the effect size for the victimization outcome was small and the only non-significant result was for the victimization SRE.

>>Insert Table 3 about here<<

Calculation of ORs (see Table 3) revealed that increases in total Vulnerability and inverted Strength scores were associated with a small but statistically significant increase in the occurrence of all five outcomes. Patients designated high risk were at increased odds of engaging in the corresponding outcome compared to those classified low risk (ORs = moderate for verbal aggression and victimization, large for physical aggression and self-neglect, and very large for self-harm/suicide). Compared to those designated low risk, patients rated moderate risk for the corresponding SRE were at increased odds of engaging in physical aggression, verbal aggression and self-harm/suicide (ORs = small, moderate and large range respectively). Those classified at high risk for self-harm/suicide, self-neglect, and

12

victimization had increased odds of engaging in the corresponding outcome (OR = large for all three cases).

3.4.2 By diagnosis

AUC values ranged from 0.49-0.95. The total Vulnerability and Strength scores significantly predicted verbal aggression as an outcome in all diagnostic groups, and physical aggression in the organic, psychosis and personality disorder groups (see Table 4). The total Strength score also significantly predicted physical aggression in the developmental disorder group. The total Vulnerability score predicted self-harm and suicide in the personality disorder and developmental disorder groups; self-neglect in the organic, psychosis, personality disorder, and developmental disorder groups; and victimization in the psychosis group. The total Strength score predicted self-harm and suicide in all groups except the organic group; self-neglect in all groups except the developmental, and psychosis and personality disorder diagnostic groups; and victimization in the organic, psychosis, and personality disorder diagnostic groups. The violence SRE significantly predicted physical aggression in all diagnostic groups, and verbal aggression in all groups except those with dual diagnoses. The self-harm SRE significantly predicted self-harm/suicide outcomes in all diagnostic groups except the organic group. The suicide SRE did not predict self-harm/suicide behaviors in any diagnostic group. The self-neglect SRE predicted related outcomes in all diagnostic groups except the organic and psychosis groups. The victimization SRE did not predict the equivalent outcome in any group. The self-neglect SRE was the only risk estimate for which accuracy differed significantly between diagnostic groups (See Figure 1 for visual representation of results).

>>Insert Figure 1 about here<<

>>Insert Table 4 about here<<

4. Discussion

The START is a relatively new SPJ tool for assessment of short-term risk for multiple outcomes. Its psychometric properties have been investigated reasonably comprehensively (O'Shea and Dickens, 2014; O'Shea and Dickens, 2015a) but its predictive validity for people with different psychiatric diagnoses has received little attention. Our analysis revealed that, for the entire cohort, Strength and Vulnerability scores significantly predicted all risk outcomes. Further, all SREs, with the exception of that for victimisation, significantly predicted their associated outcomes, and the suicide SRE was not predictive of a composite self-harm/suicidality outcome. Patients classified as high risk by the SREs had significantly increased odds of engaging in corresponding behaviours over the next three months compared with those at low risk. Overall, levels of predictive ability for the entire cohort were very similar to those revealed in the same setting in a previous study (O'Shea et al., 2016), and were within the range of those found in other individual studies rated as low risk of bias in our meta-analysis (O'Shea & Dickens, 2014).

4.1 Predictive validity by diagnosis

The diagnostic groups differed significantly on each of the potential covariates (gender, ethnicity, security level, Mental Health Act status, substance misuse and age). When controlled for in the analysis, the promising level of performance achieved for the whole cohort was substantially replicated in single-diagnosis groups.We hypothesized that the START would perform significantly in those with psychosis or personality disorder. Vulnerability and Strength scores in the psychosis group and the personality disorder group significantly predicted all outcomes except that the Vulnerability score was not predictive of self-harm/suicide in the psychosis group or victimisation in the personality disorder group In single-diagnosis groups less like validation samples results were mixed; outcomes for those with a diagnosis of developmental disorder were largely predicted in line with those for the whole cohort while self-harm was not predicted by any scale or estimate for those with an

organic diagnosis. Predictive efficacy for the dual diagnosis groups was patchy though victimization was the only outcome for either psychosis and personality disorder or psychosis and developmental disorder groups which failed to yield a significant predictor.

The START's developers recommend that the SREs, rather than the scale scores, provide the basis for appropriate risk management planning. For some outcomes this was partially supported in the current study since the SREs improved on scale scores indicating that rating teams are adding value that is not captured by scale scores alone. For other outcomes, the Strength and Vulnerability scores outperformed the SREs, suggesting that additional formulation added little value. Specifically, in the psychosis group, the Strength and Vulnerability scores predicted all outcomes significantly, with the exception of the Vulnerability score for self-harm/suicide. However, SRE formulation did not build on this since those for suicide, self-neglect and victimization did not significantly predict associated outcomes.

4.2 Risk outcome base rates

Base rates of all risk outcomes were generally lower than in other studies (Crocker et al., 2011). The more passive outcomes such as self-neglect and victimization were particularly low in comparison to those other studies, possibly reflecting issues with the validation of the SOS for these outcomes, or indicating that these types of incidents were under-reported by staff. However, in those other studies, more patients were in civil rather than secure/forensic settings so the difference may be explicable with reference to patient need and environment, for example, more acutely unwell patients in civil settings and fewer risk management practices to reduce these outcomes together with more negative symptoms and self-neglect driving the need for admission). Further, most previous studies have investigated START properties in populations with a high proportion (80-90%) of patients with a diagnosis of psychosis (Braithwaite et al., 2010; Chu et al., 2011; Desmarais et al., 2012; Nonstad et al.,

2010). The patient population in this current study was much more diagnostically diverse, with only around 50% of patients with a diagnosis of schizophrenia, and a much higher proportion of people with an organic or personality disorder. Further, only 11.8% of people had a diagnosis of substance misuse compared to higher rates (52.5% and 53.0%) reported by Desmarais et al. (2012) and Wilson et al. (2013), but, again, this could be due to differences in setting or reporting. Similar rates to those in the current study were reported from a UK setting by Gray et al. (2011).

4.3 START predictor elements

Raters made use of almost the full range of possible Strength and Vulnerability scale scores, and the whole range of SREs. For aggressive outcomes, neither the Vulnerability score or violence SRE added predictive value to the Strength scale score. However, Vulnerability scale scores were more predictive of self-neglect. Figure 1 demonstrates that reliance on the SRE across all diagnoses for all outcomes is unlikely to always result in the most accurate risk prediction. We have previously noted the lack of conceptual certitude around the relationship between protective and risk factors (O'Shea and Dickens, 2016a). It was observed that (inverted, so lower) Strength scores were more strongly predictive of the more active risk outcomes (verbal and physical aggression and self-harm/suicide) whereas higher Vulnerability scores were more strongly predictive of self-neglect (which could be considered a more passive outcome).

Although the Strength and Vulnerability scores have good predictive validity, with large AUCs for some outcomes, the AUC for the SREs was often greater. This indicates that raters completing the START are not just using the Strength and Vulnerability scores but are considering extra clinical information, using their experience and focusing on the most pertinent items, as would be expected for an SPJ tool. However, this should not automatically be assumed to be a triumph for the recommended START assessment process: O'Shea and

Dickens (2016b) found that it was the presence of similar risk outcomes in the period preceding the assessment that uniquely predicted most outcomes. Interestingly, despite the wide range of factors the START authors recommend that raters consider, consideration of recent related behaviour is not explicitly one of them.

4.4 Study limitations

This study employed a pseudo-prospective cohort method which provided a large sample size. Arguably, the naturalistic risk assessment data collection increases the ecological validity, but there are some limitations. Notably, one group in particular (those with dual diagnoses of psychosis and developmental disorder) was fairly small. It is possible that significant predictive ability within this group was missed; further, predictive ability in more sizeable groups could have been underestimated for low base rate outcomes like victimisation. Clinical diagnosis relied on routine assessment by the patient's consultant psychiatrist and could not be confirmed by structured methods. Substance abuse and unauthorized leave outcomes were not examined due to insufficiently large sample size due to their low base rates. The self-harm/suicide outcomes could not be differentiated due to difficulty in ascertaining the intentions of deliberate self-harm and suicide behaviors.

Since patients had often been inpatients for a long period prior to the introduction of the START, the mean period between admission and START assessment was in excess of 2 years. The possible limitation is that patients were already well known to their clinical team and may have been managed more effectively than if they were new admissions. It is known that START scores vary as patients progress from conditions of higher security down the care pathway towards discharge (Nicholls et al., 2011), as would be expected with dynamic risks, and also that the latest START assessment does not always predict risk most accurately (Wilson et al., 2010). It is not known if its predictive accuracy changes over time. Between one in ten and three in ten SREs were missing and we note this as a limitation of the study.

We did not explore the severity of risk outcomes, so it is not known if interventions were put in place that moderated the level of violence and prevented the more severe outcomes on the SOS scale, even if it did not prevent incidents occurring altogether. Like other methods of measuring aggression, the SOS is limited in that only objective outcomes - for example, resulting harm - can be measured, rather than intention. Finally, the study was conducted in a secure psychiatric setting and results should not be considered general sable to other settings.

4.5 Clinical implications

The somewhat complex picture emerging from our analysis of START predictive validity by clinical diagnosis leads us to conclude that there is no strong case to support the use of the START *unreservedly* or in its *entirety* either across diverse diagnostic groups, or in terms of which of its elements are used to guide risk rating, or for all five outcomes studied here. Nevertheless, with the exception of the victimization SRE, no START element was without at least some merit across diagnostic groups. The implication to consider is, therefore, whether clinicians should, when conducting a START assessment, be expected to bear in mind the evidence about which elements of the tool best predict which outcomes, and in patients with which specific diagnoses. Two of us have demonstrated elsewhere (O'Shea and Dickens, 2016) that the recommended elements of the START assessment, while numerous and including the SRE, were not better predictors of a range of outcomes than a simple heuristic based on recent pre-assessment occurrence of the same behaviour.

We suggest that the clear advantages of the START - for example, that it reminds clinicians to widen their perspective on risk, and to consider protective as well as risk factors - make perseveration worthwhile. However, for research purposes it may be helpful to view START not as a single risk assessment tool which can be conveniently categorized as having either acceptable or unacceptable psychometric properties. Rather, it is, in actuality, a

compendium or suite of assessment tools. For any risk assessment tool to be useful, it must inform treatment and risk management to reduce risk and not just predict occurrence of outcomes. For the START to claim a comprehensive evidence base, then, it must demonstrate both predictive and clinical value for all of the outcomes it intends to address. Further, to have value in informing treatment targets, it should be clear which if any elements of the assessment, are identifying causal risk factors. Thus, while the START has promise, it is currently presented as a single tool with multiple purposes, an oversimplification which leads, paradoxically, to complexity in its actual operation.

References

Abidin, Z., Davoren, M., Naughton, L., Gibbons, O., Nulty, A., Kennedy, H. G., 2013. Susceptibility (risk and protective) factors for in-patient violence and self-harm: prospective study of structured professional judgement instruments START and SAPROF, DUNDRUM-3 and DUNDRUM-4 in forensic mental health services. BMC Psychiatry 13, 197. DOI:10.1186/1471-244X-13-197

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). Int. J. Forensic. Ment. Health 9, 271-281. DOI:10.1080/14999013.2010.534378

Buchanan, A., 2013. Violence risk assessment in clinical settings: Being sure about being sure. Behav. Sci. Law, 31, 74-80. DOI:10.1002/bsl.2045

Chu, C. M., Thomas, S. D., 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.). Int. J. Forensic. Ment. Health, 10, 337-345. DOI:10.1080/14999013.2011.629715

Chu, C. M., Thomas, S. D., 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.   DOI:10.1177/1073191111418298

Cohen, J., 1992. A power primer. Psychol. Bull, 112, 155. DOI:10.1037/0033-2909.112.1.155

Crocker, A.G., Braithwaite, E., Laferrière, D., Gagnon, D., Venegas, C., Jenkins, T., 2011. START changing practice: Implementing a risk assessment and management tool in a civil psychiatric setting. Int. J. Forensic. Ment. Health 10, 13-28. DOI:10.1080/14999013.2011.55

de Ruiter, C., Nicholls, T. L., 2011. Protective factors in forensic mental health: A new frontier. Int. J. Forensic. Ment. Health, 10, 160-170. DOI:10.1080/14999013.2011.600602

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. Psychol. Assessment, 24, 685-700. DOI:10.1037/a0026668

Dickens G. L., Picchioni M., Long C., 2013. Aggression in specialist secure and forensic inpatient mental health care: incidence across care pathways. J. Forensic. Practice 15, 206-217. DOI: http://dx.doi.org/10.1108/JFP-09-2012-0017

Dolan, M., Doyle, M., 2000. Violence risk prediction. Br. J. Psychiatry 189, 303-311. DOI:10.1192/bjp.177.4.303

Doyle, M., Dolan, M., 2006. Predicting community violence from patients discharged from mental health services. Br. J. Psychiatry, 189, 520-526. DOI:10.1192/bjp.bp.105.021204

Doyle, M., Lewis, G., Brisbane, M., 2008. Implementing the Short-Term Assessment of Risk and Treatability (START) in a forensic mental health service. BJPsych Bull, 32, 406-408. DOI: 10.1192/pb.bp.108.019794

Field, A., 2009. Discovering statistics using SPSS: Sage Publications.

Fitzgerald, S., Gray, N. S., Alexander, R. T., Bagshaw, R., Chesterman, P., Huckle, P., et al., 2013. Predicting institutional violence in offenders with intellectual disabilities: the predictive efficacy of the VRAG and the HCR- 20. J. Appl. Res. Intellect. Disabil 26, 384–393. DOI: 10.1111/jar.12032

Gray, N. S., Benson, R., Craig, R., Davies, H., Fitzgerald, S., Huckle, P., et al., 2011. The Short-Term Assessment of Risk and Treatability (START): A prospective study of

inpatient behavior. Int. J. Forensic. Ment. Health, 10, 305-313.
DOI:10.1080/14999013.2011.631692

Grevatt, M., Thomas-Peter, B., Hughes, G., 2004. Violence, mental disorder and risk assessment: can structured clinical assessments predict the short-term risk of inpatient violence? J. Forensic. Psychiat. Psychol, 15: 278–292.
DOI:10.1080/1478994032000199095

Gunstone, S., 2003. Risk assessment and management of patients whom self neglect: a 'grey area' for mental health workers. J. Psychiatr. Ment. Health. Nurs 10, 287–296.
DOI:10.1046/j.1351-0126.2002.00568.x

Hanson, R., 2008. What statistics should we use to report predictive accuracy. Crime Scene 15, 15-17.

Hanson, R. K., Harris, A. J. R., 2000. Where should we intervene? Dynamic predictors of sexual offense recidivism. Crim. Justice. Behav 27, 6–35.
DOI:10.1177/0093854800027001002

Hart, S. D., Michie, C., Cooke, D. J., 2007. Precision of actuarial risk assessment instruments. Evaluating the 'margins of error' of group v. individual predictions of violence. Br. J. Psychiatry, 190, 60-65. DOI:10.1192/bjp.190.5.s60

Kettles, A. M., Moir, E., Woods, P., Porter, S., Sutherland, E., 2004. Is there a relationship between risk assessment and observation level? J. Psychiatr. Ment. Health. Nurs 11, 156–164. DOI:10.1111/j.1365-2850.2003.00699.x

McNiel, D. E., Binder, R. L., 1991. Clinical assessment of the risk of violence among psychiatric inpatients. Am. J. Psychiatry 148, 1317-1321. DOI: 10.1176/ajp.148.10.1317

McNiel, D. E., Gregory, A. L., Lam, J. N., Binder, R. L., Sullivan, G. R., 2003. Utility of decision support tools for assessing acute risk of violence. J. Consult. Clin. Psychol, 71, 945–953. DOI:10.1037/0022-006X.71.5.945

Mossman, D., 1994. Assessing predictions of violence: Being accurate about accuracy. J. Consult. Clin. Psychol 62, 783. DOI:10.1037/0022-006X.62.4.783

Muir-Cochrane, E., Mosel, K. A., 2008. Absconding: A review of the literature 1996-2008. Int. J. Mental. Health. Nurs, 17, 370-378. DOI: 10.1111/j.1447-0349.2008.00562.x

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.
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 and Addiction Services.

Nicholls, T. L., Petersen, K., Brink, J., Webster, C., 2011. A clinical risk profile of forensic psychiatric patients: Treatment team STARTs in a Canadian service. Int. J. Forensic. Ment. Health, 10, 187-199. DOI:10.1080/14999013.2011.600234

Nonstad, K., Nesset, M. B., Kroppan, E., Pedersen, T. W., Nøttestad, J. Aa., Almvik, R., 2010. Predictive validity and other psychometric properties of the Short-Term Assessment of Risk and Treatability (START) in a Norwegian high secure hospital. . Int. J. Forensic. Ment. Health 9, 294–299. DOI:10.1080/14999013.2010.534958

O'Shea, L. E., Dickens, G. L., 2014. Short-Term Assessment of Risk and Treatability (START): Systematic review and meta-analysis. Psychol. Assessment 26, 990-1002.
DOI:10.1037/a0036794

O'Shea, L. E., Dickens, G. L., 2015a. 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. . Int. J. Forensic. Ment. Health 14, 132-146. DOI:10.1080/14999013.2015.1033112

O'Shea, L. E., Dickens, G. L., 2015b. Predictive validity of the Short-Term Assessment of Risk and Treatability (START) for unauthorised leave and substance abuse in a secure mental health setting: A pseudo-prospective cohort study. Int. J. Nurs. Studies 52, 970-979. DOI:10.1016/j.ijnurstu.2015.02.007

O'Shea, L. E., Dickens, G. L, 2016a. Performance of protective factors assessment in risk prediction for adults: Systematic review and meta-analysis. Clin. Psychol-Sci. Pr 23, 126–138. DOI:10.1111/cpsp.12146

O'Shea, L. E., Dickens, G. L., 2016b. Role of assessment components and recent adverse outcomes in risk estimation and prediction: Use of the Short Term Assessment of Risk and Treatability (START) in an adult secure inpatient mental health service. Psychiatry. Res 30, 398–405. DOI:10.1016/j.psychres.2016.04.068

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. Aggress. Violent. Behav. 18, 255–270. DOI:10.1016/j.avb.2012.11.016

O'Shea, L.E., Picchioni, M.M., Mason, F.L., Sugarman, P.A., Dickens, G.L., 2014. Differential predictive validity of the Historical, Clinical and Risk Management Scales (HCR–20) for inpatient aggression. Psychiatry. Res 220, 669-678. DOI: 10.1016/j.psychres.2014.07.080

O'Shea, L. E., Picchioni, M. M., Dickens, G. L., 2016. The predictive validity of the Short-Term Assessment of Risk and Treatability (START) for multiple adverse outcomes in a secure psychiatric inpatient setting. Assessment 23,150-162. DOI:10.1177/1073191115573301

O'Shea, L. E., Picchioni, M. M., McCarthy, J., Mason, F. L., Dickens, G. L., 2015. Predictive validity of the HCR-20 for inpatient aggression: the effect of intellectual disability on accuracy. J. Intellect. Disabil. Res 59, 1042-1054. DOI:10.1111/jir.12184

Ogloff, J.R.P., Daffern, M., 2006. The dynamic appraisal of situational aggression: an instrument to assess risk for imminent aggression in psychiatric inpatients. Behav. Sci. Law 24, 799-813. DOI: 10.1002/bsl.741

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., Newcomb, P., 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am. J. Epidemiol 159, 882–890. DOI:10.1093/aje/kwh101

Rice, M. E., Harris, G. T., 1995. Violent recidivism: Assessing predictive validity. J. Consult. Clin. Psychol 63, 737-748. DOI:10.1007/s10979-005-6832-7

Rice, M. E., Harris, G. T., 2005. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law. Hum. Behav 29, 615. DOI: 10.1007/s10979-005-6832-7

Rogers, R., 2000. The uncritical acceptance of risk assessment in forensic practice. Law. Hum. Behav 24, 595–605. DOI:10.1023/A:1005575113507

Rosenthal, J. A., 1996. Qualitative descriptors of strength of association and effect size. J. Soc. Serv. Res 21, 37-59. DOI:10.1300/J079v21n04_02

SPSS Inc., 2009. PASW Statistics for Windows (Version 18). Chicago: SPSS Inc.

St Andrew's., 2015. Clinical Risk Management Policy, CRM02 (Version 8.0). Anderson, M., Natarajan, M., Long, C.

Stouthamer-Loeber, M., Loeber, R., Wei, E., Farrington, D.P., Wikströrm, P.O. 2002. Risk and promotive effects in the explanation of persistent serious delinquency in boys. J. Consult. Clin. Psychol 70, 111-123. DOI:10.1037/0022-006X.70.1.111

van den Brink, R. H. S., Troquete, N. A. C., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R.A., et al., 2015. Risk assessment by client and case manager for shared

decision making in outpatient forensic psychiatry. BMC Psychiatry 15, 120. DOI:10.1186/s12888-015-0500-3

Viljoen, S., Nicholls, T. L., Roesch, R., Gagnon, N., Douglas, K., Brink, J., 2016. Exploring gender differences in the utility of strength-based risk assessment measures. Int. J. Forensic. Ment. Health 15, 149-163. DOI:10.1080/14999013.2016.1170739

Webster, C. D., Douglas, K. S., Eaves, D., Hart, S. D., 1997. HCR-20: Assessing risk for violence version 2. Mental Health, Law and Policy Institute, Simon Fraser University.

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, Canada: BC Mental Health and Addiction Services.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., Middleton, C., 2004. Manual for the Short-Term Assessment of Risk and Treatability (START) (Version 1.0 Consultation Edition). Hamilton, Canada: St. Joseph's Healthcare; Port Coquitlam, Canada: Forensic Psychiatric Services Commission.

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 violence risk structured professional judgment scheme. Behav. Sci. Law 24, 747–766. DOI:10.1002/bsl.737

Whittington, R., Bjørngaard, J. H., Brown, A., Nathan, R., Noblett, S., Quinn, B., 2014. Dynamic relationship between multiple START assessments and violent incidents over time: a prospective cohort study. BMC Psychiatry 26, 323. DOI:10.1186/s12888-014-0323-7.

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). Int. J. Forensic. Ment. Health 9, 282-293. DOI:10.1080/14999013.2010.534694.

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., Brink, J., 2013. Predictive validity of dynamic factors: assessing violence risk in forensic psychiatric inpatients. Law. Hum. Behav 37, 377-388. DOI:10.1037/lhb0000025

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

START: Effect of Diagnosis

Table 1: Participant characteristics

	Organic	Psychosis	Personality Disorder	Developmental Disorder	Psychosis and Personality Disorder	Psychosis and Developmental Disorder	Total
Diagnostic group *n* (%)	139 (26.4)	203 (38.5)	78 (14.8)	41 (7.8)	46 (8.7)	20 (3.8)	527 (100.0)
Age (years)[a]							
Mean (*SD*)	50.1 (16.5)***	40.7 (14.5)	31.4 (12.3)***	29.9 (11.3)***	36.1 (10.9)	37.7 (14.1)	40.4 (15.8)
Range	18-86	19-90	18-68	18-59	18-70	21-80	18-90
Gender *n* (%) [b]							
Male	114 (82.0)	172 (84.7)***	18 (23.1)***	36 (86.8)	29 (63.0)	19 (95.0)	388 (73.6)
Female	25 (18.0)	31 (15.3)	60 (76.9)	5 (13.2)	17 (37.0)	1 (5.0)	139 (26.4)
Ethnicity *n* (%) [c]							
Caucasian	91 (65.5)***	74 (36.5)***	61 (78.2)***	25 (61.0)	26 (56.5)	15 (75.0)	292 (55.4)
Asian/Asian British	1 (0.7)	19 (9.4)	0 (0.0)	3 (7.3)	0 (0.0)	1 (5.0)	24 (4.6)
Black/Black British	5 (3.6)	26 (12.8)	1 (1.3)	0 (0.0)	3 (6.5)	0 (0.0)	35 (6.6)
Mixed ethnicity	4 (2.9)	12 (5.9)	2 (2.6)	0 (0.0)	2 (4.3)	0 (0.0)	20 (3.8)
Other ethnic group	0 (0.0)	3 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (0.6)
No recorded ethnicity	38 (27.3)	69 (34.0)	14 (17.9)	13 (31.7)	15 (32.6)	4 (20.0)	153 (29.0)
Security Level *n* (%) [d]							
Open ward	0 (0.0)	7 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	2 (10.0)***	9 (1.7)
Low security	124 (89.2)***	141 (69.5)	46 (59.0)	28 (68.3)	25 (54.3)	10 (50.0)	374 (71.0)
Medium security	15 (11.8)	55 (27.1)	32 (41.0)***	13 (31.7)	21 (45.7)***	8 (40.0)	144 (27.3)
Mental Health Act *n* (%) [e]							
Informal	50 (36.0)***	6 (3.0)***	1 (1.3)***	0 (0.0)***	0 (0.0)***	0 (0.0)	57 (10.8)
Civil section	63 (45.3)	91 (44.8)	39 (50.0)	15 (36.6)	18 (39.1)	8 (40.0)	234 (44.4)
Forensic section	26 (18.7)	106 (52.2)	38 (48.7)	26 (63.4)	28 (60.9)	12 (60.0)	236 (44.8)
ICD-10 Substance misuse diagnosis *n* (%) [f]	8 (5.8)	34 (16.7)	9 (11.5)	1 (2.4)	8 (17.4	2 (10.0)	62 (11.8)
START							
Mean (*SD, range*) Strength total [g]	14.8 (6.8)**	17.6 (7.0)**	17.2 (7.9)	18.5 (7.9)	18.2 (8.9)	17.5 (7.3)	24.3 (7.7, 0-40)
Mean (*SD*, range Vulnerability total [h]	24.9 (6.7)	22.8 (7.3)***	27.9 (7.8)***	22.9 (8.5)***	23.8 (9.4)***	25.0 (5.6)	16.9 (7.4, 2-40)

[a] Age (defined as below and above the mean age of 40) (χ^2 [5, N=527]=64.03, $p<0.001$): those with an organic disorder were more likely to be older, whereas those with personality disorder or developmental disorder were more likely to be younger.

[b] Gender (χ^2 [5, N=527]=132.16, $p<0.001$): those with a psychotic illness were more likely to be male, whereas those with a personality disorder were more likely to be female

[c] Ethnicity (χ^2 [10, N=527]=71.83, $p<0.001$): Caucasians were more likely to have a diagnosis of organic disorder or personality disorder compared to other ethnic groups, but less likely to have a diagnosis of psychosis

[d] Security level (χ^2 [10, N=527]=54.47, $p<0.001$): those with an organic disorder were more likely to be detained in conditions of low security and less likely to be in conditions of medium security, those with psychosis and personality disorder, or personality disorder alone were more likely to be detained in conditions of medium security and those with psychosis and developmental disorder were more likely to be inpatient on an open ward.

[e] Mental Health Act status (χ^2 [10, N=527]=143.33, $p<0.001$): those with an organic disorder were more likely to be informal patients, whereas those who had a diagnosis of psychosis, personality disorder, developmental disorder and psychosis and personality disorder were less likely to be informal patients.

[f] Substance misuse (χ^2 [5, N=527]=14.60, $p<0.012$): those with an organic disorder were less likely to abuse substances, whereas those with psychosis were more likely to abuse substances.

[g] The psychosis group scored significantly higher than the organic group on mean pro-rated Strength score, F[5, 521]=3.37, $p=0.005$.

[h] The personality disorder group scored significantly higher than the psychosis group, the developmental disorder group and the psychosis and personality disorder group on mean pro-rated Vulnerability score, F[5, 521]=5.59, $p<0.001$.

Table 2: Base rates of risk outcomes by diagnostic group

Outcome	Organic	Psychosis	Personality Disorder	Developmental Disorder	Psychosis and Personality	Psychosis and Developmental Disorder	All
	n (%)	*n* (%)	*n* (%)	*n* (%)	Disorder *n* (%)	Disorder *n* (%)	*n* (%)
Verbal Aggression	61 (43.9)	71 (35.0)	29 (37.2)	15 (36.6)	11 (23.9)	5 (25.0)	192 (36.4)
Physical Aggression[a]	63 (45.3) ***	43 (21.2) ***	25 (32.1)	14 (34.1)	9 (19.6)	5 (25.0)	159 (30.2)
Self-harm/suicide[b]	7 (5.0)	12 (5.9)	22 (28.2) ***	6 (14.6)	10 (21.7) ***	0 (0.0)	57 (10.8)
Self-neglect	17 (12.2)	7 (3.4)	5 (6.4)	3 (7.3)	2 (4.3)	1 (5.0)	35 (6.6)
Victimization	20 (14.4)	36 (17.7)	18 (23.1)	5 (12.2)	1 (2.2)	2 (10.0)	82 (15.6)

[a] Physical aggression (χ^2 [5, *N*=527]=26.08, *p*<0.001): those with an organic disorder were most likely to engage in physical aggression; those with psychosis were least likely to engage in physical aggression

[b] Self-harm/suicide (χ^2 [5, *N*=527]=43.06, *p*<0.001): those with personality disorder or psychosis and personality disorder were most likely to engage in self-harm/suicidal acts; those with organic disorder or psychosis were least likely to engage in self-harm/suicide

Table 3: Predictive validity of the START for inpatient risk outcomes

Outcome:	Verbal Aggression		Physical Aggression		Self-harm / Suicidality		Self-neglect		Victimization	
	AUC [99% CI]	OR [99%CI]	AUC [99% CI]	OR [99% CI]	AUC [99% CI]	OR [99%CI]	AUC [99% CI]	OR [99%CI]	AUC [99% CI]	OR [99%CI]
Strength Total	0.69*** [0.63, 0.75]	1.11*** [1.07, 1.15]	0.68*** [.61, .74]	1.11*** [1.06, 1.15]	0.64*** [.55, .74]	1.09*** [1.02, 1.16]	0.74*** [.65, .82]	1.14*** [1.06,1.23]	0.62*** [.54, .71]	1.08*** [1.03, 1.13]
Vulnerability Total	0.65*** [0.59, 0.71]	1.09*** [1.05, 1.13]	0.65*** [.59, .72]	1.09*** [1.05, 1.13]	0.70*** [0.61, 0.79]	1.07** [1.01, 1.14]	0.75*** [0.66, 0.84]	1.15*** [1.06, 1.25]	0.61** [0.52, 0.69]	1.07*** [1.02, 1.12]
Violence SRE	0.62*** [0.55, 0.69]	ML 2.38*** [1.33, 4.24] HL 3.59*** [1.69, 7.67] HM 1.41 [0.65, 3.04]	0.68*** [.61, .75]	ML 3.65*** [1.96, 6.80] HL 5.60*** [2.54, 12.34] HM 1.49 [0.69, 3.25]	-	-	-	-	-	-
Self-harm SRE	-	-	-	-	0.81*** [0.72, 0.90]	ML 5.78*** [1.87, 17.91] HL 19.60*** [5.48, 70.06] HM 3.06 [0.96, 9.77]	-	-	-	-
Suicide SRE	-	-	-	-	0.73*** [0.61, 0.84]	ML 2.57 [0.87, 7.60] HL 19.93*** [3.54, 112.11] HM 7.56** [1.20, 47.67]	-	-	-	-
Self-neglect SRE	-	-	-	-	-	-	0.67** [0.53, 0.81]	ML 1.68 [0.50, 5.68] HL 6.61*** [1.84, 23.81] HM 4.72** [1.25, 17.80]	-	-
Victimization SRE	-	-	-	-	-	-	-	-	0.54 [0.44, 0.64]	ML 0.77 [0.31, 1.87] HL 3.24** [1.22, 8.58] HM 4.31** [1.22, 15.25]

**p<0.01; ***p<0.001; START Strength scores inverted for the purpose of this analysis; OR, odds ratio; CI, confidence interval; SRE, specific risk estimate; ML, SRE Moderate vs. SRE Low, HL, SRE High vs. SRE Low, HM, SRE High vs. SRE Moderate

Table 4: Predictive validity of the START Strength and Vulnerability scale scores and SREs by diagnosis

Outcome	Rocreg		Organic		Psychosis		Personality Disorder		Developmental Disorder		Psychosis and Personality Disorder		Psychosis and Developmental Disorder	
	Co-efficient	99%CI	AUC	99%CI	AUC	99%CI	AUC	99%CI	AUC	99%CI	AUC	99%CI	AUC	99%CI
Verbal Aggression														
Vulnerability Scale	0.09	[-0.04, 0.23]	0.62*	[0.53, 0.70]	0.64*	[0.57, 0.70]	0.67*	[0.59, 0.73]	0.71*	[0.60, 0.81]	0.73*	[0.58, 0.84]	0.75*	[0.57, 0.88]
Strength Scale	0.07	[-0.06, 0.22]	0.67*	[0.58, 0.74]	0.69*	[0.62, 0.75]	0.70*	[0.63, 0.77]	0.74*	[0.61, 0.84]	0.75*	[0.59, 0.87]	0.77*	[0.56, 0.90]
Violence SRE	-0.01	[-0.17, 0.11]	0.64*	[0.56, 0.75]	0.64*	[0.57, 0.72]	0.63*	[0.56, 0.70]	0.63*	[0.50, 0.73]	0.62	[0.47, 0.75]	0.62	[0.44, 0.77]
Physical Aggression														
Vulnerability Scale	-0.01	[-0.18, 0.14]	0.65*	[0.56, 0.74]	0.64*	[0.57, 0.71]	0.64*	[0.56, 0.71]	0.63	[0.49, 0.75]	0.63	[0.45, 0.77]	0.62	[0.41, 0.80]
Strength Scale	-0.03	[-0.17, 0.13]	0.69*	[0.61, 0.77]	0.68*	[0.61, 0.76]	0.67*	[0.60, 0.75]	0.66*	[0.53, 0.79]	0.65	[0.49, 0.81]	0.65	[0.44, 0.83]
Violence SRE	0.04	[-0.11, 0.20]	0.67*	[0.57, 0.76]	0.68*	[0.60, 0.76]	0.69*	[0.60, 0.76]	0.71*	[0.57, 0.83]	0.72*	[0.53, 0.85]	0.73*	[0.50, 0.88]
Self-harm and Suicide														
Vulnerability Scale	0.09	[-0.16, 0.37]	0.60	[0.42, 0.76]	0.62	[0.48, 0.75]	0.64*	[0.52, 0.75]	0.69*	[0.51, 0.84]	0.71	[0.48, 0.89]	0.73	[0.46, 0.93]
Strength Scale	0.11	[-0.18, 0.40]	0.62	[0.43, 0.81]	0.64*	[0.50, 0.78]	0.67*	[0.57, 0.78]	0.73*	[0.57, 0.87]	0.75*	[0.56, 0.92]	0.77*	[0.52, 0.95]
Self-harm SRE	0.06	[-0.32, 0.71]	0.76	[0.45, 0.96]	0.77*	[0.62, 0.93]	0.79*	[0.66, 0.90]	0.82*	[0.62, 0.98]	0.83*	[0.58, 10.00]	0.85*	[0.51, 10.00]
Suicide SRE	0.11	[-0.27, 0.49]	0.56	[0.28, 0.81]	0.60	[0.42, 0.78]	0.64	[0.48, 0.77]	0.71	[0.40, 0.91]	0.75	[0.34, 0.96]	0.78	[0.28, 0.98]
Self-neglect														
Vulnerability Scale	0.15	[-0.24, 0.56]	0.71*	[0.56, 0.83]	0.73*	[0.61, 0.83]	0.76*	[0.61, 0.87]	0.81*	[0.53, 0.96]	0.84	[0.46, 0.99]	0.86	[0.42, 0.99]
Strength Scale	0.16	[-0.13, 0.42]	0.69*	[0.57, 0.79]	0.72*	[0.61, 0.80]	0.75*	[0.63, 0.83]	0.80	[0.43, 0.90]	0.82	[0.34, 0.92]	0.84*	[0.51, 0.95]
Self-neglect SRE	0.33*	[0.06, 0.99]	0.54	[0.30, 0.74]	0.64	[0.42, 0.79]	0.73*	[0.57, 0.91]	0.87*	[0.68, 10.00]	0.92*	[0.70, 10.00]	0.95*	[0.72, 10.00]
Victimization														
Vulnerability Scale	-0.08	[-0.41, 0.20]	0.65	[0.49, 0.76]	0.63*	[0.53, 0.71]	0.61	[0.49, 0.72]	0.56	[0.27, 0.78]	0.54	[0.19, 0.81]	0.52	[0.13, 0.85]
Strength Scale	-0.03	[-0.31, 0.30]	0.65*	[0.52, 0.76]	0.64*	[0.54, 0.72]	0.63*	[0.53, 0.73]	0.62	[0.40, 0.84]	0.61	[0.33, 0.88]	0.60	[0.25, 0.91]
Victimization SRE	0.19	[-0.07, 0.59]	0.49	[0.35, 0.63]	0.54	[0.45, 0.65]	0.60	[0.46, 0.73]	0.70	[0.46, 0.92]	0.75	[0.43, 0.96]	0.79	[0.40, 0.99]

*$p<0.01$; START Strength scores inverted for the purpose of this analysis; SRE, specific risk estimate