Author’s response to reviews

Title: Susceptibility (risk and protective) factors for in-patient violence and self-harm: prospective study of structured professional judgement instruments START and SPROF, DUNDURM-3 and DUNDURM-4 in forensic mental health services.

Authors:

Zareena Abidin (zareenaabidin@yahoo.ie)
Mary Davoren (davorenm@tcd.ie)
Leena Naughton (leenanaughton@gmail.com)
Olivia Gibbons (olivia.gibbons@yahoo.com)
Andrea Nulty (andrea.nulty@hse.ie)
Harry G Kennedy (kennedh@tcd.ie)

Version: 2 Date: 12 July 2013

Author’s response to reviews: see over
My co-authors and I are grateful to the two reviewers for their helpful comments. We have carefully addressed each, point by point, as follows. We hope the revised manuscript is now acceptable. We look forward to any further comments and hope to be able to satisfy any further issues that might have escaped us. The revisions (responding to all 80 queries raised by the first referee) have unfortunately lengthened the manuscript but we have also edited extensively so that the main body of the article, at 7,400 words, is not much longer than the previous draft.

We hope the manuscript is now acceptable for publication. We are of course ready and willing to respond to any further points.

Please note that in the responses to the reviewers below, we have numbered extracts from the revised manuscript in accordance with the revised page numbers. In places, because of serial revision, these may no longer be correct – the (serially and multiply) revised manuscript is the best reference point.

Reviewer I

Reviewer: Michiel de Vries Robbé

Reviewer's report:

Review comments

Thank you for the opportunity to review this very interesting manuscript. It describes an important prospective clinical study including a brought range of tools. The manuscript addresses several fairly new risk assessment tools, especially those with a strengths-based focus rather than a risk-based focus. The authors should be congratulated for their important contribution to the knowledge about the psychometric properties of these new tools and about the value of the strengths-based risk assessment for clinical practice. If the authors attend sufficiently to the recommendations below I recommend publishing the article in BMC Psychiatry.

Major Revisions

General:

1. The main focus of the manuscript is on the protective/strength factor tools the SPROF and the START. However, attention is also focused on the
DUNDRUM-3 and 4 tools. Throughout the manuscript other tools are described: HCR-20, S-RAMM, PANSS, GAF, DUNDRUM-1. The comparison with all these tools is very interesting and a strong point of this manuscript, however the focus seems to switch here and there and the manuscript could be strengthened by keeping the main focus on the two / four key instruments while the others keep serving as comparison. In this respect the HCR-20 and the S-RAMM are the most valuable for this manuscript as they represent well established measures of risk of violence to others and risk of self-harm. The other tools (PANSS, GAF, DUNDRUM-1) are of less importance for the paper. However, the GAF adds a nice qualitative comparison and shows surprisingly good results. The PANSS only adds in terms of comparison. It is unclear what the DUNDRUM-1 adds to the paper and thus it is recommended to delete the DUNDRUM-1 all together from the manuscript.

We have now explained the use of the DUNDRUM-1 in this paper in greater detail –

“The DUNDRUM-1 triage security instrument is a measure of the need for therapeutic security and is designed to be a static measure of a quality that is complimentary to and distinct from risk of violence [26]. It has been shown to influence moves between levels of therapeutic security [24] and it is used also as a benchmark to enable comparisons between studies [25].” P4 lines1-4

And on page 14 –

“Contrary to expectations, The DUNDRUM-1 triage security score predicted violence (AUC=0.743, p=0.007, OR=1.238 95% CI 1.060-1.447) and though the AUC for the prediction of self harm did not reach significance, the odds ratio did (OR=1.226 95% CI 1.012-1.486).”

And on page 17 -

“Table 10 shows that three of the eleven DUNDRUM-1 triage security items predicted violence (TS4 ‘immediacy of risk of suicide or self harm’ OR=1.5, TS9 ‘complex risk of violence’ OR=3.3, TS10 ‘institutional behaviour’ OR=2.7) and 3 items predicted self-harm (TS2 ‘seriousness of self harm’ OR=1.4, TS4 ‘immediacy of risk of suicide / self harm’ OR=1.5, TS10 ‘institutional behaviour’ OR=2.7).”

At page 19 we have discussed this -
“The DUNDRUM toolkit instruments were not designed as risk assessment instruments, they were designed to be complimentary to risk assessments. The DUNDRUM-1 was included only as a benchmarking measure to enable future replication and meta-analysis. In spite of this, the DUNDRUM-1 triage security scale predicted violence and some of its items predicted violence and also self harm. This may be because items such as suicidal behaviour, complex needs and institutional behaviour are indicators of the seriousness of the behaviour that follows and such acts are easier to detect.”

2. Throughout the text there are very many inconsistencies regarding punctuation (font, periods, comma’s, spaces etc missing or in the wrong place). The entire text needs to be checked very carefully for this, to ensure all details are corrected. Please used APA.

Done

Background:

3. P3/P4: the description of the DUNDRUM tools could be more clear, perhaps a table would help? For example, why are 1, 3 and 4 for described in detail and not 2? It would be better to only describe the ones which are relevant for the current paper: 3 and 4.

The DUNDRUM-2 is used to prioritise those on waiting lists for admission to secure mental health services. It cannot be used for those already admitted. We would be happy to add in a sentence explaining this but it would be completely irrelevant to this paper. The handbook for the DUNDRUM toolkit is extensively referenced here and explains this fully. We have in response to the reviewer’s suggestion included an explanatory sentence regarding the use of the DUNDRUM1, DUNDRUM-3 and DUNDRUM-4 as follows –

“The DUNDRUM-1 was used to provide a benchmark for comparative purposes so that others replicating this study or carrying out meta-analyses can compare patients according to their assessed need for therapeutic security. The scale includes items for seriousness of violence and self harm, immediacy of risk of violence and self harm, specialist forensic need, absconding, preventing access to contraband, victim sensitivity and public confidence, complex risk of violence, institutional behaviour and legal process. The DUNDRUM-3 is a measure of programme completion in domains relevant to risk and harm reduction such as physical and mental health, substance misuse, problem behaviours, self-care and activities of daily living, education, occupation and creativity and family and social networks. The DUNDRUM-4 recovery items include stability, insight, therapeutic rapport, leave, dynamic risk and victim sensitivity.” Page 6

And in the interests of balance we have included a similar short summary of the SAPROF and START. We hope this answers the points raised by the referee.
“The START takes a list of risk factors and treats each one as both a risk factor and a protective factor. The SAPROF includes items thought to be protective against violence such as intelligence, secure attachment in childhood and empathy that are not included in existing risk assessment instruments.” Page 6

4. P5, line 4: Conceptually it would make sense to use the S-RAMM as ‘golden standard’ for self-harm risk assessment. Please describe why the HCR was chosen in stead.

We have re-worded this sentence as follows –

“We compared these to existing validated instruments for the assessment of risk of violence (HCR-20) and self-harm (S-RAMM) and examined whether they accounted for any element of statistical prediction over and above an existing ‘gold standard’ instrument for the assessment of risk of violence, the HCR-20. We compared these to existing validated instruments for the assessment of risk of violence (HCR-20) and self-harm (S-RAMM) and examined whether they accounted for any element of statistical prediction over and above an existing ‘gold standard’ instrument for the assessment of risk of violence, the HCR-20.” Page 5

We have previously shown that the HCR-20 predicts both violence and self harm, while the S-RAMM also predicts self harm and violence [34]. It is more economical to use one.

Method:

5. P5: Bottom 2 lines: ‘...so that the location... patient is located [24].’ Is irrelevant as security level is not used as a variable, can be deleted.

Security level (like length of stay) is used as a variable to check for possible confounding for the reasons demonstrated in [24] and also in [34] –

“The location at baseline (for eight locations from the most to least secure) predicted harm to others (AUC=0.812, 95% confidence interval 0.677 to 0.948, p<0.001) as expected, since we have previously shown that location is a proxy for measures of risk [33] and recovery [23,24]. Length of stay at the beginning of the observation period did not predict harm to others (AUC=0.504, 95% CI 0.343-0.665, p=0.963). Location at baseline also predicted self-harm (AUC=0.838, 95% CI 0.689-0.987, p=0.003) while length of stay did not predict self harm or the absence of it (AUC=0.578, 95% CI=0.383-0.722, p=0.495).” pages 13-14

We believe it is important to leave this in, since it is again relevant to the ability of others to replicate this and similar studies. We would however remove it if this is thought to be unhelpful.
6. P6, line 5: it should be mentioned that the START has a strength and a vulnerability scale which are coded simultaneously, as you are addressing this on page 7.

Agreed –

“The START takes a list of risk factors and treats each one as both a risk factor and a protective factor. “ page 6

7. P6: in general an explanation needs to be added of the START, the SAPROF and the DUNDRUM tools. As these are the new instruments they need to be addressed in more detail, how many items / scales etc.

Agreed, see above responses to points 3, 4 & 6

8. P6, line 19: what does the DUNDRUM-1 add to the paper? If nothing, it should be deleted.

See above, response to points 1 & 3. We believe the DUNDRUM-1 is methodologically essential.

9. P6, line 23: how can IRR be assessed if only one rater rated all files? Please provide more details.

We agree that the inter-rater reliability data is not necessary but as there are so few independent validations of this form of reliability we thought it would be helpful to do it and publish it.

10. P8: so the sample size was actually 98, not 100?

Yes, 100 eligible, 98 assessed.

11. P8: what was the mean follow-up time for the 6 discharged patients?

The mean follow-up time for those who were discharged before they could be fully assessed was 21.7 days.

12. P8, line 20 and line 25: location in analyses? Seems irrelevant to paper, suggest to delete this from the paper. Here: lines 18-22 and lines 24 (from ‘Means’) and 25.

Regarding location, see above. There remainder of this comment is unclear – we would be happy to address if clarified.
13. P9, line 14: please describe analysis in more detail

We apologise that we are unable to respond to this suggestion because of its vagueness. We would be happy to respond to it if clarified.

14. P9: what analysis was used for the interaction?

Univariate analysis of variance, reporting main effects and interactive effects, see page 13-14

“Interactive effects between risk factors and protective factors

Univariate analysis of variance was used to test for the presence of interactive effects between risk measures and protective measures. Since the SAPROF and START-S are dynamic measures, the HCR-20-dynamic score was taken as the ‘standard’ risk measure.

Tables 3 and 4 show that although the SAPROF, START-S, DUNDRUM-3 and DUNDRUM-4 were all significantly different for the 13 patients who were violent when compared to the non-violent and likewise for the 7 who self harmed, when these scores were adjusted for the HCR-20 dynamic score the differences were no longer significant.

For harm to others, the SAPROF and HCR-20-dynamic had significant main effects (HCR-20-dynamic F=3.97, df=17, p=0.003, SAPROF F=4.67, df=25, p<0.001) and a significant interaction effect (F=2.973, df=38, p=0.008) indicating that the SAPROF had a ‘true’ protective effect. The SRAMM-dynamic score also had a significant interaction with the HCR-20-dynamic score (HCR-20-dynamic F=3.828, df=18, p=0.001, SRAMM-dynamic F=3.909, df=20, p<0.001, interaction F=2.794, df=33, p=0.003) apparently indicating a synergistic effect. The START-S, DUNDRUM-1, DUNDRUM-3, DUNDRUM-4, PANNS positive, PANSS negative, PANSS general and GAF scores did not have significant interactions with the HCR-20-dynamic score.

The SAPROF did not have significant interactive effects with the START-S, DUNDRUM-1, DUNDRUM-3, DUNDRUM-4, SRAMM-dynamic, PANSS-positive, PANSS-negative or PANSS-general scores, but had a marginal interactive effect with the GAF (main effects SAPTOF F=4.05, df=25, p<0.001, GAF F=5.78, df=21, p<0.001, interactive effect F=1.98, df=33, p=0.059). “ pages 13-14

15. P9, line 17, table 3 and 4 mention HCR-20 total scores, here it states C+R only.

The text is correct and the labelling of the tables has now been up-dated to reflect this. The mislabelling was left over from an earlier draft.

Results:

16. Throughout: please pay attention to punctuation details, missing spaces,
comma’s etc.

The distinguished referee has confused the use of apostrophes in the pleural and the possessive. Please note that the pleural does not take an apostrophe in English – “comma’s” above is incorrect and should be rendered “commas”. See Lynne Truss ‘Eats, Shoots and Leaves: the Zero Tolerance Approach to Punctuation. London, Profile Books 2003.

17. P10 line 21 (Similarly...) through P11 line 5: the internal consistency of the well-know measures is less important and takes away from the main focus, I’d suggest deleting all this and just mentioning START, SAPROF and DUNDRUM-3 and DUNDRUM-4.

While we agree that the internal consistency of well-established instruments is well known, we believe it is correct to report our findings in this data set because the internal consistency of these instruments will be important in a subsequent paper. Also, the S-RAMM did not perform well in this study and had lower than usual internal consistency, suggesting some problem for future exploration. It also demonstrates the multiple co-linearity which had to be taken into account when deciding on regression analysis, since Cronbach’s alpha statistic is derived from a correlation matrix. We have now mentioned this under statistical methods, page 10 –

“High internal consistency also indicates multiple co-linearity for items within a scale.”

18. P11-P12: the different sections on concurrent validity should be combined and shortened, all details are also show in table 1 (table and text are redundant), please just highlight the main findings. Why is concurrent validity with DUNDURM-3 and 4 not mentioned? (The section on GAF has no p-values, the section on PANSS has no statistics at all. Perhaps referring to the table is better here for all sections, but it needs to be consistent).

Agreed and done.

19. P13 first paragraph on location and length of stay: this section takes away from the main focus of the paper and should be deleted.

We believe it is an essential observation to record that length of stay was not a confounder. It is a related point that location was a significant predictor. This demonstrates that patients moved though the hospital’s levels of care according to risk-need-responsivity and not simply according to some form of sentence tariff. We decided not to use location as a co-variate in this study because it is merely a secondary or proxy covariate for risk. We believe the referee would find the same in the hospital where he works and hope he and his colleagues will one day publish on this. The
inclusion of length of stay and location data is therefore a methodologically necessary point for future replication studies and meta-analysis. We have now made this clear in the discussion -

“...predicted harm to others (AUC=0.812, 95% confidence interval 0.677 to 0.948, p<0.001) as expected, since we have previously shown that location is a proxy for measures of risk [33] and recovery [23,24]. Length of stay at the beginning of the observation period did not predict harm to others (AUC=0.504, 95% CI 0.343-0.665, p=0.963). Location at baseline also predicted self-harm (AUC=0.838, 95% CI 0.689-0.987, p=0.003) while length of stay did not predict self harm or the absence of it (AUC=0.578, 95% CI=0.383-0.722, p=0.495).” page 14

And in ‘limitations’ -

“...determined by risk and need for therapeutic security, not by a simple chronological waiting list or tariff for movement from more secure to less secure locations. The location was therefore a proxy for risk and an indication that placement and milieu were appropriate to individual need, to manage and reduce risk [40-42].” Pages 22-23

20. P13-P14: CI’s can be read in table, don’t need to be repeated in text.

DONE

21. P13: HCR and S-RAMM: add AUC’s, subscales don’t have to be mentioned in text.

DONE

22. Actually, if OR’s are mentioned in text, it should be explained what this means for different scales, that is an OR on a scale with 5 items is not comparable to an OR on a scale of 20 items as this simply reflects the increase in odds if the score on the scale was increased by 1 unit. It would be helpful to explain what an increase of let’s say 10 points on each tool (total scores) would mean for the increase in violence likelihood. But for this to make sense to the reader, the tools need to be explained in more detail in the Method section. Or just mention AUC values in text and leave OR’s for table.

We are grateful for this valuable suggestion. We have now added the following explanatory sentence in ‘Statistical Methods’ -
“The odds ratio for a scale such as the GAF which is rated 0 to 100 will be inherently smaller when comparing like for like with the HCR-20, a scale rated 0 to 20. Likewise, the odds ratio for items rated 0 to 1 as in the HCR-20, S-RAMM, SAPROF and START will appear larger when comparing like for like with items from the DUNDRUM-1, DUNDRUM-3 and DUNDRUM-4 where items are rated 0 to 4. “ Page 10

And we have further amended the earlier section describing each scale used to include the item and scale ratings. This is set out in the ‘objectives’ sessions as a null hypothesis -

“We also examined the predictive properties of measures of symptoms (PANSS [29,30]) and global function (GAF [31]) treating these as another standard to be beaten – are specific risk assessment instruments and their constituent items better than assessing symptoms and function?” Page 5

- the articles referenced support the proposition that symptoms underpin risk in the mentally ill.

23. P14: DUNDRUM-1 can be taken out. The PANSS as well? Or just mentioned very briefly.

See above regarding the necessity of retaining the DUNDRUM-1. The PANSS is also included because like the GAF, it represents a ‘null hypothesis’ – are risk assessment instruments better than this? They proved to be better so that the value of the risk assessment instruments is demonstrated. This is set out in the ‘objectives’ sessions as a null hypothesis -

“We also examined the predictive properties of measures of symptoms (PANSS [29,30]) and global function (GAF [31]) treating these as another standard to be beaten – are specific risk assessment instruments and their constituent items better than assessing symptoms and function?” Page 5

- the articles referenced support the proposition that symptoms underpin risk in the mentally ill.

This is set out in the ‘objectives’ sessions as a null hypothesis -

“We also examined the predictive properties of measures of symptoms (PANSS [29,30]) and global function (GAF [31]) treating these as another standard to be beaten – are specific risk assessment instruments and their constituent items better than assessing symptoms and function?” Page 5

- the articles referenced support the proposition that symptoms underpin risk in the mentally ill.

24. P15, 3rd paragraph: what analysis was used for the interaction?

We explain this in the ‘statistical methods’ section –

“To examine the extent to which the protective instruments SAPROF, START-S, and recovery instruments DUNDRUM-3 and DUNDRUM-4 were protective in the presence of risk factors, we first carried out an analysis of variance in SPSS-18 with SAPROF, START-S, DUNDRUM-3 and DUNDRUM-4 as dependent variables, violence to others or self-harm as fixed factors (in separate analyses) and the HCR-20-dynamic score as covariate. To
examine for interactive effects, we then carried out univariate analysis of variance to examine for with main effects and interactive effects.” Page 11

The actual models are discussed in detail in the results section.

25. P15, 3rd paragraph: interaction HCR and S-RAMM not relevant for paper

We respectfully believe this significant interactive effect is relevant. The paper is about the validation of ‘protective’ risk assessment instruments and interactive effects. A synergistic effect between measures of risk of violence and risk of self-harm is relevant to the declared objectives of the study and is of great interest. When these results were presented in preliminary form to the Forensic Psychiatry Research society in Oxford earlier this year this point was commented on by the audience members. We have now drawn attention to this in the results section

“The S-RAMM, an assessment of risk of suicide and self-harm, was a predictor of violence and the S-RAMM dynamic score had a synergistic interaction with the HCR-20-dynamic score. The S-RAMM, an assessment of risk of suicide and self-harm, was a predictor of violence and the S-RAMM dynamic score had a synergistic interaction with the HCR-20-dynamic score.” Page 22

26. P15, bottom lines: SAPROF x GAF interaction is not significant.

We described this as marginal, not significant. We believe it is necessary to report this. The interaction could not be described as absent or negligible -

“The SAPROF did not have significant interactive effects with the START-S, DUNDRUM-1, DUNDRUM-3, DUNDRUM-4, SRAMM-dynamic, PANSS-positive, PANSS-negative or PANSS-general scores, but had a marginal interactive effect with the GAF (main effects SAPTOF F=4.05, df=25, p<0.001, GAF F=5.78, df=21, p<0.001, interactive effect F=1.98, df=33, p=0.059).” page 17

It is worth reporting because both SAPROF and GAF are shown to be protective and both are strongly correlated. The relationship between the two is therefore of interest – is there a true synergy or even a true antagonism – we hope to return to this in future work, and others may wish to study it too.

27. Basically, apart from the first 3 lines on SAPROF x HCR-dyn, the rest of the text can be said in one sentence: ‘No other interaction effects were observed.’

As above

28. P16-17: the item analysis for all tools seems to carry too far for the current paper. I would suggest only discussing the SAPROF, START, DUNDRUM-3 and DUNDRUM-4, mentioning how many significantly predicting items each tool has
and which items were good predictors. All on HCR, S-RAMM, PANSS and DUNDRUM-1 should be taken out. Also the tables are probably too elaborate and should be deleted?

We believe there are important advantages in this form of presentation for the reasons already set out. The measures of symptom severity (PANSS) and seriousness of risk (DUNDRUM-1), like the S-RAMM and HCR-20 are used as ‘control’ measures – the research question is ‘can the START, SAPROF and DUNDRUM-3 and DUNDRUM-4 do better’? The reasons why they are valid comparators is set out above and in literature references [29-31]. We have now clarified this in the ‘objectives’ section –

“Objectives

In this prospective study we set out to assess psychometric properties, concurrent validity and criterion outcome measures of the validity of the START and SAPROF. We prospectively tested whether START and SAPROF, DUNDRUM-3 and DUNDRUM-4 predict adverse events (or the absence of adverse events), violence or self harm. We compared these to existing validated instruments for the assessment of risk of violence (HCR-20) and self-harm (S-RAMM) and examined whether they accounted for any element of statistical prediction over and above an existing ‘gold standard’ instrument for the assessment of risk of violence, the HCR-20. We also examined the predictive properties of measures of symptoms (PANSS [29,30]) and global function (GAF [31]) treating these as another standard to be beaten – are specific risk assessment instruments and their constituent items better than assessing symptoms and function?” pages 3-4

While the tables are lengthy, they meet modern standards for such validation, giving both the ROC-AUC and the odds ratios so that strength of association of individual items can be compared within each instrument. We believe strongly that the ability to present tabular data in this way rather than restricted to subjective summaries is an important advantage of internet only journals such as BMC Psychiatry.

Discussion:

We have now added a first paragraph which summarises the results in the light of the ‘objectives’ paragraph earlier in the article –

“This paper presents validation studies for ‘fourth generation’ risk assessment instruments. We have examined the utility of these instruments for assessing risk and protective factors for both violence and self-harm. We have identified both overlaps and differences in the risk factors that contribute to predictions of risk of violence and self-harm. We have included some methodological approaches intended to facilitate future researchers who might replicate this work or include it in meta-analyses. These include stating the base rates for violence and self-harm and giving the DUNDRUM-1 triage security ratings as a means of benchmarking the background need for therapeutic security.
We believe the most important finding is confirmation that true protective effects can be identified. The SAPROF, a protective scale does more than assess the absence of risk – the SAPROF also had an interactive effect with the HCR-20, offsetting risk. “ page 20

29. P19/20: Bottom lines P19 and first paragraph P20: this paragraph seems a little out of place, could be revised.

This comment refers to a paragraph concerning self-harm and instruments such as the S-RAMM.

“The S-RAMM, an assessment of risk of suicide and self-harm, was a predictor of violence and the S-RAMM dynamic score had a synergistic interaction with the HCR-20-dynamic score. The GAF and PANSS scales (other than the PANSS negative score) also performed well as predictors of violence. Although designed as assessments of risk and protective factors for violence, most scales were also predictive of self-harm. The S-RAMM-C scale performed well but the S-RAMM-B (background or fixed historical risk factors for suicide), S-RAMM-F (‘future’ risk factors for suicide), START-S and START-V were notable for their lack of predictive capability for self harm in this study.” Now Pages 21-22

For the reasons outlined above, we have included this because it is a comparator for the SAPROF and START as assessments of risk of self-harm – a claim advanced by the authors of the START in the handbook for that instrument. This is made clear in the following paragraph –

30. The following two paragraphs are also somewhat unclear and the focus on individual items too detailed (see comment on individual items later on). Perhaps summarize these paragraphs and just make the point that its (Does the referee mean ‘it’s’ or ‘its’? the sentence can be read with two different meanings depending on which was intended. We assume the referee meant ‘it’s’, the abbreviation of ‘it is’) different factors which are good predictors for violence and for self-harm.

We have now clarified the two paragrpahs and we are grateful for the suggested need for this –

“The overlap between risk factors for violence and self-harm, and the need to assess both has been established [38,39,33,34]. An item analysis shows considerable overlap of the content of each of the scales examined here. Of these, the DUNDRUM-3 programme completion items appeared particularly strong predictors of self-harm or the absence of it, perhaps because of an underlying element of positive motivation that is inherent in the way each item is defined and rated…………

“For violence, only some items in each scale were predictive. The highest AUC results were obtained for lack of progress in treatment programmes such as education, occupation and creativity, a low GAF score, conduct problems, lack of progress in mental health programmes, impulse control, adverse institutional behaviour, leisure activities, external
triggers, negative attitudes, poor attention, financial problems, hyperactivity and self-control, relationship problems, stability, empathy and hostility.

For self-harm a different selection of items predicted adverse events with the highest AUC results for the GAF, poor attention, conceptual disorganisation, lack of progress in mental health programmes, disturbance of volition, unresponsiveness to treatment, adverse institutional behaviour, preoccupation, leisure activities, hyperactivity, tension, problem solving deficits, stability, self-control and negative attitudes." Page 22

31. The focus of the discussion should be more on summarizing the findings from the present study and comparing these to findings in other studies. Please make more reference to (findings in) the literature. For example, it would be interesting to note that the some tools intent to measure violence to others as well as self-harm, while others only focus on either one. This can be related to the current findings. Also, comparison can be made regarding the short follow-up time and inpatient aggression as outcome, in the light of findings in other studies with the tools.

We have now explained this point in greater detail, as follows –

“……Of greatest relevance is that most scale scores were predictors of both violence and self-harm, though this was often because of different items within each scale. Much of this appears to be contextual. In a group made up predominantly of forensic patients admitted to a forensic hospital because of severe mental illness and violence, it is not surprising that items such as the first item of the HCR-20 ‘past violence’ should be poor predictors of further violence in a group where all score positive. It is important to note that items such as the first S-RAMM item ‘past self-harm’ are such good predictors of violence to others, though not self-harm, while items such as HCR-20 C5 ‘unresponsiveness to treatment’ and S-RAMM item C3 ‘psychological symptoms’ and C8 ‘problem solving deficits’ predicted both harm to others and self-harm. Page 22

32. P21 2nd paragraph, first line: many items in HCR appeared not predictive?

HCR-20 AUC total score was .87/.88! (see table 2), which can hardly be described as ‘not predictive’. See comment lateron (Tables) the focus on individual items is less strong given the relatively small sample size and low base-rate. I’d suggest focusing more on the scales in general. Last two lines on
We are grateful for this observation by the referee - we believe this is exactly why it is important to give the results for items as well as scales. In a scale with strong internal consistency (multiple co-linearity) many weakly predictive items will yield a scale score that is predictive, but the items are a poor guide to treatment need. This is contrary to the declared purpose of structured professional judgement and raises the need for revisions of the scales as evidence such as this (based on many small studies) accumulates. And we do not know of many studies with more than 94 subjects. We have now dealt with this in the discussion section –

“A notable feature emerges when unadjusted odds ratios are examined, comparing AUC and odds ratios. Scales and items with significant AUC statistics may have better than random sensitivity and specificity, but may still be weak predictors. While this reflects the reality of multiple co-linearity, any argument that a risk assessment scale made up of just a few of the strongest items would be sufficient is at odds with the clinical need to take notice of a much wider range of risk and protective factors when planning care and treatment [7,8] and when making recommendations or decisions regarding discharge [25]. However it is also the case that using structured professional judgement instruments to assess treatment needs would be invalid if many of the items were poor predictors on their own. We believe the poor performance of many scale items should lead to two forms of revision of these scales. The first would be to specify that some items are useful only for specific contexts – such as in-patient setting, outpatient community placements or prisons. The second would be to drop some items or refine the handbook definitions.”

We hope this is acceptable.

33. P22 ‘Finally...’: longer observation period would also have been informative!

We have now amended this paragraph as follows –

“Finally, this prospective study covered a six month observation period, though a shorter observation period may have been more meaningful, at least for symptom ratings. A longer observation period might have improved statistical power with more adverse events emerging, but a longer observation period would also raise the problem of the validity of ratings of dynamic risk items which might have changed in the interim. Much remains to be learned about the time course over which dynamic and some static risk items might change.” Pages 25-26

We hope this is acceptable.

34. The fact that different tools were rated by different raters (see Authors’ contributions) might have an impact on the findings, this should be addressed in the limitations. Also, the fact that only one rater rated all files for every tool may have influenced the findings.
The methods section sets out clearly that the instruments were variously rated by a single research clinician or else collated by a single researcher (see below). This has the important advantage that it eliminates inter-rater reliability problems which in our experience are substantial outside of research settings. Each researcher was blind to the ratings of others – an essential methodological check against bias. The ratings were made just before the period of observation for outcomes so that the ratings were blind to outcome. Each rater was supervised by the lead investigator to ensure fidelity to the various handbooks. We do not see how having a single rater for a scale, blind to others, could have an adverse effect?

35. The HCR-20 and S-RAMM were collated from team assessments, while the other tools were rated by a single rater. Team assessments have shown to have better predictive validity (e.g. De Vogel et al), this too needs to be addressed in the limitations.

We have been unable to confirm the findings of De Vogel et al that team assessments have better predictive validity than single raters. In the naturalistic setting, we have found that the extent of inter-rater variation between different teams, differently constituted at different times can largely invalidate predictive power despite regular retraining. We have found in a series of studies that having a single experienced clinician collate team assessments for consistency, mentoring the teams, is necessary to maintain quality and consistency. We will report this elsewhere.

36. P24. First few lines, more findings have been published on the START in recent years, which should be made reference to.

37. P24. Line 3: ‘De’ should be added: ‘De Vries Robbé’. Here too, reference should be made to previous findings of good predictive validity which are very much in line with the current paper.

Done

Reference list:

38. Please mind punctuation and consistency. Almost all references have something incorrect: missing spaces, bold/italic, positioning etc. Please check very carefully for this.

Done

39. Ref. 44: Author order incorrect; switch ‘De Vries Robbé’ and ‘De Vogel’
Tables:

40. Although most of the information in the tables is informative, the 37 (!) pages of tables seem quite extravagant for the manuscript. I’ll leave it up to the editors to comment on how many pages are acceptable in an online supplement and whether extensive tables like these would be desirable.

*We are happy to abide by the views of the editor. As set out above, we believe there are advantages in publishing this information and that on-line journals are best suited to this.*

41. All tables, please use APA style: no vertical lines, proper headers etc.

*Done as set out by BMC Psychiatry style*

42. Pages and headers have varying fonts. Please be consistent in font, header lay-out etc.

*Done as per journal style*

43. Tables 3 and 4: these concern HCR-dynamic scores, not total scores?!

*Corrected, see above*

44. Table 3: DUNDRUM-3: higher (in stead of lower) total programme completion scores for Violent group after controlling for HCR?

*What is important is that the difference between violent and non-violent was no longer significant. This is explained in the text.*

45. Table 4: START-S: higher (in stead of lower) total strength scores for the Self harming group after controlling for HCR? -> should be mentioned in discussion.

*See previous paragraph.*

46. Tables 5 through 11: although informative to authors of these tools, it seems it might be better to delete these tables from the manuscript. Beside the fact that this is far too much information for one manuscript, the item AUC analyses of this many individual items in a relatively small sample and low recidivism base-rate is not recommended. It might be best to mention the best predictors of the main tools (START, SAPROF, DUNDRUM-3 and -4) in the text in stead and perhaps
the number of significant predicting items per tool.

This has been done. The base rate for adverse outcomes violence and self-harm is made clear in the results section -

“There are thirteen individuals had adverse incidents concerning harm to others (broadly defined, as above) during the follow up period and 7 individuals had incidents involving self harm (broadly defined, as above). There was a significant overlap between self-harm and harm to others (X²=35.2, df=1, p<0.001, phi = 0.593, p<0.001). The rate of events of harm to others was 7.1 per 10,000 patient-days at risk (95% confidence interval 3.8 to 12.2/10,000) and the rate of self-harming events was 3.8 per 10,000 patient-days at risk (95% CI 1.5 to 7.9/10,000).”

We believe this should be essential whenever such studies are reported, to enable readers to form their own judgement concerning the base rate and along with the odds ratios, to enable other researchers to carry out metanalysis.

Minor essential revisions

Title:

47. The title is quite long and could benefit from more focus. For example:

‘Risk and protective factors for inpatient violence and self-harm: A prospective study of the START, SAPROF, DUNDRUM-3 and DUNDRUM-4’

While we agree that this might be in keeping with other publications, the referee may be overlooking the discussion of the importance of susceptibility factors –

“Rutter [27] pointed out that a protective or resilience factor should do more than simply predict the absence of harm or adverse outcomes, since this is merely the absence of risk. Risk or vulnerability factors (or their reciprocals, measuring the absence of risk) and protective or resilience factors can be validated as predictive or not using the receiver operating characteristic, as a means of taking into account base rate variations between samples [28]. The strength of association in the population and setting studied can be assessed with unadjusted odds ratios. A truly protective factor however would interact with risk factors to reduce the probability of an adverse event or outcome, even when risk factors were present [27]. This requires a form of analysis of interactive effects additional to that normally used to validate risk factors. “ page 4

And -

“According to Rutter [27] a proper analysis of protective effects or resilience would have to involve examining for the effects of protective factors and the interactions they might have not just with risk factors but also with adverse events and difficulties. We found some evidence for an interactive effect between the SAPROF and the HCR-20 dynamic score” page 25
We intend to return to this topic in our next paper

Background:

48. P3 2nd paragraph last line: ‘..resilience factors that reduce the risk of violence and self-harm.’

Done

Method:

49. P6, line 9: Historical-Clinical-Risk-20 should be: ‘Historical, Clinical, Risk management-20’, but actually other tool names are not spelled out either, so just change to ‘HCR-20’ and spell out when first mentioned earlier on. The same goes for the S-RAMM.

Done

50. P6, line 10: R stands for ‘Risk management’

Not in general use

51. P6, line 13: ‘Similarly’ should be ‘Similar to the HCR-20’

noted

52. P6, line 16: ‘risks’ should be ‘risk of suicide’

noted

53. P6, line 17: ‘These’ should be ‘the HCR-20 and the S-RAMM..’

Done

54. P6, line 24: ‘individual risk or protective factors’

Noted

55. P7, line 1: ‘S-RAMM’

done

56. P7, Outcome measures, line 4: ‘.. between violence to others and..’

This is not clear?

57. P8, line 1: bracket wrong place

done
58. P8, line 4: the header ‘Bias’ should be changed into ‘Study size’ (and that header should be deleted below)

Done

59. P9, line 6: not a new paragraph

Done

Results:

60. P10, line 3: bracket

Done

61. P10, line 10: add range of follow-up. Number of patient days at risk is not very informative.

The number of patient-days at risk is necessary to calculate the base rate of adverse events, see above.

62. P11 line 2: HCR, C = Clinical, R = Risk management

This is not the customary way to describe the ‘R’ items in published literature and the use of the term ‘risk management’ in the HCR-20 manual is debatable, hence it is not followed generally.

63. P11, lines 10-12: The sentence ‘If the....scales.’ should be deleted (is repeated in line 14/15).

It would not be helpful to delete the first occurrence of this essential explanation. In the interests of clarity we believe it is necessary to leave both.

64. P12 last 3 lines: ‘the rate of events... 10,000).’ Seems irrelevant, can be deleted.

We respectfully disagree. It is essential to give the base rate for the reasons stated above. The previously published validation studies of many of the instruments studied here are weakened by not having stated the base rate. This paper advances the methodology of validation.

65. P15, bottom lines: ‘SAPROF’

Done

Figures:

66. Page 32 is empty?

Fixed
Discretionary revisions

Background:

67. P2, 2nd paragraph: If the length of the manuscript is too great, the description of the RA tools history can be summarized in a few lines.

Note

68. P3, last paragraph: this section describes some previous results, while the same is not done for the START and SAPROF (these are mentioned in the discussion in stead), perhaps it would be better here to just explain what the DUNDRUM tools are and what they intend to measure.

Note

Method:

69. P7, line 4-5: move to Variables section P6

Done

70. P8, line 17: ‘the line of zero information’ -> ‘chance’

All probability is ‘chance’. Random probability is often expressed as p=0.5. The line of random (or zero) information corresponds to an AUC = 0.5. The line of random/zero information is a term specific to the receiver operating characteristic which plots sensitivity (true positive rate) against one minus specificity (specificity is the true negative rate) – this is not the same as chance though we agree that chance is an approximation of what it means.

71. P9, lines 18-24: merge with top page 9 or delete altogether if the item analyses are deleted from the paper.

Note

72. P9, bottom two lines: ‘Because... attempted.’ is irrelevant and can be deleted.

Note. We believe it is useful to leave this very short explanation in, since other regression models are presented and a curious reader might wonder why this was not attempted for items within scales.

Results:

73. P14, 3rd paragraph bottom line: ‘Lower... factors.’ Should be deleted here or moved to method.
74. P15, 2nd paragraph: New sentence ‘Likewise for the 7 who self-harmed compared to those who did not. However, when…’

Done

75. P16, line 3: ‘Eleven’... the item Life goals had a p of .050, was this rounded up or down? (if up, this makes 12).

This is an exact result. We have changed the sentence to ‘twelve’ though an AUC of 0.727 is marginal and the odds ratio was not significant.

Discussion:

76. Limitations: Omitting HCR-20 H7 (Psychopathy) and the 6% females in the sample could also be mentioned.

Done.

77. P22 Future studies: A recommendation could be made for studies including repeated measures in order to be able to assess the influence of treatment change as measured by the different tools on reductions in violence risk.

The referee has already published such a study concerning the SAPROF, listed as ‘submitted’ in his 2012 review in ‘Psychology’. We look forward to seeing it. We have such a study under way at present looking at a range of instruments. We have now added the following -

“Serial assessments to demonstrate change and assess the effect of change would be of interest.” Page 26

78. P23. 2nd paragraph (about S-RAMM B1) is not relevant enough for the discussion, suggest to delete.

It is worth quoting the next sentence, which we believe justifies the sentence queried – we have added a further explanatory sentence.

“The S-RAMM item B1 “history of deliberate self harm” predicted harm to others as well as harm to self (AUC 95% CI greater than 0.5) in this population. In contrast, the HCR-20 H1 item “previous violence” was a poor discriminator in this forensic population, because almost all patients scored positive. This may be an unexpected benefit of using diverse assessments and may also be understandable in psychological terms.” Page 26

79. P23. 3rd paragraph what does ‘these instruments’ refer to, the instruments in
the 1st paragraph? I’d suggest merging paragraph 1 and 3 (and deleting 2).

We have now divided the paragraphs differently.

80. P25. Last line could be deleted/changed: ‘Many...outcomes.’

Done

Level of interest:

An article of importance in its field

Quality of written English:

Needs some language corrections before being published

Statistical review:

No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I am co-author of the SAPROF, one of the tools that is the focus of this manuscript. I do not receive any direct financial benefits from this tool, however the hospital I work for holds the patent to the SAPROF.
Reviewer's report II

Title:
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.
Version:
1
Date:
21 June 2013
Reviewer:
Danny Sullivan

Reviewer's report:

This is a well-constructed study which investigates a range of assessment tools which might be used in a secure forensic psychiatric hospital. The research questions are clearly-defined and seek to detail the elements of these instruments which might be effective in predicting various adverse outcomes. The methods are described clearly. They are appropriate to investigate the research questions. The analysis explores a range of psychometric properties of the instruments which are relevant to establishing a basis for their use in this population, and investigates their usefulness to predict events of clinical relevance. The data appear sound and are set out comprehensively. Discussion of the study and its findings elaborates clearly its limitations and the potential utility of the instruments. The writing is clear and strikes a balance between the necessity of explaining the methods of analysing the tools, and plain language discussion of the practical relevance of the findings. This study will be of interest to those tasked with the clinical governance of secure psychiatric facilities, and further adds to study of the properties of a number of tools which are increasingly used in contemporary risk assessment and management planning.

Minor Essential Revisions
Minor typographical errors are noted:
In the penultimate paragraph of the Background section, complimentary should read complementary; and once more in the first paragraph of the Discussion section.

Done

Under the Results section, subheading internal consistency, lead should read led.

Done

Discretionary Revisions
The Background section could do with minor rewriting to flow more clearly into
the research question.

 Done