Reporting Guidance for Violence Risk Assessment Predictive Validity Studies: The RAGEE Statement

Jay P. Singh
Singh Institute of Forensic Research and
Universität Konstanz

Suzanne Yang
Veterans Affairs Pittsburgh Healthcare System and
University of Pittsburgh School of Medicine

Edward P. Mulvey
University of Pittsburgh School of Medicine

The RAGEE Group

Available reporting guidelines for prognostic and diagnostic accuracy studies apply primarily to biological assessment and outcomes, overlooking behavioral issues with major public health and safety implications such as violence. The present study aimed to develop the first set of reporting guidance for predictive validity studies of violence risk assessments: the Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement. A systematic search of 8 electronic databases prior to September 2012 identified 279 reporting guidelines for prognostic and diagnostic accuracy studies. Unique items were extracted and modified to make them relevant to risk assessment. A 4-wave Delphi process involving a multidisciplinary team of 37 international experts resulted in a 50-item reporting checklist. The panelists endorsed the RAGEE Statement checklist as being highly satisfactory and as indicating study features that should be reported routinely in manuscripts. Use of these proposed standards has the potential to improve the quality of the risk assessment literature.

Keywords: reporting guidance, risk assessment, violence, checklist, RAGEE

Supplemental materials: http://dx.doi.org/10.1037/lhb0000090.supp

Study quality has been shown to account for variation in clinical research findings (Rutjes et al., 2006). Because it is difficult to assess and compare study quality without transparent and consistent reporting of methodology, investigators in prognostic (McShane et al., 2006) and diagnostic medicine (Bossuyt et al., 2003) have developed well-received guidelines for methodological reporting in accuracy studies. Evidence suggests that the implementation of such guidelines has resulted in an improvement in reporting practices (Plint et al., 2006; Prady, Richmond, Morton, & MacPherson, 2008; Smidt et al., 2006; Smith et al., 2008). However, available guidance of this type is limited to research assess-

ing the risk of biological outcomes, overlooking behavioral issues with major public health and safety implications such as violence.

Given the mortality rate and economic burden associated with violence, the World Health Organization (2002) has designated violence prevention as one of its priorities. This perspective is shared both by the mental health and criminal justice systems of numerous countries, including the United States and the United Kingdom, and is reflected in clinical guidelines for psychologists (American Psychological Association Presidential Task Force on Evidence-Based Practice, 2006), psychiatrists (American Psychiatric Association, 2004; National Institute for Health & Clinical

Excellence, 2009), and nurses (Nursing & Midwifery Council, 2004) that recommend using evidence-based methods to assess violence risk. While the research base on the predictive validity of structured risk assessment instruments has grown exponentially (Buchanan, Binder, Norko, & Swartz, 2012), recent evidence from systematic reviews suggests that this literature has not achieved the same transparency and consistency as fields with established reporting guidelines. Considerable variability has been found in the reporting of essential sample- and study-level information in risk assessment studies published between 1990 and 2011 (Singh, Desmarais, & Van Dorn, 2013), making it difficult to assess the internal and external validity of their findings.

The development of reporting standards for violence risk assessment predictive validity studies could allow more informed comparisons between primary investigations, as well as sounder meta-analyses. This would, in turn, support the development of a cumulative science and potentially increase the reliability, utility, and impact of research in this area (Simera et al., 2010). Hence, to address the limitations of available reporting guidelines for prognostic and diagnostic accuracy studies when applied to the area of violence risk assessment, we used the Delphi technique to develop a novel reporting checklist: the Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement. Following published guidelines for developers of health research reporting guidance (Moher, Schulz, Simera, & Altman, 2010), our aim is to promote consistency and transparency for this important area of the behavioral sciences.

Method

Design

Consistent with the development of previous reporting standards (Hutchings, Raine, Sanderson, & Black, 2006), a multi-wave Delphi process was used to select the item content of the RAGEE Statement. The Delphi method is based on the premise that group decisions are necessary when the scope of a problem is such that no single individual has sufficient expertise and knowledge to effect a solution. It is a structured communication technique that relies on the anonymous feedback of a panel of experts in an iterative process to establish consensus (Powell, 2003). By maintaining the anonymity of panelists and controlling their interactions, the Delphi technique avoids the disadvantages of more conventional consensus-based roundtable discussions and committees (Hasson, Keeney, & McKeena, 2000). Ethical review was waived by the University of South Florida Institutional Review Board; therefore, informed consent was not sought.

Participants

The Delphi panel consisted of 37 experts in the field of violence risk assessment (Table 1). This group included a multidisciplinary set of clinicians, researchers, legal professionals, and journal editors from 10 countries: Australia, Belgium, Canada, Germany, The Netherlands, Norway, Sweden, Switzerland, the United Kingdom, and the United States. The principal investigator (JPS) and coinvestigators (SY, EPM) organized, but were not members of, the Delphi panel. Potential panel members were identified by using

recent reviews of the risk assessment literature (e.g., **>**; Hanson & Morton-Bourgon, 2009; Singh, Serper, Reinharth, & Fazel, 2011; Skeem & Monahan, 2011) and were recruited to serve as experts if they met Farmer and Richman's (1965) criteria for Delphi panelist selection:

1. Extensive knowledge of the problem area and the ability to apply that knowledge
2. Good performance record in their particular area
3. High degree of objectivity and rationality
4. Time available to participate
5. Willingness to participate

Materials

To identify a pool of items for consideration by the Delphi panel, a systematic search was performed to identify existing reporting guidance for prognostic and diagnostic accuracy studies. We searched the Cochrane Methodology Register, Database of Abstracts of Reviews of Effects, NHS Economic Evaluation Databases, Health Technology Assessment Databases, US National Criminal Justice Reference Service Abstracts, PROSPERO, PsycINFO, and MEDLINE prior to September 2012 using combinations of the following Boolean keywords: *prognos**, *diagnos**, *guid**, *checklist*. Additional guidelines were identified using the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network (Altman, Simera, Hoey, Moher, & Schulz, 2008), annotated bibliographies (e.g., Sanderson, Tatt, & Higgins, 2007), and discussion with experts.

Using this search strategy, we identified 279 published checklists (Figure 1). Items from each were extracted by the first and second authors with a high level of interrater agreement as established using a randomly selected subsample of 28 (10.0%) checklists (κ = 0.92). Items addressing the same methodological principle (e.g., the inclusion of a structured abstract) were combined, and the wording of select items was modified to make them relevant to risk assessment (e.g., descriptions of biological tests were changed to descriptions of risk assessment instruments). This procedure, combined with a review of the literature on violence (including sexual violence) and criminal recidivism risk assessment, resulted in the identification of 130 unique items.

Procedure

A four-wave Delphi process was conducted between September 2012 and February 2013 to select which of the 130 initially identified items would be included in the final RAGEE Statement. The Delphi process was conducted electronically using Qualtrics survey software (www.Qualtrics.com), thus effectively managing the geographic dispersion of panelists and overcoming the time constraints related to physical meetings. Qualtrics has been used in recent research with forensic mental health professionals (e.g., Kimonis Fanniff, Borum, & Elliott, 2011; Singh, 2013) and has a number of benefits, including data collection through a secure server, libraries of customizable question templates, and a continuous file saving function to minimize data loss because of browser crashes.

In both the first and second waves of the Delphi process, panelists voted to definitely include, definitely exclude, or abstain from voting

Table 1

Members of the RAGEE Statement Delphi Panel

Name	Affiliation at time of Publication
Stål Bjørkly, PsyD	Institute of Health Sciences, Molde University College, Molde, Norway
Marcus T. Boccacini, PhD	Department of Psychology and Philosophy, Sam Houston State University, Huntsville, Texas
Randy Borum, PsyD	School of Information, University of South Florida, Tampa, Florida
Alec Buchanan, PhD, MD	Department of Psychiatry, Yale University, New Haven, Connecticut
David J. Cooke, PhD	Psychology Department, Glasgow Caledonian University, Scotland, United Kingdom
Sarah L. Desmarais, PhD	Department of Psychology, North Carolina State University, Raleigh, North Carolina
Kevin S. Douglas, LLB, PhD	Department of Psychology, Simon Fraser University, Burnaby, BC, Canada
Michael Doyle, PhD, RMN	Institute of Brain, Behaviour and Mental Health, University of Manchester, Manchester, United Kingdom
John F. Edens, PhD	Department of Psychology, Texas A&M University, College Station, Texas
Eric B. Elbogen, PhD	Department of Psychiatry, University of North Carolina-Chapel Hill School of Medicine, Chapel Hill, North Carolina
Jerome Endrass, PhD	Department of Justice, Psychiatric/Psychological Service, Canton of Zurich, Switzerland
Seena Fazel, MD	Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, United Kingdom
Martin Grann, PhD, MBA	Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden
Laura S. Guy, PhD	Department of Psychiatry, University of Massachusetts Medical School, Worcester, Massachusetts
R. Karl Hanson, PhD	Public Safety Canada, Ottawa, ON, Canada
Robert D. Hare, PhD	Department of Psychology, University of British Columbia, Vancouver, BC, Canada
Grant T. Harris, PhD	Waypoint Centre for Mental Health Care Penetanguishene, ON, Canada
Stephen D. Hart, PhD	Department of Psychology, Simon Fraser University, Burnaby, BC, Canada
Kirk Heilbrun, PhD	Department of Psychology, Drexel University, Philadelphia, Pennsylvania
Mark A. Larsen, JD	Committee for Public Counsel Services, Commonwealth of Massachusetts, Boston, Massachusetts
John Monahan, PhD	School of Law, University of Virginia, Charlottesville, Virginia
Daniel F. Montaldi, PhD	Sexually Violent Predator Program, Florida Department of Children and Families, Tallahassee, Florida
Douglas Mossman, MD	Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati College of Medicine, Cincinnati, Ohio
Tonia L. Nicholls, PhD	Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada
James R. P. Ogloff, JD, PhD	Centre for Forensic Behavioural Science, Monash University and Forensicare, Victoria, Australia
Randy K. Otto, PhD	Department of Mental Health Law & Policy, University of South Florida, Tampa, Florida
John Petrila, JD, LLM	Department of Health Policy and Management, University of South Florida, Tampa, Florida
Thierry H. Pham, PhD	Centre de Recherche en Defense Sociale, Tournai, Belgium
Martin Rettenberger, PhD	Department of Psychology, Johannes Gutenberg University Mainz, Mainz, Germany
Marnie Rice, PhD	Waypoint Centre for Mental Health Care Penetanguishene, ON, Canada
Corine de Ruiter, PhD	Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands
Astrid Rossegger, PhD	Department of Justice, Psychiatric/Psychological Service, Canton of Zurich, Switzerland
Nicholas Scurich, PhD	Department of Psychology & Social Behavior and Department of Criminology, Law & Society, University of California, Irvine, California
Jennifer L. Skeem, PhD	Department of Psychology & Social Behavior, University of California, Irvine, California
Robert L. Trestman, PhD, MD	Correctional Managed Health Care, University of Connecticut Health Center, Farmington, Connecticut
Frank Urbaniok, PhD	Department of Justice, Psychiatric/Psychological Service, Canton of Zürich, Switzerland
Jodi L. Viljoen, PhD	Department of Psychology, Simon Fraser University, Burnaby, BC, Canada

on each of the items. An inclusion threshold of 75% approval and an exclusion threshold of 25% disapproval were set (cf. Campbell, Piaggio, Elbourne, Altman, & the CONSORT Group, 2000). Items falling between these thresholds were retained for a further round of voting. Panelists had the opportunity to suggest new items, as well as to recommend modifications in wording. In the third wave, the panel was asked to dichotomously vote to either include or exclude remaining items. In the final wave, panelists used seven-item Likert scales to register their degree of satisfaction with the finished checklist (1 = *very dissatisfied*; 7 = *very satisfied*), as well as whether the guidance statement should be routinely used as reporting standards for risk assessment predictive validity studies (1 = *strongly disagree*; 7 = *strongly agree*). Upon the completion of each wave, approved items were summarized and panelists were given access to the voting results for each item if requested.

Response rates in each wave were maximized using the Dillman Total Design Method (Dillman, Smyth, & Christian, 2009). In accordance with this approach, an initial e-mail with an active Qualtrics link was sent to panelists on a Friday requesting participation in the given wave. Three reminder e-mails were sent at seven day intervals after the initial distribution. Using this strategy, a 100% panelist response rate was achieved for each wave (Figure 2).

Results

The RAGEE Statement Checklist Criteria

The completed RAGEE Statement includes 50 items and contains minimal reporting standards for the abstract (k items = 4), introduction (k = 2), methods (k = 30), results (k = 6), and discussion (k = 4) sections of risk assessment predictive validity manuscripts, as well as guidance on recommended disclosures (n = 4; Table 2). The methods criteria are divided into six subsections: participants (k = 5), instrument design (k = 7), instrument administration (k = 5), study design (k = 5), predicted outcome (k = 2), and statistical analysis (k = 6). The results criteria are divided into two subsections: participant outcomes (k = 2) and predictive validity (k = 4). The checklist version of the

PROGNOSTIC

Cochrane Methodology Register
(*RECORDS* = 32)

Database of Abstracts of Reviews of Effects
(*RECORDS* = 945)

NHS Economic Evaluation Database
(*RECORDS* = 525)

Health Technology Assessment Database
(*RECORDS* = 209)

US National Criminal Justice Reference Service
(*RECORDS* = 354)

PROSPERO
(*RECORDS* = 40)

PsycINFO
(*RECORDS* = 4932)

MEDLINE
(*RECORDS* = 19935)

Additional Records via Other Sources
(*RECORDS* = 184)

DIAGNOSTIC

Cochrane Methodology Register
(*RECORDS* = 402)

Database of Abstracts of Reviews of Effects
(*RECORDS* = 1328)

NHS Economic Evaluation Database
(*RECORDS* = 3487)

Health Technology Assessment Database
(*RECORDS* = 1567)

US National Criminal Justice Reference Service
(*RECORDS* = 4513)

PROSPERO
(*RECORDS* = 221)

PsycINFO
(*RECORDS* = 6794)

MEDLINE
(*RECORDS* = 9311)

Additional Records via Other Sources
(*RECORDS* = 97)

Checklists Screened for Eligibility
(*N* = 586)

Checklists Excluded as Duplicates or Not Relevant to Diagnostic or Prognostic Methodology
(*N* = 307)
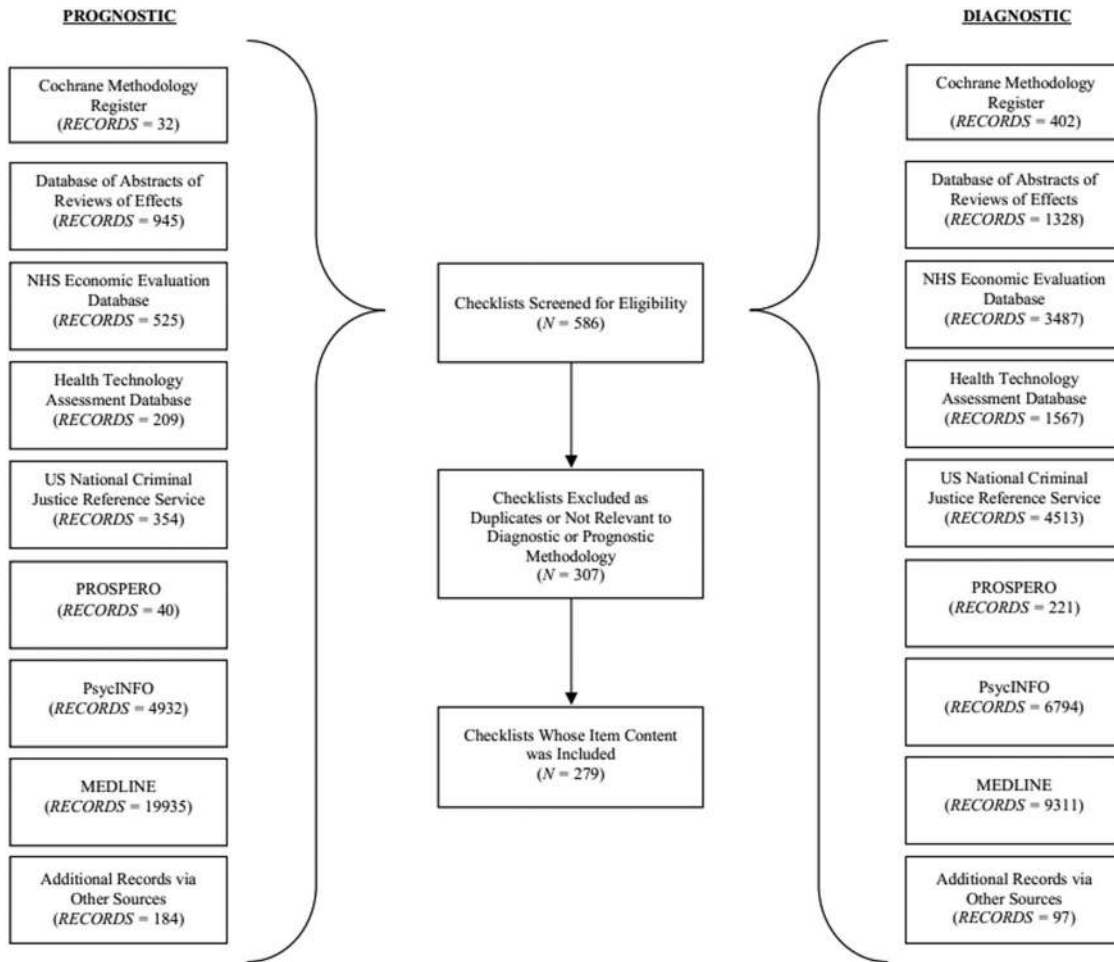
Checklists Whose Item Content was Included
(*N* = 279)

Figure 1. Identification of methodological reporting checklists for prognostic and diagnostic accuracy studies.

RAGEE Statement criteria can be found in the Supplemental Materials (Supplement 1). All criteria in the most comprehensive section of the checklist, concerning methods, can be met in fewer than 250 words (a sample methods section is available upon request), suggesting that the checklist does not place a substantial burden on authors. An elaboration document including exemplars for each item from the peer-reviewed predictive validity literature on risk assessment instruments was also developed to increase the usefulness of the checklist (Supplement 2).

Perceived Usefulness of the RAGEE Statement Checklist

Using 7-point Likert scales, the average satisfaction rating with the checklist was 6.00 (*SD* = 1.04), and the average support rating for using the checklist as reporting standards for risk assessment predictive validity studies was 5.84 (*SD* = 1.31). Narrative comments revealed that lower ratings were due to the desire of some panelists to include mandatory reporting of calibration performance indicators (e.g., positive and negative predictive values) as an item rather than just discrimination performance indicators (e.g., area under the curve and correlation coefficients), the belief

that no guidance should be given for introduction and discussion sections, and uncertainty about whether minimum reporting standards would exclude from consideration studies that merit publication.

Discussion

The development of health research guidance has resulted in increased transparency and consistency in the methodological reporting of diagnostic and prognostic accuracy studies. None of this work, however, has been done in the critical and rapidly growing area of violence risk assessment. The creation of general guidelines for research studies such as the American Psychological Association *Journal Article Reporting Standards* (American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) has been a positive development for the social sciences, but such standards do not provide adequately specific guidance on sample- and study-level characteristics that should be reported to maximize the clinical relevance of the risk assessment literature.

In the present report, we developed the first set of methodological reporting standards for predictive validity studies in violence
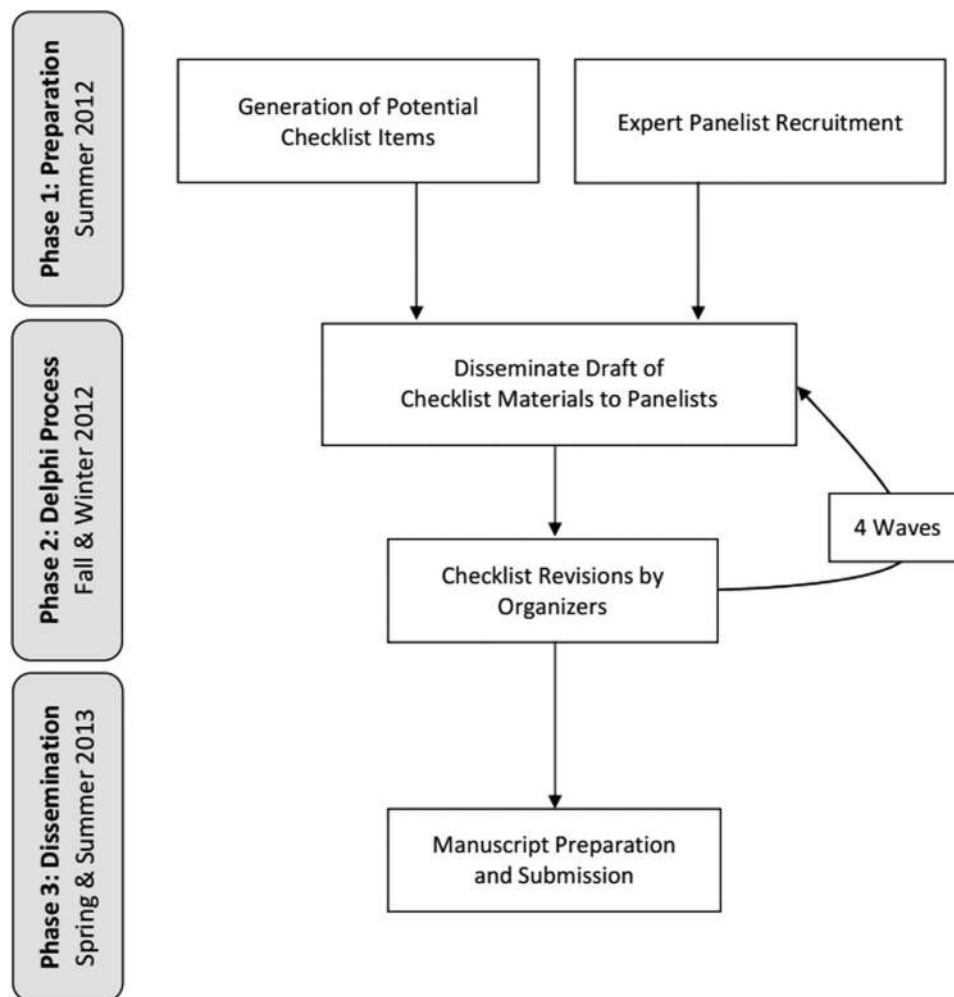
Figure 2. Development of the Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) statement.

risk assessment. A four-wave Delphi process involving 37 international experts from diverse fields resulted in a 50-item reporting checklist. Because the guidance statement was voted highly satisfactory and appropriate for routine use as a reporting standard for risk assessment predictive validity studies, researchers may wish to reference the RAGEE Statement checklist while preparing manuscripts. In addition to being useful for manuscript authors, the use of the checklist by reviewers has the potential to expedite and increase agreement in the peer-review process.

Just as health research reporting guidance for other specialties has been adapted to related fields (Campbell, Piaggio, Elbourne, Altman, & the CONSORT Group, 2012; Ioannidis et al., 2004), the RAGEE Statement checklist may provide a useful basis for the development of methodological standards in other fields of behavioral prediction, such as suicide risk assessment. In its current form, however, the RAGEE Statement is designed for use only in studies of violence (including sexual violence) and criminal recidivism risk assessment.

Adherence to the RAGEE Statement guidance has the potential to resolve and overcome these obstacles to innovation, rigor, and relevance. It is important to note that the items on the RAGEE

Statement checklist represent a *minimum* of what should be reported in risk assessment predictive validity studies at this time. Other valuable demographic, design, and performance information should continue to be reported where appropriate. For example, it is reasonable to assume that a brief summary of past predictive validity and reliability information will be reported in manuscripts. And as the field continues to develop, additional statistical approaches may enrich our picture of an instrument's predictive validity and expand the domain of study features that are desirable to report. It is our aim to update the RAGEE criteria as these developments arise. Hence, the RAGEE should be viewed as a living document. Meanwhile, when RAGEE Statement reporting criteria conflict with a journal's *Instructions for Authors*, please follow the latter.

Conclusion

Mental health professionals are routinely called upon to assess the violence risk presented by their clients, frequently aided by structured instruments. Though a considerable literature exists on the predictive validity of these instruments, such studies are often

Table 2

Risk Assessment Guidelines for the Evaluation of Efficacy (RAGEE) Statement Checklist

Section	Item	Description	Endorsed (N of 37, %)
Abstract	1	Include a structured abstract describing the study	30 (81.1%)
	2	Identify the article as a risk assessment study in which predictive validity is measured	30 (81.1%)
	3	Identify the risk assessment instrument(s) whose predictive validity is measured	37 (100.0%)
	4	State the nature of the principal outcome (e.g., violence, sexual violence, criminal offending, institutional misconduct)	33 (89.2%)
Introduction	5	Provide the rationale and a summary of the scientific/theoretical background for the study	37 (100.0%)
	6	State the research questions and/or study aims	37 (100.0%)
Methods			
Participants	7	Report the sample size	37 (100.0%)
	8	Report the sex/gender composition of the sample	37 (100.0%)
	9	Report the average age at assessment (with dispersion parameter)	31 (83.8%)
	10	Report the index offense composition of the sample	29 (78.4%)
	11	Report the characteristics of groups that underwent subgroup analysis	28 (75.7%)
Instrument design	12	Report the acronym(s) and full name(s) of the instrument(s) under investigation with appropriate reference to source document	37 (100.0%)
	13	Report the number of items on the instrument(s) under investigation	30 (81.1%)
	14	Report the approach by which the assessment information from the instrument(s) under investigation is organized into an overall evaluation of risk	28 (75.7%)
	15	Report the population for which the instrument(s) under investigation was intended to be used	34 (91.9%)
	16	Report the outcome(s) that the instrument(s) under investigation was intended to assess	35 (94.6%)
	17	Report the length of follow-up for which manual-recommended probability estimates of risk were derived for the instrument(s) under investigation	30 (81.1%)
	18	Report the cut-off score(s) and/or risk categories that the instrument(s) under investigation was designed to use to classify risk level	30 (81.1%)
Instrument administration	19	Report whether risk assessments were conducted in the context of research or practice	28 (75.7%)
	20	Identify when risk assessments occurred (e.g., pre-admission, admission, release, post-release)	37 (100.0%)
	21	Report the number of assessors in the study as well as their training in the administration of the instrument(s) under investigation	34 (91.9%)
	22	Identify the source(s) of information used to administer the instrument(s) under investigation	37 (100.0%)
	23	Describe any modifications made to the instrument(s) under investigation	37 (100.0%)
Study design	24	Report the geographical location and clinical setting in which risk was assessed	34 (91.9%)
	25	Describe the method(s) used to recruit participants	34 (91.9%)
	26	Identify the temporal design of the study (prospective or quasi-prospective)	36 (97.3%)
	27	Identify the setting in which participants were followed to ascertain whether the outcome(s) of interest had occurred	37 (100.0%)
	28	Report the average length of follow-up and time at risk (with dispersion parameter, if not fixed), including a description of periods subtracted from follow-up time (e.g., incarceration and/or hospitalization)	35 (94.6%)
Predicted outcome	29	Specify the event(s) coded as meeting outcome criteria (e.g., assault, rape, homicide)	34 (91.9%)
	30	Identify the type (e.g., arrest, charge, conviction, incarceration) and source (e.g., criminal records, self-report, collateral) of information used to detect outcome occurrence	37 (100.0%)
Statistical analysis	31	Describe the statistical methods used to conduct all analyses, and report the purpose of each analysis	30 (81.1%)
	32	Report whether risk scores and/or risk categories of the instrument(s) under investigation were used as an independent variable in analyses	32 (86.5%)
	33	Identify the statistical significance level used	34 (91.9%)
	34	Describe any subgroup analyses planned *a priori*	32 (86.5%)
	35	Report inter-rater reliability for administration of the instrument(s) under investigation (if conducted). If inter-rater reliability was not assessed, clarify why not	28 (75.7%)
	36	Include at least one discrimination performance indicator when measuring predictive validity	32 (86.5%)
Participant outcomes	37	Report the rate of attrition	32 (86.5%)
	38	Report the outcome occurrence rate for the entire sample as well as for relevant subgroups	34 (91.9%)
Predictive validity	39	Report predictive validity performance indicators for each outcome of interest as specified in the Methods with associated dispersion parameters	36 (97.3%)
	40	Report the number of participants with each risk score and/or in each risk category and how many went on to engage in the outcome(s) of interest	29 (78.4%)
	41	Report the results of subgroup analyses planned *a priori* as specified in the Methods	32 (86.5%)
	42	Describe and report the findings of any *post hoc* analyses conducted	28 (75.7%)

(*table continues*)

Table 2 (*continued*)

Section	Item	Description	Endorsed (*N* of 37, %)
Discussion	43	Provide a summary of the principal findings, including a discussion of their relevance in the context of the current literature	35 (94.6%)
	44	Discuss limitations of the study design	35 (94.6%)
	45	Discuss the generalizability of study findings	30 (81.1%)
	46	Discuss future research directions based on study findings	29 (78.4%)
Disclosures	47	Report any commercial interests and/or source(s) of funding as well as their role(s) in the conduct of the study	30 (81.1%)
	48	Report whether an author or translator of the risk assessment instrument(s) under investigation was also a study author	29 (78.4%)
	49	Report whether the study presented in the article has been published in an alternative form (e.g., government report)	28 (75.7%)
	50	Report whether the sample or a portion thereof has been studied in other publications	33 (89.2%)

plagued by inconsistent methodological reporting, limiting their reproducibility and clinical utility. The use of reporting guidelines has the potential to resolve and overcome these obstacles to innovation, rigor, and relevance.

References

Altman, D. G., Simera, I., Hoey, J., Moher, D., & Schulz, K. (2008). EQUATOR: Reporting guidelines for health research. *Lancet, 371,* 1149–1150. doi:10.1016/S0140-6736(08)60505-X

American Psychiatric Association. (2004). *Practice guidelines for the treatment of patients with schizophrenia.* Arlington, VA: Author.

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61,* 271–285. doi:10.1037/0003-066X.61.4.271

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851. doi:10.1037/0003-066X.63.9.839

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glaziou, P. P., Irwig, L. M., . . . the STARD Group. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326,* 41–44. doi:10.1136/bmj.326.7379.41

Buchanan, A., Binder, R., Norko, M., & Swartz, M. (2012). Resource document on psychiatric violence risk assessment. *The American Journal of Psychiatry, 169,* 1–10. doi:10.1176/appi.ajp.2012.169.3.340

Campbell, M. K., Piaggio, G., Elbourne, D. R., Altman, D. G., & the CONSORT Group. (2012). CONSORT 2010 statement: Extension to cluster randomised trials. *British Medical Journal, 345,* e5661. doi:10.1136/bmj.e5661

Campbell, S. M., Cantrill, J. A., & Roberts, D. (2000). Prescribing indicators for UK general practice: Delphi consultation study. *British Medical Journal, 321,* 425–428. doi:10.1136/bmj.321.7258.425

Dillman, D., Smyth, J., & Christian, L. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley.

Farmer, R. N., & Richman, B. M. (1965). *Comparative management and economic progress.* Homewood, IL: Irwin.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21,* 1–21. doi:10.1037/a0014421

Hasson, F., Keeney, S., & McKeena, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing, 32,* 1008–1015. doi:10.1046/j.1365-2648.2000.t01-1-01567.x

Hutchings, A., Raine, R., Sanderson, C., & Black, N. (2006). A comparison of formal consensus methods used for developing clinical guidelines. *Journal of Health Services & Policy Research, 11,* 218–224. doi:10.1258/135581906778476553

Ioannidis, J. P. A., Evans, S. J. W., Gotzsche, P. C., O'Neill, R. T., Altman, D. G., Schulz, K., Moher, D., & the CONSORT Group. (2004). Better reporting of harms in randomized trials: An extension of the CONSORT Statement. *Annals of Internal Medicine, 141,* 781–788. doi:10.7326/0003-4819-141-10-200411160-00009

Kimonis, E. R., Fanniff, A., Borum, R., & Elliott, K. (2011). Clinician's perceptions of indicators of amenability to sex offender-specific treatment in juveniles. *Sexual Abuse: A Journal of Research & Treatment, 23,* 193–211. doi:10.1177/1079063210384278

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., Clark, G. M. (2006). Reporting recommendations for tumor Marker prognostic studies (REMARK). *Breast Cancer Research & Treatment, 100,* 229–235. doi:10.1007/s10549-006-9242-8

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Medicine, 7,* e10000217. doi:10.1371/journal.pmed.1000217

National Institute for Health and Clinical Excellence. (2009). *Core interventions in the treatment and management of schizophrenia in primary and secondary care.* London, United Kingdom: National Institute for Health and Clinical Excellence.

Nursing and Midwifery Council. (2004). *Standards of proficiency for specialist community public health nurses.* London, United Kingdom: Author.

Plint, A. C., Moher, D., Morrison, A., Schulz, K. F., Altman, D. G., Hill, C., & Gaboury, I. (2006). Does the CONSORT checklist improve the quality of reports of randomized controlled trials: A systematic review. *Medical Journal of Australia, 185,* 263–267.

Powell, C. (2003). The Delphi technique: Myths and realities. *Journal of Advanced Nursing, 41,* 376–382. doi:10.1046/j.1365-2648.2003.02537.x

Prady, S. L., Richmond, S. J., Morton, V. M., & MacPherson, H. (2008). A systematic evaluation of the impact of STRICTA and CONSORT recommendations on quality of reporting for acupuncture trials. *PLoS ONE, 3,* e1577. doi:10.1371/journal.pone.0001577

Rutjes, A. W. S., Reitsma, J. B., Di Nisio, M., Smidt, N., van Rijn, J. C., & Bossuyt, P. M. M. (2006). Evidence of bias and variation in diagnostic accuracy studies. *Canadian Medical Association Journal, 174,* 469–476. doi:10.1503/cmaj.050090

Sanderson, S., Tatt, I. D., & Higgins, J. P. T. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36,* 666–676. doi:10.1093/ije/dym018

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *BMC Medicine, 8,* 24. doi:10.1186/1741-7015-8-24

Singh, J. P. (2013, March). *The International Risk Survey (IRiS) project: Perspectives on the practical application of violence risk assessment tools*. Paper presented at the Annual Conference of the American Psychology-Law Society, Portland, Oregon.

Singh, J. P., Desmarais, S., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law, 31,* 55–73. doi:10.1002/bsl.2053

Singh, J. P., Serper, M., Reinharth, J., & Fazel, S. (2011). Structured assessment of violence risk in schizophrenia and other disorders: A systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin, 37,* 899–912. doi:10.1093/schbul/sbr093

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20,* 38–42. doi:10.1177/0963721410397271

Smidt, N., Rutjes, A. W., van der Windt, D. A., Ostelo, R. W., Bossuyt, P. M., Reitsma, J. B., . . . de Vet, H. C. (2006). The quality of diagnostic accuracy studies since the STARD statement: Has it improved? *Neurology, 67,* 792–797. doi:10.1212/01.wnl.0000238386.41398.30

Smith, B. A., Lee, H. J., Lee, J. H., Choi, M., Jones, D. E., Bausell, R. B., & Broome, M. E. (2008). Quality of reporting randomized controlled trials (RCTs) in the nursing literature: Application of the Consolidated Standards of Reporting Trials (CONSORT). *Nursing Outlook, 56,* 31–37. doi:10.1016/j.outlook.2007.09.002

World Health Organization. (2002). *World report on violence and health.* Geneva: Author.

Section	Item	Description	Page	NR	N/A
ABSTRACT	1	Include a structured abstract describing the study			
	2	Identify the article as a risk assessment study in which predictive validity is measured			
	3	Identify the risk assessment instrument(s) whose predictive validity is measured			
	4	State the nature of the principal outcome (e.g., violence, sexual violence, criminal offending, institutional misconduct)			
INTRODUCTION	5	Provide the rationale and a summary of the scientific/theoretical background for the study			
	6	State the research questions and/or study aims			
METHODS					
Participants	7	Report the sample size			
	8	Report the sex/gender composition of the sample			
	9	Report the average age at assessment (with dispersion parameter)			
	10	Report the index offense composition of the sample			
	11	Report the characteristics of groups that underwent subgroup analysis			
Instrument Design	12	Report the acronym(s) and full name(s) of the instrument(s) under investigation with appropriate reference to source document			
	13	Report the number of items on the instrument(s) under investigation			
	14	Report the approach by which the assessment information from the instrument(s) under investigation is organized into an overall evaluation of risk			
	15	Report the population for which the instrument(s) under investigation was intended to be used			
	16	Report the outcome(s) that the instrument(s) under investigation was intended to assess			
	17	Report the length of follow-up for which manual-recommended probability estimates of risk were derived for the instrument(s) under investigation			
	18	Report the cut-off score(s) and/or risk categories that the instrument(s) under investigation was designed to use to classify risk level			
Instrument Administration	19	Report whether risk assessments were conducted in the context of research or practice			
	20	Identify when risk assessments occurred (e.g., pre-admission, admission, release, post-release)			
	21	Report the number of assessors in the study as well as their training in the administration of the instrument(s) under investigation			
	22	Identify the source(s) of information used to administer the instrument(s) under investigation			
	23	Describe any modifications made to the instrument(s) under investigation			
Study Design	24	Report the geographical location and clinical setting in which risk was assessed			
	25	Describe the method(s) used to recruit participants			
	26	Identify the temporal design of the study (prospective or quasi-prospective)			
	27	Identify the setting in which participants were followed to ascertain whether the outcome(s) of interest had occurred			

	28	Report the average length of follow-up and time at risk (with dispersion parameter, if not fixed), including a description of periods subtracted from follow-up time (e.g., incarceration and/or hospitalization)			
Predicted Outcome	29	Specify the event(s) coded as meeting outcome criteria (e.g., assault, rape, homicide)			
	30	Identify the type (e.g., arrest, charge, conviction, incarceration) and source (e.g., criminal records, self-report, collateral) of information used to detect outcome occurrence			
Statistical Analysis	31	Describe the statistical methods used to conduct all analyses, and report the purpose of each analysis			
	32	Report whether risk scores and/or risk categories of the instrument(s) under investigation were used as an independent variable in analyses			
	33	Identify the statistical significance level used			
	34	Describe any subgroup analyses planned *a priori*			
	35	Report inter-rater reliability for administration of the instrument(s) under investigation (if conducted). If inter-rater reliability was not assessed, clarify why not			
	36	Include at least one discrimination performance indicator when measuring predictive validity			
RESULTS					
Participant Outcomes	37	Report the rate of attrition			
	38	Report the outcome occurrence rate for the entire sample as well as for relevant subgroups			
Predictive Validity	39	Report predictive validity performance indicators for each outcome of interest as specified in the Methods with associated dispersion parameters			
	40	Report the number of participants with each risk score and/or in each risk category and how many went on to engage in the outcome(s) of interest			
	41	Report the results of subgroup analyses planned *a priori* as specified in the Methods			
	42	Describe and report the findings of any *post hoc* analyses conducted			
DISCUSSION	43	Provide a summary of the principal findings, including a discussion of their relevance in the context of the current literature			
	44	Discuss limitations of the study design			
	45	Discuss the generalizability of study findings			
	46	Discuss future research directions based on study findings			
DISCLOSURES	47	Report any commercial interests and/or source(s) of funding as well as their role(s) in the conduct of the study			
	48	Report whether an author or translator of the risk assessment instrument(s) under investigation was also a study author			
	49	Report whether the study presented in the article has been published in an alternative form (e.g., government report)			
	50	Report whether the sample or a portion thereof has been studied in other publications			

NOTE. NR = not reported but applicable to study; N/A = not applicable to study.

RAGEE

Risk Assessment Guidelines
for the Evaluation of Efficacy

Elaboration Document
with Item Exemplars

R1 (Abstract). Include a structured abstract describing the study

Objectives: An actuarial model was developed in the MacArthur Violence Risk Assessment Study to predict violence in the community among patients who have recently been discharged from psychiatric facilities. This model, called the multiple iterative classification tree (ICT) model, showed considerable accuracy in predicting violence in the construction sample. The purpose of the study reported here was to determine the validity of the multiple ICT model in distinguishing between patients with high and low risk of violence in the community when applied to a new sample of individuals.

Methods: Software incorporating the multiple ICT model was administered with independent samples of acutely hospitalized civil patients. Patients who were classified as having a high or a low risk of violence were followed in the community for 20 weeks after discharge. Violence included any battery with physical injury, use of a weapon, threats made with a weapon in hand, and sexual assault.

Results: Expected rates of violence in the low- and high-risk groups were 1 percent and 64 percent, respectively. Observed rates of violence in the low- and high-risk groups were 9 percent and 35 percent, respectively, when a strict definition of violence was used, and 9 percent and 49 percent, respectively, when a slightly more inclusive definition of violence was used. These findings may reflect the "shrinkage" expected in moving from construction to validation samples.

Conclusions: The multiple ICT model may be helpful to clinicians who are faced with making decisions about discharge planning for acutely hospitalized civil patients.

Monahan, J, Steadman, H., Robbins, P., Appelbaum, P., Banks, S., Grisso, T., Heilbrun, K., Mulvey, E., Roth, L., & Silver, E. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services, 56,* 810-815.

Comment: A brief description including the study aims, sample, methods, principal results, and implications should be reported, in keeping with the specific journal's requirements.

R2 (Abstract). Identify the article as a risk assessment study in which predictive validity is measured

Background: Current violence risk assessment instruments are time-consuming and mainly developed for forensic psychiatry. A paucity of violence screens for acute psychiatry instigated the development and validation of the V-RISK-10. **The aim of this prospective naturalistic study was to test the predictive validity of the V-RISK-10 as a screen of violence risk after discharge from two acute psychiatric wards.**

Methods: Patients were screened with V-RISK-10 before discharge, and incidents of violence were recorded 3, 6, 9 and 12 months after discharge. A total of 381 of the 1017 patients that were screened completed the follow up.

Results: The ROC-AUC values for any violent behaviour were 0.80 and 0.75 (p < 0.001) for the 3 and 12 months follow-up periods, respectively, and significant for both genders. The most accurate risk estimates were obtained for severe violence. For persons without a known history of violence prior to the screening, AUCs were 0.74 (p = 0.004) and 0.68 (p = 0.002).

Conclusions: Results indicate that the V-RISK-10 is a valid and clinically useful screen for violence risk after discharge from acute psychiatry, and even significant for patients without a known previous history of violence.

Roaldset, J. O,, Hartvig, P., & Bjørkly, S. (2011). V-RISK-10: Validation of a screen for risk of violence after discharge from acute psychiatry. *European Psychiatry, 26,* 85-91.

Comment: Predictive validity is defined as the extent to which an instrument-based estimate (e.g., total risk score, actuarial risk bin, or final risk judgment) predicts an outcome measure (e.g., arrest, conviction, or incarceration for a violent offense).

R3 (Abstract). Identify the risk assessment instrument(s) whose predictive validity is measured

Objective: This study tested the interrater reliability and criterion-related validity of structured violence risk judgments made by using one application of the structured professional judgment model of violence risk assessment, **the HCR-20 violence risk assessment scheme**, which assesses 20 key risk factors in three domains: historical, clinical, and risk management.

Methods: The HCR-20 was completed for a sample of 100 forensic psychiatric patients who had been found not guilty by reason of a mental disorder and were subsequently released to the community. Violence in the community was determined from multiple file-based sources.

Results: Interrater reliability of structured final risk judgments of low, moderate, or high violence risk made on the basis of the structured professional judgment model was acceptable (weighted kappa=.61). Structured final risk judgments were significantly predictive of postrelease community violence, yielding moderate to large effect sizes. Event history analyses showed that final risk judgments made with the structured professional judgment model added incremental validity to the HCR-20 used in an actuarial (numerical) sense.

Conclusions: The findings support the structured professional judgment model of risk assessment as well as the HCR-20 specifically and suggest that clinical judgment, if made within a structured context, can contribute in meaningful ways to the assessment of violence risk.

Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services, 54,* 1372-1379.

Comment: Report the acronym(s) and/or full name(s) of the violence risk assessment instrument(s) whose predictive validity was examined in the study.

R4 (Abstract). State the nature of the principal outcome (e.g., violence, sexual violence, criminal offending, institutional misconduct)

Purpose: To investigate the validity of risk factors and established risk measures in predicting community violence in an acute mental health sample up to 20 weeks post-discharge.

Method: Prospective cohort follow-up study conducted between January 2006 and August 2007. Baseline assessments were conducted while participants were inpatients. The measures were rated following interview with the participants, record review and speaking to someone who knows the person well (e.g. friend, relative, carer). Baseline measures were then compared with frequency and severity of **violence in the community** post-discharge at 20 weeks.

Results: In the 20-week period post-discharge, 29 (25.4%) of the 114 participants were violent. All the risk measures and measures of impulsiveness and anger were predictive of violence where $p < 0.05$. The HCR-20 total, psychopathy and clinical factors were strongly correlated with the frequency of violence where $p < 0.05$.

Conclusions: The risk factors and risk measures that have been found to be predictive in forensic samples are also predictive in acute mental health samples, although the effects are not as large. Future research needs to be con-ducted with a larger sample to include investigation of differences in risk factors based on gender and social support. Services and clinicians need to consider how to integrate findings into useful frameworks to support decisions and contribute to managing risk. This should assist in identifying interventions aimed at preventing community violence.

Doyle, M., Carter, S., Shaw, J., & Dolan, M. (2011). Predicting community violence from patients discharged from acute mental health units in England. *Social Psychiatry & Psychiatric Epidemiology, 47*, 627-637.

Comment: The principal outcome is defined as that which was used as the primary dependent variable in predictive validity analyses.

R5 (Introduction). Provide the rationale and a summary of the scientific/theoretical background for the study

> START has experienced quick uptake into clinical practice: We are aware of implementations in at least 10 countries, and the manual has been translated into four languages, with an additional four translations underway. **However, only a handful of studies have examined the reliability and validity of START assessments (Braithwaite, Charette, Crocker, & Reyes, 2010: Chu et al., 2011; Gray et al., in press; Nicholls, Brink, Desmarais, Webster, & Martin, 2006; Nonstad et al., 2010). The original validation study was published by Nicholls et al. (2006). Evaluating START assessments completed by nurses, social workers, and psychiatrists regarding 137 male forensic psychiatric patients, Nicholls et al. found excellent interrater agreement overall (intraclass correlation coefficient, $ICC_2 = .87$, $p < .001$) and within professional disciplines nursing = .88; social work = .92; and psychiatry = .80; all $ps < .001$). The authors also reported significantly higher START total scores for patients who engaged in aggression over the 12-month follow-up: any aggression to others ($M = 75.66$ vs.65.86), verbal aggression ($M = 75.86$ vs. 66.82), aggression against objects ($M = 77.90$ vs. 68.00), physical aggression against others ($M = 76.32$ vs. 68.25), violence against others ($M = 81.82$ vs. 69.12), and sexual aggression ($M = 80.63$ vs. 70.24; all $ps < .05$). Receiver operating characteristic (ROC) analyses of a subsample of 50 patients who remained hospitalized throughout follow up revealed good validity in predicting verbal aggression (area under the curve [AUC] = .72, $SE = 0.07$, $p < .001$), physical aggression against objects (AUC = .67, $SE = 0.08$, p _ .05), physical aggression against others (AUC = .70, $SE = 0.08$, $p < .01$), and sexually inappropriate behavior (AUC = .92, $SE = 0.10$, $p < .05$).**
>
> **Results of this research are promising, but further evaluation is necessary for several reasons. The predictive validity of the final risk estimates, one of the identifying features of the structured professional judgment approach (Singh, Grann, & Fazel, 2011; Skeem & Monahan, 2011), has only been examined in two studies (Braithwaite et al., 2010; Gray et al., in press), and the validation samples have been quite small ($ns = 34$-50). Furthermore, there have been several significant changes to the instrument since these evaluations. Now, each of the 20 START items is scored in terms of *both* vulnerability and strength, and final risk estimates of low, moderate, or high are made across seven outcome domains (Webster et al., 2004). In 2009, the authors published a text revision of the START manual (Version 1.1; Webster et al., 2009), which included content updates to three items (Mental State, Emotional State, and Treatability), explicit operationalization of START components left undefined in the 2004 consultation edition of the manual (e.g., each of the risk outcome domains, strengths and vulnerabilities, and key and critical items), and specification of coding time frames (i.e., item ratings based on functioning over the past 2 to 3 months or since the previous START assessment).**

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: reliability and validity of START assessments. *Psychological Assessment, 24*, 685-700.

Comment: The rationale for the study is defined as the reason why the present investigation is necessary given existing evidence. Describing the scientific/theoretical background for the study serves to situate the investigation in light of previous research, contemporary legal statutes, organizational and/or government reports, and relevant clinical guidelines.

R6 (Introduction). State the research question(s) and/or study aim(s)

The present study is part of a large-scale research project concerning the clinical, criminological, and legal aspects of sexual murderers using a comprehensive and comparatively large sample of offenders convicted of sexually motivated murder in Germany (e.g., Berner et al, 2008; Briken, Habermann, Kafka, Berner, & Hill, 2006; Briken, Nika, & Berner, 1999; Hill, Habermann, Berner, & Briken, 2006; Hill, Habermann et al., 2008; Hill, Ujeyl et al., 2008; Ujeyl et al., 2008). **The main aim of the present study was to examine the predictive accuracy of four well established risk assessment instruments: the Static-99 (Hanson & Thornton, 2000), the Historical-Clinical-Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997), the Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster., 1997), and the Psychopathy Checklist-Revised (PCL-R; Hare, 2003).** In order to achieve comparability with the existing status of risk assessment research we used commonly used effect sizes for the investigation of the predictive accuracy of the instruments. **Furthermore, we examined different recidivism criteria in order to prove differential effects, and we investigated the predictive accuracy of the different subscales and the individual items of the instruments as well.**

Hill, A., Rettenberger, M., Habermann, N., Berner, W., Eher, R., & Briken, P. (2012). The utility of risk assessment instruments for the prediction of recidivism in sexual homicide perpetrators. *Journal of Interpersonal Violence, 27*, 1-26.

Comment: Identify the main points of inquiry that the present research sought to answer. Whereas the rationale and scientific/theoretical background serve to provide larger context within the field, the research question(s) and specific aims(s) focus more precisely on the particular issues addressed in the present study.

R7 (Methods). Report the sample size

> Participants were **99** male adolescents who had been released from custody. They had been at liberty for a minimum of 12 months and were traced on the Home Office Police National Computer (HOPNC). The 99 participants constituted 80.5% of a baseline sample of 123 male adolescents who had been assessed on measures of personality and risk while in custody.

Dolan, M. C., & Rennie, C. E. (2008). The Structured Assessment of Violence Risk in Youth as a predictor of recidivism in a United Kingdom cohort of adolescent offenders with conduct disorder. *Psychological Assessment, 20*, 35-46.

Comment: Identify the number of participants in the study.

R8 (Methods). Report the sex/gender composition of the sample

> Between April 15, 2005 and December 31, 2007, the clinical staff completed 258 START assessments of 61 patients. Of these, 47 could be included in the validation study because they had been in the hospital for three months after their first START assessment. **These were 39 men (83%) and eight women (17%).**

Nonstad, K., Nesset, M. B., Kroppan, E., Pedersen, T. W., Nottestad, J. A., Almvik, R., & Palmstierna, T. (2010). Predictive validity and other psychometric properties of the Short-Term Assessment of Risk and Treatability (START) in a Norwegian high secure hospital. *International Journal of Forensic Mental Health, 9*, 294-299.

Comment: Provide the number and/or percentage of the participants who were men/male and/or women/female.

R9 (Methods). Report the average age at assessment (with dispersion parameter)

> **Average age at discharge from inpatient forensic psychiatric care, release from prison or onset of probation was 34.07 (SD = 11.00, range 16-58) years.** … The study had a retrospective follow-up design. Subjects were followed from release until first event of a sexual or a violent non-sex reoffence resulting in a new criminal conviction, or the end of follow-up (June 1st, 1999).

Sjöstedt, G., & Långström, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime & Law, 8*, 25-40.

Comment: The mean age should be reported in years. A dispersion parameter should be reported to describe spread, using a standard deviation or confidence interval. In cases where the distribution is skewed, a median and interquartile range may be substituted.

R10 (Methods). Report the index offense composition of the sample

> **The breakdown of the sample by current offence at the start of sentence was as follows: 61 violent (28.2%), 28 burglary (13.0%), 21 theft (9.7%), 15 sexual (6.9%), 7 fraud and forgery (3.2%), 3 criminal damage (1.4%), and 81 (36.1%) offenders were convicted for a diverse range of offences, including motoring, drugs, and financial crimes.**

Dowdy, E. R., Lacy, M. G., Unnithan, N. P. (2002). Correctional prediction and the Level of Supervision Inventory. *Journal of Criminal Justice, 30,* 29-39.

Comment: When the study sample involves a correctional and/or forensic population, provide the number and/or percentage of the participants who engaged in each type of act that resulted in contact with the criminal justice or forensic mental health setting where risk assessments took place. This item may not be relevant for non-forensic samples.

R11 (Methods). Report the characteristics of groups that underwent subgroup analysis

Variable	MIC (*n* = 101)	MIOF (*n* = 100)	Jail (*n* = 107)	Total (*n* = 308)
Race and ethnicity				
African American	18 (17.8)	14 (14.0)	28 (26.2)	60 (19.5)
Hispanic	54 53.5)	46 (46.0)	48 (44.9)	147 (47.9)
White	29 (28.7)	40 (40.0)	31 (29.0)	100 (32.6)
Gender				
Male	44 (43.6)	71 (71.0)	61 (57.0)	175 (57.0)
Female	57 (56.4)	29 (29.0)	46 (43.0)	132 (43.0)
Age				
18-39	59 (58.4)	83 (83.0)	68 (63.6)	210 (68.4)
40-61	42 (41.6)	17 (17.0)	39 (46.4)	97 (31.6)
Marital status				
Single or divorced	86 (85.1)	83 (83.0)	97 (90.7)	265 (86.3)
Married	15 (14.9)	17 (17.0)	10 (9.3)	42 (13.7)

Note: Values are *n* (%). MIC = Mentally Impaired Caseload; MIOF = Mentally Impaired Offender Facility; SSI = Supplemental Security Income; Unavail = Data not

Castillo, D. E., & Alarid, L. F. (2011). Factors associated with recidivism among offenders with mental illness. *International Journal of Offender Therapy & Comparative Criminology, 55*, 98-117.

Comment: Provide the same (or a relevant subset of) descriptive characteristics for groups that underwent subgroup analysis as were provided for the overall sample. This may be done in the text or in a table as in the example above.

R12 (Methods). Report the acronym(s) and full name(s) of the instrument(s) under investigation with appropriate reference to source document

> **The Psychopathy Checklist-Revised (PCL-R; Hare, 1991)** contains 20 items, each scored on a 3-point scale from 0 to 2, giving a total score ranging from 0 to 40. It was developed as a measure of the extent to which an individual matches Cleckley's (1941) description of the prototypical psychopath, and has been found to be a good predictor of violent recidivism (Dolan & Doyle, 2000).

Thomson, L., Davidson, M., Brett, C., Steele, J., & Darjee, R. (2008). Risk assessment in forensic patients with schizophrenia: The predictive validity of actuarial scales and symptom severity for offending and violence over 8-10 years. *International Journal of Forensic Mental Health, 7*, 173-189.

Comment: A source document is a report that initially presents an instrument's items, discusses its appropriate use, and provides information on psychometric validation. Appropriate source documents include articles, books, government reports, Masters theses, doctoral dissertations, and conference presentations, with preference given to published manuals and peer-reviewed instrument development studies.

R13 (Methods). Report the number of items on the instrument(s) under investigation

> The VRAG (Quinsey et al., 1998) **comprises 12 items**, including such item s as the Psychopathy Checklist Revised (PCL- R; Hare, 2004) score (which in turn has 20 items), elementary school adjustment, offender's age at time of index of-fence, etc. If we could not score a particular item then that item was rated as a '0'. We note that the updated manual (Quinsey et al., 2006) pro-rating is now recommended. We did not score the VRAG if more than four items could not be scored.

Snowden, R. J., Gray, N. S., Taylor, J., & MacCulloch, M. J. (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine, 37,* 1539-1549.

Comment: This value should represent the total number of originally published items on each instrument whose predictive validity was measured in the study. Modifications to the instrument(s), such as systematically omitted items, should be reported separately.

R14 (Methods). Report the approach by which the assessment information from the instrument(s) under investigation is organized into an overall evaluation of risk

> *J-SORRAT-II.* The J-SORRAT-II is a 12-item **actuarial tool** designed for assessing risk of violence among male juvenile offenders who were 12 to 18 years old at the time of their index sexual offense (Epperson et al., 2005). A number of items on the J-SORRAT-II focus on the youths' sexual and nonsexual offense history (e.g., number of adjudications as a sex offender, number of victims in sex offenses). Other variables examine youths' treatment history (i.e., completion of sex offender treatment), school history (e.g., special education), and past victimization experiences (e.g., number of physical abuse victimization events).

Viljoen, J. L., Scalora, M., Cuadra, L., Bader, S., Chavez, V., Ullman, D., & Lawrence, L. (2008). Assessing risk for violence in adolescents who have sexually offended: A comparison of the J-SOAP-II, J-SORRAT-II, and SAVRY. *Criminal Justice & Behavior, 35*, 5-23.

Comment: Common approaches in forensic risk assessment combine information gathered at the item level using one of two strategies to arrive at a risk estimate. The first of these approaches is actuarial assessment, which involves the estimation of the likelihood of future antisocial behavior by assigning numerical values to factors associated with offending. These numbers are then combined using a statistical algorithm to translate an individual's total score into a group-based probabilistic estimate of future antisocial behavior. The second approach is referred to as structured professional judgment (SPJ), which involves *aide-mémoires* that guide assessors to estimate risk after reviewing empirically- and theoretically-based risk and/or protective factors.

R15 (Methods). Report the population for which the instrument(s) under investigation was intended to be used

The START is a structured professional judgment guide for the assessment of seven often interrelated risks associated with mental, substance use, and personality disorders in adults: violence to others, self-harm, suicide, unauthorized leave, substance abuse, self-neglect, and being victimized. The instrument consists of 20 dynamic factors that are assessed for both Strength and Vulnerability on a 3-point ordinal scale from 0 (minimally present) to 2 (maximally present). Strength and Vulnerability ratings should be scored independent of one another, and a patient may be scored high (or low) on both Strength and Vulnerability for any particular item. For example, a patient may receive a high Vulnerability rating for relationships (Item 2) if he or she is involved in an abusive intimate relationship but also may receive a high Strength rating if he or she has a warm, loving, and reciprocal relationship with his or her parents, other family members, or peers. Based on item ratings, identification of key and critical items (i.e., items that are particularly relevant, either recently or historically, to individual risk), and consideration of historical factors, assessors estimate risk as low, moderate, or high for each of the seven outcome domains. Strength and Vulnerability total scores can be calculated for research purposes by summing the item ratings (possible range = 0–40). **START is intended for use with both inpatient and outpatient populations in civil psychiatric, forensic psychiatric, and correctional settings.**

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: reliability and validity of START assessments. *Psychological Assessment, 24*, 685-700.

Comment: Relevant description of the population may include: setting (institutional/inpatient vs. community/outpatient), forensic status (forensic vs. non-forensic), age (adult vs. juvenile), sex (men vs. women), and whether the instrument was designed to be used in mental health and/or correctional contexts.

R16 (Methods). Report the outcome(s) that the instrument(s) under investigation was intended to assess

The Violence Risk Appraisal Guide (VRAG) is an actuarial violence risk assessment developed on 618 violent offenders evaluated in a maximum security forensic psychiatric facility. Most in this development sample were convicted before or after the evaluation while a minority was found not guilty by reason of insanity; about a quarter met the diagnostic criteria for a psychotic disorder. In development, the VRAG's items were selected for their ability to provide independent and incremental information about the likelihood that subjects later met **the operational definition of violent recidivism – a criminal charge for a violent offense or reinstitutionalization for violent conduct that would otherwise have resulted in a criminal charge.**

Harris, G. T., & Rice, M. E. (2007). Adjusting actuarial violence risk assessments based on aging or the passage of time. *Criminal Justice & Behavior, 34*, 297-313.

Comment: Provide both the type of outcome (e.g., violent, sexual, general offending) as well as the source of outcome detection (e.g., criminal charges, arrest, conviction, incarceration, self-report, collateral interviews). As risk assessment instruments other than those adopting the actuarial approach may not have been intended to assess the likelihood of an outcome detected via a particular source, the latter criterion may not be relevant for all risk assessment instruments.

R17 (Methods). Report the length of follow-up for which manual-recommended probability estimates of risk were derived for the instrument(s) under investigation

The SORAG is an actuarial risk assessment tool for sexual offenders that was developed by Canadian forensic researchers. This instrument is a modification of the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006), which was developed to predict violent and sexual recidivism among adult male offenders; 10 of the 14 items of the SORAG are the same as in the VRAG. The SORAG is conceptualized for sexual offenders to assess violent recidivism risk, which includes sexual offences involving physical contact with the victim. The instrument consists of 14 weighted items: lived with biological parents up to age 16, elementary school maladjustment, history of alcohol problems, marital status, criminal history for nonviolent offences, criminal history for violent offences, previous convictions for sexual offences, sexual offences against girls under age 14 only, failure on prior conditional releases, age at index offence, *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.) (American Psychiatric Association, 1980) criteria for any personality disorder, *DSM-III* criteria for schizophrenia, phallometric test results indicating pedophilia or sexual sadism, and PCL-R score. Based on the total score the evaluator can allocate the offender to one of nine risk categories. **By means of these risk categories, it is possible to infer to empirically calculated probabilities of violent (including sexual) recidivism after 7 and 10 years at risk, respectively.**

Rettenberger, M., Matthes, A., Boer, D. P., & Eher, R. (2010). Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *International Journal of Offender Therapy & Comparative Criminology*, *54*, 169-186.

Comment: This information is routinely available in instrument manuals and development studies. This item may not be relevant for non-actuarial risk assessment instruments, as the generation of probabilistic risk estimates is unique to the actuarial approach.

R18 (Methods). Report the cut-off score(s) and/or risk categories that the instrument(s) under investigation was designed to use to classify risk level

> The Static-99 is composed of 10 historical risk factors (see Table 2) that have to be coded from file information. **The factors add up to a maximum total score of 12 that is subsequently translated into four risk categories: low (0,1), medium low (2-3), medium high (4-5) and high (6 or more; Hanson & Thornton, 1999).**

de Vogel, V., de Ruiter, C., van Beek, D. & Mead, G. (2004). Predictive validity of the SVR-20- and Static-99 in a Dutch sample of treated sex offenders. *Law & Human Behavior, 28,* 235-251.

Comment: Cut-off scores and risk categories established by the developers of the instrument(s) should be reported. If the study used different cut-off scores and/or risk categories than those identified by the instrument developers, this should also be stated.

R19 (Methods). Report whether risk assessments were conducted in the context of research or practice

> **One hundred thirty-two youth in custody were invited to participate in the study. Of these, parent/legal guardians refused consent for 28 youth (21%), five youth refused consent (4%), and one youth withdrew partway through the study (<1%). Furthermore, 19 of these youth (14%) did not receive SAVRY ratings due to insufficient collateral information to code the measure or because the youth did not complete a research interview. We invited 102 youth from the mental health assessment center to participate. Of these, 19 youth refused consent (19%) and two youth withdrew partway through the study (2%). Sixteen of these youth (16%) did not receive SAVRY ratings for the same reasons listed above.** The gender and age composition of youth who did not participate in the study was not significantly different from youth who consented to participate (for gender, $\chi^2 = 0.31$, $P > .05$; for age, $F(1,226) = .78$, $P > 4.05$).

Penney, S. R., Lee, Z., & Moretti, M. M. (2010). Gender differences in risk factors for violence: An examination of the predictive validity of the Structured Assessment of Violence Risk in Youth. *Aggressive Behavior, 36*, 390-404.

Comment: Studies conducted in a research context are those in which the risk assessment instruments under investigation *were not* administered as part of routine practice. Studies conducted in a practice context are those in which the risk assessment instruments under investigation *were* administered as part of routine practice. As the example above suggests, this criterion can be met by stating that risk assessments were conducted as part of a voluntary research process.

R20 (Methods). Identify when risk assessments occurred (e.g., pre-admission, admission, release, post-release)

The HCR–20 consists of 20 items: 10 items related to historical factors (e.g. employment problems, history of mental illness), 5 items related to current clinical presentation (e.g. lack of insight, current symptoms of major mental illness) and 5 items related to future risk factors (e.g. lack of personal support, non-compliance with remediation attempts). Each item was scored as 0 (not present), 1 (partially or possibly present) or 2 (present), leading to a maximum total score of 40, and maximum sub-scale scores of 20 for the historical scale and 10 for the clinical and risk scales. If insufficient information was available we omitted the item score but pro-rated the scale and sub-scales (by taking the average score on scale or sub-scale). If too many items were omitted (more than five in total, two for the historical scale and one for the clinic and risk scales), then the assessment was considered invalid and omitted from the analysis. In all we were able to score 887 patients **at their point of discharge.**

Gray, N. S., Taylor, J., & Snowden, R. J. (2008). Predicting violent reconvictions using the HCR-20. *British Journal of Psychiatry, 192*, 384-387.

Comment: Describe the time point at which risk assessments were systematically performed. If the study design calls for repeated assessments at multiple time points, report the completion rate at each time point.

R21 (Methods). Report the number of assessors in the study as well as their training in the administration of the instrument(s) under investigation

The Static-99 was scored at the time of the 1989/90 release using only that information available at the release date. **The Static-99 was scored for each case by one investigator trained in the coding rules who had extensive experience using the Static-99 in sexually violent predator evaluations.**

Sreenivasan, S., Garrick, T., Norris, R., Cusworth-Walker, S., Weinberger, L. E., Essres, G., Turner, S., & Fain, T. (2007). Predicting the likelihood of future sexual recidivism: Pilot study findings from a California sex offender risk project and cross-validation of the Static-99. *Journal of the American Academy of Psychiatry & the Law*, *35*, 454-468.

Comment: Relevant information regarding training includes formal certification through attendance of workshops and/or seminars on the administration of the instrument and/or supervision by formally trained assessors. If different instruments under investigation were administered by different assessors, specify this.

R22 (Methods). Identify the source(s) of information used to administer the instrument(s) under investigation

All measures were coded from the correctional files of participants, an acceptable research method according to the measures' manuals. The PCL measures and HCR-20 recommend the use of an interview for clinical purposes, although file-based coding is acceptable and permissible for research purposes. Further, although the HCR-20 recommends the use of a "low, moderate, high" structured final risk judgment, the manual also states that risk is generally assumed to increase with increases in the number of risk factors, making the evaluation of HCR-20 scores a necessary component of its overall evaluation (a procedure described by Douglas & Kropp, 2002). **The files are detailed and often voluminous, containing social, psychological, psychiatric, medical, criminal, and legal reports and information.**

Douglas, K. S., Yeomans, M., & Boer, D. P. (2005). Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice & Behavior, 32,* 479-510.

Comment: Sources of information used to administer risk assessment instruments may include: criminal justice records, clinical records, school records, interviews (with evaluee, family members, probation officers, service providers), mental health examination reports, and neuropsychological testing results.

R23 (Methods). Describe any modifications made to the instrument(s) under investigation

> **As it was considered by the present authors to be unclear whether an offense was a noncontact sex offense or not based on the CPIC files alone and the very low frequency of official charges for noncontact sex offenses in our sample, a zero was assigned to every offender for this Static-99 item.** Sex, age, number of victims, and relationship to victim (i.e., stranger, acquaintance, relative) were gathered from self-report data and used to score the corresponding Static-99 and SORAG items. **Item 8 of the SORAG ("female victims under 14 years") was scored as yes if the offender had only female victims under age 16, as this was the closest approximation of the item available for the present sample.**

Nunes, K. L., Firestone, P., Bradford, J. M., Greenberg, D. M., & Broom, I. (2002). A comparison of modified versions of the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research & Treatment, 14,* 253-269.

Comment: Modifications made to risk assessment instruments may include: systematically adding or removing certain items or subscales, changing the scoring or weighting of items, altering cut-off thresholds or risk categories, and the use (or lack thereof) of prorating.

R24 (Methods). Report the geographical location and clinical setting in which risk was assessed

All subjects suspected of having committed a sexual offence, who underwent a formal forensic psychiatric evaluation (FPE) for the court at **the Department of Forensic Psychiatry, Aarhus University Hospital, or at the Clinic of Forensic Psychiatry, Ministry of Justice in Copenhagen, Denmark**, between 1 January 1978 and 31 December 1992 ($n = 445$), were followed for 10.25 years as part of a larger retrospective follow-up study.

Bengtson, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime & Law, 14,* 85-106.

Comment: Geographic location may be reported by specifying the country, region, state, city, and/or institution as appropriate. Clinical settings may include hospitals and clinics (inpatient or outpatient), correctional institutions, and/or community corrections (parole or probation offices).

R25 (Methods). Describe the method(s) used to recruit participants

> **Participants were randomly selected from a list of all male offenders released from the two facilities during the specified period. Researchers randomly chose a starting position on the list and then selected every other file.**

Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice & Behavior, 35*, 1095-1108.

Comment: Recruitment methods may include purposive selection of a subsample with specific characteristics (e.g., individuals with a substance use disorder), sequential sampling of eligible individuals during a particular time frame, convenience sampling, random selection, opt-in, opt-out, or the evaluation of total cohort. If any payments or other compensation was offered for participation, specify these.

R26 (Methods). Identify the temporal design of the study (prospective or quasi-prospective)

> The **prospective cohort follow-up design** chosen was modelled on the MacVRAS to evaluate the predictive validity of historical, dispositional and clinical risk factors and to test the hypothesis that the non-forensic participants with high baseline scores on the VRAG, HCR-20 and VRS will be significantly more likely to be violent up to 20 weeks post-discharge than participants with low scores.

Doyle, M., Carter, S., Shaw, J., & Dolan, M. (2011). Predicting community violence from patients discharged from acute mental health units in England. *Social Psychiatry & Psychiatric Epidemiology, 47*, 627-637.

Comment: A prospective design is defined as a study design in which risk assessment instruments are administered, and participants are subsequently examined at a future time to determine whether the outcome(s) of interest occurred. A quasi-prospective design (sometimes referred to as "postdiction" or "retrospective prediction") is defined as a study design in which risk assessment instruments are completed by a rater using information available at a past time point, and whether the outcome(s) of interest occurred is assessed at a second time point either in the future or in the past based on available archival information. If the study was quasi-prospective, state whether assessors were blind to outcome occurrence at the time when the risk assessment instrument was administered.

R27 (Methods). Identify the setting in which participants were followed to ascertain whether the outcome(s) of interest had occurred

> **Twenty weeks after hospital discharge was chosen as the time frame for the analysis here because this was the period during which the prevalence of violence by patients in the community was at its highest (Steadman *et al*, 1998).** Research interviewers attempted two follow-up interviews with enrolled patients in the community during this period, approximately 10 weeks apart. … Patients and collaterals independently were asked whether the paitent had been involved in any of several categories of violent behaviour in the past 10 weeks (Lidz *et al*, 1993).

Monahan, J., Steadman, H. J., Appelbaum, P., Robbins, P. C., Mulvey, E. P., Silver, E., Roth, L., & Grisso, T. (2000). Developing a clinically useful actuarial tool for assessing violence risk. *British Journal of Psychiatry, 176,* 312-319.

Comment: Commonly used follow-up contexts include intra-institutional settings (e.g., jail, prison, hospital) or the community (including outpatient clinic settings). Hospital or outpatient care may be forensic (under court jurisdiction) or non-forensic. If participants were followed across multiple settings, specify each.

R28 (Methods). Report the average length of follow-up and time at risk (with dispersion parameter, if not fixed), including a description of periods subtracted from follow-up time (e.g., incarceration and/or hospitalization)

Inpatient violence was coded from the date of the index admission. **The average length of stay for patients in our sample was 108 days (SD=871.14 days, range=8–6366 days).** At the time of completing the HCR-20, PCL:SV, and VSC, the raters were blind to whether or not patients were violent following community release. **Depending on when they were discharged, patients were tracked in the community for an average of almost two years (M=690.26 days, SD=184.31 days, range from 312 days to 1053 days).** To increase the likelihood that comprehensive and reliable follow-up data were obtained, multiple sources were used (e.g. psychiatric hospital records, review panel office records, coroner records, additional psychiatric hospital and unit records from 16 general hospitals throughout the province with designated psychiatric units, BC Forensic Psychiatric Services records, and corrections/criminal records). Despite the limitations inherent in not using follow-up interviews with the patient and collaterals (see Steadman et al., 1998), prior research has evidenced an acceptable base rate of follow-up community violence using strictly archival methods (see Harris, Rice, & Quinsey, 1993; McNiel et al., 2003; Menzies & Webster, 1995). The University Research Ethics Review Committee at Simon Fraser University and the Ministry of the Attorney General Corrections Branch approved the research protocol.

….

To incorporate the length of time from the day of release from the hospital until the first incident of community violence, four survival analyses were carried out for men and women. Cutting scores of _20 on the HCR-20 total and _8 on the PCL:SV total were used. **Analyses for community violence took into account any time the patient spent institutionalized (e.g. hospitalized, jail).** The outcome variables for these analyses included any violence, physical violence, any crime, and violent crime following hospital discharge.

Nicholls, T.L., Ogloff, J.R.P., Douglas, K.S. (2004). Assessing risk for violence among male and female civil psychiatric patients: The HCR-20, PCL:SV, and VSC. *Behavioral Sciences and the Law, 22,* 127-158.

Comment: Length of follow-up is defined as the amount of time during which participants were observed to determine whether the outcome(s) of interest occurred. Time at risk is defined as the amount of time during which participants had the opportunity to engage in the outcome(s) of interest. Thus, in some studies (e.g., fixed follow-up studies) the length of follow-up and time at risk can be equivalent. However, time at risk can be decreased if other events intervene to reduce the participant's opportunity to engage in the outcome. For example, if the outcome is community violence, then periods of time in detention or hospital may be subtracted. Dispersion parameters, such as a standard deviation or confidence interval, should be reported to provide a measure of spread.

R29 (Methods). Specify the event(s) coded as meeting outcome criteria (e.g., assault, rape, homicide)

> Violent recidivism was operationally defined as any criminal charge for **a violent offense against persons (assault, assault causing bodily harm, wounding, attempted homicide, homicide, kidnapping, forcible confinement, armed robbery, and all "hands-on" sexual offenses)** that occurred subsequent to the index offense. Also included were **any actions that resulted in patients being returned to maximum security (only males could qualify) that, in the judgment of the research assistants, would otherwise have resulted in such a criminal charge** (<15% of violent

Harris, G. T., Rice, M. E., & Cormier, C. A. (2002). Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law & Human Behavior, 26*, 377-394.

Comment: If the outcome of interest was not restricted to specific events (e.g., any violent offense), state this and provide examples of events included in the definition (e.g., assault, homicide, rape, robbery). In cases where the outcome of interest was restricted to specific events, list these.

R30 (Methods). Identify the type (e.g., arrest, charge, conviction, incarceration) and source (e.g., criminal records, self-report, collateral) of information used to detect outcome occurrence

> **Reconviction data was obtained from the Offenders Index database**, a national database covering convictions incurred in England and Wales. **Offenders were classified according to whether or not they had a sexual reconviction.** Note that those reconvicted for other kinds of offense were included in the "not sexually reconvicted"

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research & Treatment, 14,* 139-153.

Comment: The *type* of information refers to the measurable and observable criterion that is used as an outcome. The *source* of information is defined as the documentation used to detect whether outcomes had occurred. In select cases, both the type and source of information will be the same (e.g., self-report of violence).

R31 (Methods). Describe the statistical methods used to conduct all analyses, and report the purpose of each analysis

To determine the predictive validity of the revised 20-item version of the DA with its weighted scoring, the investigators used data reported from the attempted femicide cases and abused controls on the revised 20-item DA. The ability of the revised DA to correctly identify the attempted femicide cases, an independent sample, was evaluated through plots of ROC curves. Receiver operating characteristic curves represent the sensitivity and 1-specificity of a measure at each successive unit that could be a potential threshold for high risk designation. We developed estimates of the area under the ROC curve (AUC), and tested whether the AUC was greater than the chance diagonal (.500), the average value under random prediction methods. In addition, we calculated sensitivity and specificity using each of the top three levels of danger (increased danger, severe danger, and extreme danger) as a threshold for being designated high risk for attempted femicide. In addition, we also compared the mean scores on the revised 20-item DA between the three study groups (e.g., femicide, attempted femicide, and abused controls) using analysis of variance (ANOVA) and Tamhane's T2 statistic for testing whether pairwise comparisons of group means were equal when the variance differs between the groups.

Campbell, J. C., Webster, D. W., & Glass, N. (2009). The Danger Assessment: Validation of a lethality risk assessment instrument for intimate partner femicide. *Journal of Interpersonal Violence, 24*, 653-674.

Comment: Describe statistical procedures in sufficient detail as to allow for replication of the main steps of analyses, and briefly provide a rationale for the choice of procedures.

R32 (Methods). Report whether risk scores and/or risk categories of the instrument(s) under investigation were used as an independent variable in analyses

The predictive validity of the violence risk assessment method was studied by logistic regression analysis. Separate analyses were performed (1) for the occurrence of any incident of violent or criminal behavior in the subsequent observational period (regardless of the occurrence of any risk enhancing behavior) and (2) for the occurrence of any risk enhancing behavior (regardless of the occurrence of any violent or criminal behavior). **The predictors studied consisted of the mean scores on the historical, clinical and situational subscales of the HKT-30, the final risk judgment by the case manager based on the HKT-30, the HoNOS-MDO mean score, and the total numbers of needs and unmet needs as assessed by the case manager on the CANFOR.**

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

Comment: In cases where risk scores were used for some analyses and risk categories for others, specifically report what was used as the independent variable in predictive validity analyses.

R33 (Methods). Identify the statistical significance level used

The data were analysed by SPSS 16.0 for Windows. Intraclass correlation coefficients were used to estimate the interrater reliability for the V-RISK-10 sum score and for each item. Receiver operating characteristics (ROC) analysis was used to measure the predictive accuracy of the instrument, as ROCs are recommended for use in risk assessment studies for they are less dependent on the base rate of aggression (Mossman 1994, Douglas et al. 1999). This analysis forms a function of the true positive rate (Sensitivity) and false positive rate (Specificity). The area under the curve (AUC) displays a summary measure for the discrimination efficiency of a scale. This can range from 0.5 to 1.0 indicating the chance of a perfect ability of discrimination. One-way ANOVA was employed to analyse the possible differences between sum scores and the selected category of violence risk level assessed (low, moderate, high). The possible differences in sum scores between three groups in terms of outcome recommendation were also analysed using one-way ANOVA, and t-tests and chi-square were also used. A **conventional 5% significance level** and 95% confidence interval (CI) were employed for all analysis.

Yao, X., Li, Z., Arthur, D., Hu, L., & Cheng, G. (2011). The application of a violence risk assessment tool among psychiatric service users: A preliminary study. *Journal of Psychiatric & Mental Health Nursing, 19*, 438-445.

Comment: The statistical significance level is the threshold the investigators used to identify a likely false positive effect.

R34 (Methods). Describe any subgroup analyses planned *a priori*

The first analytic objective was to compare the prevalence of antisocial attitudes among the adolescent offenders to the CTS norms reported for adults in Knight et al. (2006). First, the adolescent means on each scale were compared with the adult means reported in Knight et al. (2006). Comparisons were made using an effect size statistic (Cohen's *d*) rather than a traditional test of significance. Because a significance test is heavily dependent on the specific sample size for the comparison, an index that was independent of the sample size was chosen. **Part of this objective was to determine how the study sample compared to the adult norms in terms of the variability of responses about the mean for each scale. To accomplish this objective, the percentage of the adolescent sample that was greater or less than two adult norm thresholds, the 33rd and the 67th percentiles (Knight et al., 2006), was compared.**

Dembo, R., Turner, C. W., & Jainchill, N. (2007). An assessment of criminal thinking among incarcerated youths in three states. *Criminal Justice & Behavior, 34*, 1157-1167.

Comment: *A priori* analyses are defined as those planned before the start of data analysis.

R35 (Methods). Report inter-rater reliability for administration of the instrument(s) under investigation (if conducted). If inter-rater reliability was not assessed, clarify why not

> **Reliability of the actuarial instruments was assessed by comparing scores generated by the two independent codings. Intraclass correlation coefficients were .96 (95% confidence interval [CI] = .84, .99) for the VRAG, .95 (95% CI = .81, .99) for the RRASOR, .87 (95% CI = .57, .97) for the Static-99, and .96 (95% CI = .86, .99) for the SORAG.**

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment, 15,* 413-25.

Comment: Commonly reported performance indicators for measuring inter-rater reliability in the forensic risk assessment literature include Cohen's kappa coefficient (κ) and the intraclass correlation coefficient (ICC).

R36 (Methods). Include at least one discrimination performance indicator when measuring predictive validity

> To test Hypothesis 1 on validity, **the predictive validity was assessed by means of ROC analysis (Mossman, 1994; Rice & Harris, 1995). We chose this statistical method because it is less reliant than other statistical analyses (like correlation coefficients) on base rates for recidivism and the particular cut-off score chosen to classify cases. Also, normality need not be assumed (Rice & Harris, 1995). ROC analyses result in a plot of the true positive rate (sensitivity) against the false positive rate (1 – specificity) for every possible cut-off score of the instrument. The resulting AUC can be interpreted as the probability that a randomly selected reoffender would score higher on the instrument than a randomly selected nonreoffender.**

Lodewijks, H. P. B., Doreleijers, T. A. H., & de Ruiter, C. (2008). SAVRY risk assessment in violent Dutch adolescents: Relation to sentencing and recidivism. *Criminal Justice & Behavior, 35,* 696-709.

Comment: The predictive validity of risk assessments can be divided into two components: discrimination and calibration. In the context of structured risk assessment, *discrimination* describes an instrument's ability to retrospectively differentiate between those who engaged in the outcome of interest and those who did not. Examples of discrimination indicators include odds ratios, correlation coefficients, and the area under the receiver operating characteristic curve. In contrast, *calibration* describes the instrument's level of fit between prospectively predicted and observed risk. Examples of calibration indicators include the positive and negative predictive values as well as the number needed to detain and number safely discharged. For a review of discrimination and calibration performance indicators, see Singh (2013).

R37 (Results). Report the rate of attrition

Fifty-nine of the 390 participants could not be reached for follow-up interviews. Thus, there is complete follow-up data for 331 patients (67% of all approached) where of 34 (10%) participated only in the 10 week follow-up, 23 (7%) only in the 20 week follow-up, and 274 (83%) gave interviews on both occasions. For 83 participants (25%), additional collateral follow-up data was retrieved. A comparison of the patients with and without any follow up interview is shown in Table 1.

Sturup, J., Kristiansson, M., & Lindqvist, P. (2011). Violent behaviour by general psychiatric patients in Sweden: Validation of Classification of Violence Risk (COVR) software. *Psychiatry Research, 188*, 161-165.

Comment: Attrition is defined as the loss of participants over the course of the study. Sources of attrition should be described, if known, including death, emigration/deportation, name changes, voluntary drop-out, inability to contact the participant and inability to obtain information from records upon follow-up.

R38 (Results). Report the outcome occurrence rate for the entire sample as well as for relevant subgroups

> **The recidivism rate at one year post-release was 64%, with significantly more males (*n=146, 70.5%*) than females (*n=41, 49.4%*)** reporting either being arrested or having committed an undetected crime ($\chi^2 = 11.55$, $p<.01$). **The rate of violent recidivism was 17%, with significantly more males (*n=42, 20%*) than females (*n=8, 9.6%*)** reporting having been arrested or having committed undetected violent criminal acts ($\chi^2 = 4.78$, $p<.05$).

Hastings, M. E., Krishnan, S., Tangney, J. P., & Stuewig, J. (2011). Predictive and incremental validity of the Violence Risk Appraisal Guide scores with male and female jail inmates. *Psychological Assessment, 23*, 174-183.

Comment: Specify the number of individuals who completed follow-up (with a percentage of the total sample, for comparison) who were found to have engaged in the outcome of interest. Relevant subgroups are those that underwent subsequent analyses.

R39 (Results). Report predictive validity performance indicators for each outcome of interest as specified in the Methods with associated dispersion parameters

ROC analyses were performed to assess the predictive validity of the Static-99 and the SORAG. **For sexual recidivism, the Static-99 yielded an area under the curve (AUC) of .70 (SE = .05, CI = .60-.79) and the SORAG yielded an AUC of .65 (SE = .06, CI = .52-.76).** The corresponding correlation coefficients were .18 and .17 for the Static-99 and SORAG, respectively. To assess the relative predictive accuracy of the Static-99 and SORAG, their respective AUCs were compared. The difference was nonsignificant (Z = .97; p > .30) indicating that performance was similar for both measures. ... **For violent (including sexual) recidivism, both the Static-99 and the SORAG yielded an AUC of .69 (SE = .04, CI = .60-.77).** The corresponding correlation coefficient was .23 for both the Static-99 and SORAG. As is evident from visual inspection of these results, the difference between the AUCs of the respective measures was nonsignificant (Z = .04; p > .90).

Nunes, K. L., Firestone, P., Bradford, J. M., Greenberg, D. M., & Broom, I. (2002). A comparison of modified versions of the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research & Treatment, 14,* 253-269.

Comment: Performance indicators statistically measure an instrument's ability to either prospectively or retrospectively assess the likelihood of an outcome of interest. Dispersion parameters, such as a standard deviation of confidence interval, provide a measure of spread.

R40 (Results). Report the number of participants with each risk score and/or in each risk category and how many went on to engage in the outcome(s) of interest

Static-99			
Risk category	**n (%)**	**Recidivists**	**Non-recidivists**
Low	0 (0)	0	0
Medium-low	11 (41)	2	9
Medium-high	10 (37)	3	7
High	6 (22)	2	4

Wilcox, D., Beech, A., Markall, H. F., & Blacker, J. (2009). Actuarial risk assessment and recidivism in a sample of UK intellectually disabled sexual offenders. *Sexual Aggression, 15*, 97-106.

Comment: Present the rates of the relevant outcomes by category of assessed risk. This outcome information may be displayed in a contingency table as above or in the text.

R41 (Results). Report the results of subgroup analyses planned *a priori* as specified in the Methods

> In the PD cohort, the VRAG predicted 2-year violent failures with an AUC of the ROC of .68 (95% CI=.62 to .73). At a cutoff of 13 points, sensitivity was .57 and specificity was .71. Positive and negative predictive values were .40 and .83, respectively. The H-10 had an AUC of the ROC of .71 (95% CI=.66 to .76). At the inflexion cutoff of 12, the sensitivity of H-10 was .72 and the specificity was .60. The positive predictive value was .38, and the negative predictive value was .86.
> In the schizophrenia cohort, the receiver operating curve of the VRAG mounted an area of only .60, which is not significantly larger than that of the .50 line of no information (95% CI=.50 to .69). At a cutoff of 0 points, sensitivity was .68, specificity was .53, positive predictive value was .20, and negative predictive value was .91. The H-10's AUC of the ROC was .66 (95% CI=.56 to .75). At a cutoff of 8 points on the H-10, its sensitivity was .88 and its specificity was .36 (positive and negative predictive values were .19 and .95, respectively).

Grann, M., Belfrage, H., & Tengström, A. (2000). Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Criminal Justice & Behavior, 27,* 97-114.

Comment: *A priori* analyses are defined as those planned before the start of data analysis.

R42 (Results). Describe and report the findings of any *post hoc* analyses conducted

> **Since the aim of the studied instruments is to predict future violence in violent offenders, analyses were repeated only for the respondents who committed violent offences before TBS treatment (*n* = 30). Six of the seven recidivists belong to this group.** Of these 30 respondents, 16 had serious addiction problems. Five of the seven recidivists belonged to the group of ex-patients with past violence combined with severe addiction. **Table 4 gives an overview of the sensitivity and specificity values connected with these variables when predicting future criminal violence. Designating violent female offenders with addiction problems as possible recidivists reduces the false positive rate by a factor of two. However, as specificity rises, sensitivity is reduced, which is a general problem in risk assessment.**

Schaap, G., Lammers, S., & de Vogel, V. (2009). Risk assessment in female forensic psychiatric patients: A quasi-prospective study into the validity of the HCR-20 and PCL-R. *Journal of Forensic Psychiatry & Psychology, 20*, 354-365.

Comment: *Post hoc* analyses are defined as those not planned before the start of data analysis.

R43 (Discussion). Provide a summary of the principal findings, including a discussion of their relevance in the context of the current literature

> **The primary goal of this research was to see if two well-established instruments for the prediction of future violence, the VRAG and the HCR-20, were applicable to a black population in the UK. Our results show that the accuracy of prediction of future violent reconvictions were good for both instruments and closely resembled the prediction efficacy for the white population.** As such we believe that this study provides a sound evidence base for the use of both of these instruments within this particular ethnic minority group discharged from medium secure psychiatric services within the UK. Of course, this does not necessarily mean that these instruments will also be effective in other ethnic minority groups, or that other instruments will also be effective in this ethnic group. Clearly further research is needed to secure a sound evidence-base for these populations. **However, the converging evidence, to which this study is supportive, appears to be that the combination of a few out of several possible risk factors can be used to produce an effective prediction instrument and that these factors appear to cross most racial and cultural boundaries.** Hence, while this direct evidence is needed, it would seem likely that these instruments will be effective in other ethnic groups.

Snowden, R. J., Gray, N. S., & Taylor, J. (2010). Risk assessment for future violence in individuals from an ethnic minority group. *International Journal of Forensic Mental Health, 9*, 118-123.

Comment: After summarizing the results of the study, the discussion section should articulate the findings with those of other studies in the literature and suggest potential implications for relevant stakeholder groups (e.g., researchers, practitioners, and policymakers).

R44 (Discussion). Discuss limitations of the study design

There are several limitations to this study. First, prospective predictive research is hampered by the clinical goals of risk assessment, i.e. risk management and prevention (Dernevik et al., 2002; Hart, 1998). Hart (1998) stated that predictions of violence are not passive assessments, but decisions that influence services delivered to individuals: ''Clinicians are bound - morally, ethically, and legally - to try to prove themselves wrong when they predict violence and take every reasonable action to prevent violence'' (p. 365). **In our study, clinicians were able to use the results of the HCR-20 ratings, for instance, for decisions concerning leave. Thus, it is very likely that risk management was influenced by the results of the risk assessment, for instance, high risk patients were not released from the hospital, or were separated if the risk of inpatient violence was judged to be high. So, the AUC values we obtained were already high, but might have been even higher if the results had not been used to manage risk. Second, the sample was derived from only one Dutch forensic psychiatric hospital, thereby limiting generalization.** Nevertheless, we consider this group to be representative of Dutch offenders with a tbs order, because they are largely similar in demographic, psychiatric and criminal characteristics to the total population of tbs offenders (see van Emmerik & Brouwers, 2001). **Third, the mean follow-up period of this study was somewhat limited; some patients had a very short follow-up period of only 1 or 2 months. Also, the range of follow-up periods was rather large (1-37 months), which complicates comparison between patients.** Nevertheless, the survival analyses we conducted take differences in time-at-risk into account. **Fourth, we found a rather low base rate of violence.** Although we conducted ROC analyses that are insensitive to base rates, the low base rate might have had an effect on the Cox regression analyses. **A final limitation is that data regarding violent outcome were not always reliable. Incidents of physical violence are not always reported on the information bulletins.** For example, it is possible that incidents of physical violence between patients are not observed by staff or told to staff. This is the case for inpatients, but even more so for patients who are in the transmural treatment phase or who can go outside the hospital without supervision. It should be noted, however, that most of these limitations would have had a negative effect on the predictive accuracy of the HCR-20, thus the findings might have been even stronger without these limitations.

de Vogel., V., & de Ruiter, C. (2006). Structured professional judgment of violence risk in forensic clinical practice: A prospective study into the predictive validity of the Dutch HCR-20. *Psychology, Crime & Law. 12*, 321-336.

Comment: Limitations are defined as potential weakness in the study's design and conduct that may have influenced findings.

R45 (Discussion). Discuss the generalizability of study findings

>   Whether our findings would be replicated if the numerical findings were stronger is an empirical question. **Another important question is whether the findings would be generalizable if more serious forms of violence could be measured (4). The results of our study were similar whether physical (more serious) or nonphysical (less serious) violence was used as a criterion. Similar findings have been reported for other HCR-20 studies (12).**
>
>   **Measures were coded for research purposes, so HCR-20 scores did not follow the patients. However, the treating psychiatrists probably would have included an HCR-20 completed independently for clinical practices, or a risk assessment of some kind, in their discharge summaries. It is unclear what, if any, effect this practice would have on the validity of the HCR-20 indexes collected in this study for research purposes. A higher HCR-20 score could cause increased surveillance, leading to observation of more violence. It also reasonably could lead to more effective risk management and treatment, leading to fewer episodes of violence to observe.** Whatever the effect, it was indirect because the outpatient clinicians did not have the HCR protocols used in this study.

Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services, 54,* 1372-1379.

Comment: Generalizability is defined as the extent to which the results of the study may be applicable to other populations and/or jurisdictions. Included in this is the extent to which research findings can be applied in practice.

R46 (Discussion). Discuss future research directions based on study findings

In spite of these limitations, the present study is one of the first evaluations of violence risk assessments completed using the revised version of the START and one of the only studies to include the START final risk estimates in the prediction models. These findings add essential new information to the growing evidence supporting START, and structured professional judgment more broadly, as approaches that clinicians can use to assess risk for a range of aggressive outcomes among adults with mental, substance use, and personality disorders. Findings also contribute to an emerging body of literature supporting the value of considering both risk and protective factors to inform assessments of violence risk. **An important next step will be to examine whether consideration of dynamic risk and protective factors, and use of START in particular, improves risk management efforts and, ultimately, reduces the prevalence and severity of aggressive outcomes. Finally, a comprehensive assessment of violence risk should include consideration of service-level (e.g., staff de-escalation tactics and training, shift change) and system-level (e.g., restraint policy) factors that may increase or decrease the likelihood of patient aggression (Gadon, Johnstone, & Cooke, 2006; Hamrin, Iennaco, & Olsen, 2009), in addition to the more "traditional" client-level assessment of violence risk (whether guided by the START or some other instrument). Although there have been some recent efforts in this area (e.g., Johnstone & Cooke, 2010), continued work is needed.**

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

Comment: After drawing out the implications and limitations of the study, the investigators should then point to the further research that is needed to clarify and further answer the questions raised by the study and its findings.

R47 (Disclosures). Report any commercial interests and/or source(s) of funding as well as their role(s) in the conduct of the study

Singh, J. P., Grann, M., Lichtenstein, P., Långström, N., & Fazel, S. (2012). A novel approach to determining violence risk in schizophrenia: Developing a stepped strategy in 13,806 discharged patients. *PLoS ONE, 7,* e31727.

Comment: A commercial interest is defined as a financial conflict such as ownership of the rights to a copyrighted instrument available for purchase, or ownership of stock in a company whose value may potentially be influenced by the results of the study. Examples of roles that funders could play in the conduct of a study include salary support for co-authors and investigators; extent of control over the study design, data collection and analysis, continuation of funding or publication of the study results; access to participants (patients or inmates); permission to use an instrument or a treatment intervention such as a medication in the study.

R48 (Disclosures). Report whether an author or translator of the risk assessment instrument(s) under investigation was also a study author

All the data from the original MacArthur Violence Risk Assessment Study were placed on the Web at http://macarthur.virginia.edu in 2001. The study reported here began data collection after the original data were made publicly available. **The software was developed by COVR, Inc., in which some of the authors have a financial interest (PA, SB, TG, JM, EM, PR, LR, ES, and HS).**

Monahan, J., Steadman, H. J., Robbins, P. C., Appelbaum, P., Banks, S., Grisso, T., Heilbrun, K., Mulvey, E. P., Roth, L. & Silver, E. (2005) An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric services 56*(7): 810-815.

Comment: Disclose those study investigators who were authors or translators of the manuals and/or development studies of instruments whose predictive validity was measured.

R49 (Disclosures). Report whether the study presented in the article has been published in an alternative form (e.g., government report)

> **Portions of the research described in this article were part of a dissertation (Langton, 2003) submitted by the first author in partial fulfillment of the requirement of the PhD degree at the University of Toronto.** The dissertation research was supervised by the second author (H.E.B.). **Portions of this research were presented at the 2002 annual meeting of the Association for the Treatment of Sexual Abusers (Langton, Barbaree, Harkins, Seto, & Peacock, 2002).**

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice & Behavior, 34,* 37-59.

Comment: Alternative forms of publication that warrant disclosure include book chapters, government reports, Masters theses, doctoral dissertations, and conference presentations.

R50 (Disclosures). Report whether the sample or a portion thereof has been studied in other publications

> **The present sample included the 178 child molesters and rapists described in earlier articles (Quinsey et al., 1995; Rice, Harris, & Quinsey, 1990; Rice, Quinsey, & Harris, 1991), plus 14 sex offenders from the same samples who had been re-leased since the earlier studies. These 192 sex offenders were supplemented by 96 sex offenders from other studies of released offenders (Rice, Harris, Lang, & Bell, 1990; Rice, Harris, & Cormier, 1992).**

Rice, M. E., & Harris, G. T. (1997). Cross-validation and extension of the Violence Risk Appraisal Guide for Child Molesters and Rapists. *Law & Human Behavior, 21*, 231-241.

Comment: This item refers to the specific case in which study participants have already been included in other analyses presented in a previous journal article.

Closing Notes

1) Information to meet checklist criteria may be provided in the text, tables, and/or figures of a manuscript. Where necessary, some criteria can be met by including information in electronic supplements that are routinely accessible to readers.

2) If an item is not relevant to the study (e.g., "Report the index offense composition of the sample" when the study involves a non-forensic sample), please select the "N/A" option on the checklist.

3) Authors of manuscripts concerning unstructured risk assessments can also benefit from the use of the RAGEE Statement. In such studies, checklist items that concern instruments should be marked "N/A".

4) When RAGEE Statement reporting criteria conflict with a journal's *Instructions for Authors*, please follow the latter.