Use of risk assessment instruments to predict violence in forensic inpatient settings: a systematic review and meta-analysis

Taanvi Ramesh

Magdalen College

A thesis submitted in fulfilment of the requirements for the degree:

Master of Science by Research in Psychiatry

Trinity Term 2017

Word Count: 23,999

Acknowledgements

I would like to thank everyone who played a role in helping this project and thesis reach completion. Firstly, I am extremely grateful to my supervisor Professor Seena Fazel, who has been a key source of guidance and supervision. I am grateful to Dr. Artemis Igoumenou for assisting with the data extraction for this review. My gratitude is also owed to the other members of the Oxford Forensic Psychiatry research team – Monique Ewen, Helen Hailes, Jelle Lamsma, Dr. Mark Toynbee, Dr. Rongqin Yu and Achim Wolf – for answering any and every question I posed them and providing support throughout this year. I am also thankful to Professors Klaus Ebmeier and Mary-Jane Attenburrow for their assessment during my transfer viva.

I would like to thank my friends and family, in both Oxford and London, for their support during this year and for making my final year in Oxford an extremely memorable one.

I am also extremely grateful to a number of external researchers who were kind enough to help with this review by sending me data that was not reported in their paper manuscripts. Thanks are owed to the following: Dr. Kaoru Arai, Dr. Oliver Chan, Professor Geoff Dickens, Dr. Óscar Herrero, Dr. Helen Miles, Professor Robert Snowden, Professor Lindsay Thomson and Dr. Vivienne de Vogel.

I would like to disclose no conflicts of interest of funding sources for this review.

Table of Contents

Abstract

Name: Taanvi Ramesh

Affiliations: Department of Psychiatry & Magdalen College, University of Oxford

Degree: Master of Science by Research in Psychiatry

Term: Trinity Term 2017

Thesis Title: Use of risk assessment instruments to predict violence in forensic inpatient settings: a systematic review and meta-analysis

BACKGROUND & AIMS: Violent behaviour by psychiatric inpatients can have many negative consequences for the physical and mental health of both psychiatric staff and other inpatients. The problem is of particular concern on forensic psychiatric wards, where the prevalence of violence is higher than on other inpatient wards. In order to assess, manage and potentially prevent this type of violence, instruments have been developed with the aim of identifying individuals who are at increased risk of violence within a certain timeframe. This systematic review and meta-analysis aims to investigate the accuracy of these risk assessment tools for the prediction of violence on forensic inpatient wards.

METHODS: The nine most commonly used violence risk assessment instruments that have been used in inpatient settings were included. A systematic search of five databases (CINAHL, Embase, Global Health, PsycINFO and PubMed) was conducted to retrieve all studies examining the predictive accuracy of these tools in forensic inpatient settings. A range of accuracy estimates and descriptive study- and sample-related variables were extracted. A quality assessment was performed for each eligible study using the QUADAS-2 (a tool designed to assess methodological quality for systematic reviews of studies investigating diagnostic or prognostic accuracy). Summary performance measures and HSROC curves were produced and meta-regression analyses investigated study and sample effects on accuracy.

RESULTS: Fifty-two eligible publications were identified, of which 43 provided information on tool accuracy in the form of AUC statistics. These provided data on 78 samples, with 7,705 participants. Due to lack of sufficient data reporting, 35 samples (3,306 participants from 19 publications) could be included in the full meta-analysis of all performance measures. Risk assessment instruments were separated into those designed for imminent violence prediction and those designed for longer-term prediction (such as the HCR-20 and PCL-R). Imminent tools performed well for the screening out of low risk individuals, with a summary specificity of 0.99 (95% confidence interval [CI]: 0.80-1.00) and negative predictive value (NPV) of 0.99 (interquartile range [IQR]: 0.85-1.00). Identification of higher-risk patients was poorer, with a summary sensitivity of 0.59 (95% CI: 0.29-0.83) and a median positive predictive value (PPV) of 0.36 (IQR: 0.10-0.93). For longer-term tools, the summary accuracy estimates were as follows: sensitivity = 0.75 (95% CI: 0.65-0.83); specificity = 0.56 (95% CI: 0.46-0.66); PPV = 0.56 (IQR: 0.30-0.75); and NPV = 0.75 (IQR: 0.58-0.95). As an overall measure of accuracy, the median area under the curve (AUC) value for the wider group of 78 samples indicated better performance for imminent tools – AUC = 0.83 (IQR: 0.71-0.85) – compared with longer-term tools with an AUC of 0.68 (IQR: 0.62-0.75). Meta-regression analyses indicated that no study- or sample-related variables were associated with between-study differences in AUCs.

DISCUSSION & CONCLUSIONS: This is a systematic review and meta-analysis of the predictive accuracy of violence risk assessment instruments for forensic inpatient violence. Imminent risk assessment instruments, including the Brøset Violence Checklist (BVC) and the Dynamic Appraisal of Situational Aggression (DASA), were found to be more accurate for the prediction of inpatient violence in forensic wards than those designed for longer-term prediction, such as the HCR-20 and VRAG.

Chapter 1: Introduction

1.1 The problem of inpatient violence

Violence in correctional and psychiatric populations has been a widely-researched topic. Violent incidents involving inpatients are recorded by nursing staff who are constantly monitoring patients and estimates of prevalence are therefore more accurate and reliable than outpatient proxies which often rely on self or informant report. In inpatient settings, violence is generally divided into three categories: physical violence, verbal aggression and property damage. This aligns with the definition of violence used by the Dynamic Appraisal of Situational Aggression (DASA), a risk assessment instrument designed to predict inpatient violence: "actual physical acts against other patients, staff or objects and any verbal threat" (Ogloff & Daffern, 2002). The Historical Clinical Risk Management-20 (HCR-20) is a risk assessment tool used for violence prediction in a wider range of settings; the HCR-20 definition of violence excludes property damage and defines violence as "actual, attempted or threatened harm to a person or persons. Threats of harm must be clear and unambiguous. Violence is behaviour that obviously is liable to cause harm to another person or persons. Behaviour which would be fear-inducing to the average person may be counted as violence" (Douglas, Hart, Webster, & Belfrage, 2013; Webster, Douglas, Eaves, & Hart, 1997).

Violence in inpatient psychiatric wards is a major problem facing national health services. Nationwide surveys in the UK have reported that over three-quarters (78%) of psychiatric nursing staff have been subject to inpatient violence, threats or have been made to feel unsafe (Chaplin, McGeorge, & Lelliott, 2006). Reported rates of such victimization were lower for service-users themselves (37%), other clinical staff (44%) and other non-clinical staff (33%).

Being victim to a violent attack by a patient has been found to have the capacity to cause short- or long-term physical (Bowers, Allan, Simpson, Nijman, & Warren, 2007; Daffern & Howells, 2002) and psychological (Caldwell, 1992; Inoue, Tsukano, Muraoka, Kaneko, & Okamura, 2006; Lauvrud, Nonstad, & Palmstierna, 2009; Richter & Berger, 2006; Wildgoose, Briscoe, & Lloyd, 2003) damage to the individual. These effects can be immediate and related to a single incident, but can also arise cumulatively from frequent experience of aggressive incidents; significantly lower levels of general psychiatric health have been found in nurses who reported "frequent" exposure to inpatient violence (Wildgoose et al., 2003). Previous research has also found that, following assaultive incidents, 10-22% of responding psychiatric staff met the DSM-III-R criteria for Post-Traumatic Stress Disorder and over 20% of registered psychiatric nurses reported re-experiencing of intrusive memories (Caldwell, 1992; Inoue et al., 2006; Lauvrud et al., 2009; Richter & Berger, 2006). Patients who are the victims of violence from fellow service-users can also find the effects to be damaging (Daffern & Howells, 2002); being the target of violence could, for example, reinforce paranoid delusions about other people trying to harm the patient, serving as an obstruction to treatment which may cause a deterioration in symptoms or a severe psychotic episode. More generally, it has been shown that inpatient violence can disrupt the therapeutic environment of the ward and may also impose a financial strain on the institution (Daffern & Howells, 2002).

A review of the literature on the prevalence of psychiatric inpatient violence, with a total sample of 69,249 patients, found that 32.4% of inpatients had been violent or aggressive one or more times and there was a lot of variation in prevalence between countries and between settings. The highest rates were reported for forensic wards, with 48% of patients being violent; this was nearly double that for acute wards (26%) and over double that for general psychiatric hospitals (22%). For rates of events per 100 patients, forensic wards were again the highest (411), over three times that for general psychiatric hospitals (121) and nearly six times that of acute wards (72) (Bowers et al., 2011).

Forensic wards are reserved for individuals with a history of criminal, and often violent, charges or convictions, as well as individuals who have been transferred from other wards due to their difficult behaviour. As such, it is not unexpected that rates of violence are higher on forensic wards than on other psychiatric wards. Forensic patients are a complex population, sharing a mixture of characteristics with individuals in prison populations and general psychiatric wards; they often have multiple comorbid diagnoses (American Psychiatric Association, 2013) and many have experienced a period of incarceration following criminal activity. Forensic inpatients tend to have similar psychosocial backgrounds to those in correctional settings; instability during childhood, substance abuse, dysfunctional networks of social support and previous instances of interpersonal aggression are common (Ferguson et al., 2005).

Given the high prevalence of violence on forensic psychiatric wards and the multitude of potential negative consequences of inpatient violence for both patients and staff, it is important for health professionals to be able to predict and pre-empt potential violent behaviour from high-risk patients. Predicting violence in inpatient settings enables staff to prepare for the appropriate management or prevention of violent incidents to avoid harm to others in the ward environment. Furthermore, a lack of preparation for violent incidents can lead to staff being less willing to work with patients who have the potential to be aggressive; this could ultimately lead to a shortage of supporting staff available to help those in great need (Anderson, Bell, Powell, Williamson, & Blount, 2004). Therefore, the assessment of patients' risk of violence has become routine practice on forensic wards.

1.2 History of violence risk assessment

Structured violence risk assessment has been one approach to tackle the high prevalence of violent incidents. However, the field has been limited by the need for improved information about the risk factors and predictors of violent behaviour in inpatients.

1.2.1 Risk factors for inpatient violence

Although not the focus of this review, the identified risk factors for violence warrant discussion as they are the basis for the development of violence risk assessment instruments. The literature on violence risk factors distinguishes between static and dynamic risk factors. Static variables refer to those contributory factors that are unchangeable over time; examples of static factors include age, gender, prior violent convictions, childhood abuse and history of substance abuse. These unchanging factors will always contribute a certain amount of risk to an individual's overall estimate of violence risk. Dynamic factors, however, are malleable and can change on a day-to-day basis; examples include current mental state (for example, psychosis), current substance abuse, adherence to medication, social support and irritability. Both static and dynamic factors have been shown to contribute to an individual's risk of violence (Douglas & Skeem, 2005; McDermott, Edens, Quanbeck, Busse, & Scott, 2008; Witt, Van Dorn, & Fazel, 2013).

1.2.1.1 Static factors

Investigation of static factors in relation to violence risk has a long history. Three static factors (history of antisocial behaviour, antisocial personality patterns and antisocial associates) are included in the four variables that are termed the "big four" criminogenic risk factors for offenders (Andrews & Bonta, 1994) and all three have received further support as validated static risk factors for violence and offending (Wang & Diamond, 1999; Witt et al., 2013). Psychopathy has also been repeatedly identified as a significant risk factor for future aggression and violence (Douglas, Vincent, & Edens, 2006; Gendreau, Goggin, & Smith, 2002; Grann, Långström, Tengström, & Kullgren, 1999; Tengström, Grann, Långström, & Kullgren, 2000). Other static risk factors identified for violence include history of assault, history of

polysubstance misuse, violent victimization in adulthood, history of imprisonment for any offence and recent arrest for any offence (Witt et al., 2013). However, static risk factors for violence are now recognised as less clinically useful; as they cannot be changed, they do not lead to the development of potential interventions.

1.2.1.1 Dynamic factors

A number of dynamic factors have been investigated in relation to violence risk. One review identified dynamic risk factors in the empirical literature based on two key criteria: the factors must be related to violence and they must change over time (Douglas & Skeem, 2005). An example of a dynamic factor considered to be a strong predictor of future aggression and violence is impulsiveness (Douglas & Skeem, 2005; Grisso, Davis, Vesselinov, Appelbaum, & Monahan, 2000; Monahan et al., 2001; Wang & Diamond, 1999), primarily defined as a lack of control over one's emotion, cognitive processing or behaviour. Although some consider impulsiveness to be a steady dispositional trait, it can change over time and is listed as a symptom for several mental disorders in the DSM-V (American Psychiatric Association, 2013). Other dynamic factors associated with violence and aggression risk include anger (Kay, Wolkenfeld, & Murrill, 1988; Menzies & Webster, 1995; Wang & Diamond, 1999), psychosis (Arango, Calcedo Barba, González-Salvador, & Calcedo Ordóñez, 1999; Bartels, Drake, Wallach, & Freeman, 1991; Cheung, Schweitzer, Crowley, & Tuckwell, 1997; McNiel & Binder, 1994; Swanson, Borum, Swartz, & Monahan, 1996), antisocial attitudes (Andrews & Bonta, 1994; Gendreau, Goggin, & Law, 1997; Gendreau, Little, & Goggin, 1996), substance misuse (Lipsey, Wilson, Cohen, & Derzon, 2002; Monahan et al., 2001; Swanson, Holzer III, Ganju, & Jono, 1990; Witt et al., 2013) and treatment non-adherence (Monahan et al., 2001; Schwartz et al., 1998; Witt et al., 2013).

1.2.2 Violence risk assessment methods

For many decades, violence risk assessment involved evaluating the risk of violence in inpatients based on the judgement of a single clinician or clinical team. This raised issues concerning whether all risk factors were being considered with sufficient care. After some time, many researchers and clinicians began to see this unstructured clinical method of risk assessment as insufficient for the appropriate prediction of violent outcomes. Furthermore, clinical judgement does not maintain a uniform level of predictive accuracy across all services, as prediction outcomes are entirely dependent on the particular clinician making the judgement.

As an alternative to unstructured clinical risk prediction, an approach based on statistical probabilities grew in popularity, leading to the development of actuarial risk assessment tools. Findings from empirical investigations into risk factors for violence were incorporated into structured instruments that could be administered to patients to give an overall probabilistic estimate of violence risk over a specified time period. The use of actuarial instruments results in consistent assessment methods and considered criteria both across services and those administering the assessment. However, one criticism of actuarial tools is that they do not take into account particular elements of personality or nuances in behaviour that may be detected by a clinician who is familiar with the individual.

This critique has led to the development of structured professional judgement (SPJ) instruments, which aim to combine the useful elements of clinical judgement and actuarial instruments. SPJ tools allow for a clinical judgement of overall violence risk to be made, after taking into account a series of items (as actuarial tools do) that are scored as present or absent for the individual being assessed. The extent of detail involved in the process of clinical judgement can vary greatly; some tools require a judgement of low, moderate or high risk categorisation taking into account the individual's actuarial score whilst others require a more substantial process of formulation, wherein the prior circumstances of violent behaviour are analysed and a future context in which violence could arise is hypothesised. In these more detailed clinical formulations, the risk categorisation of low, moderate or high becomes less

relevant, as the formulation process itself produces a plan for action and management to monitor risk. This has been seen as a benefit in terms of shifting the literature from a focus on the prediction of violence alone, to a collective assessment of violence risk along with planning management strategies (Whittington et al., 2013).

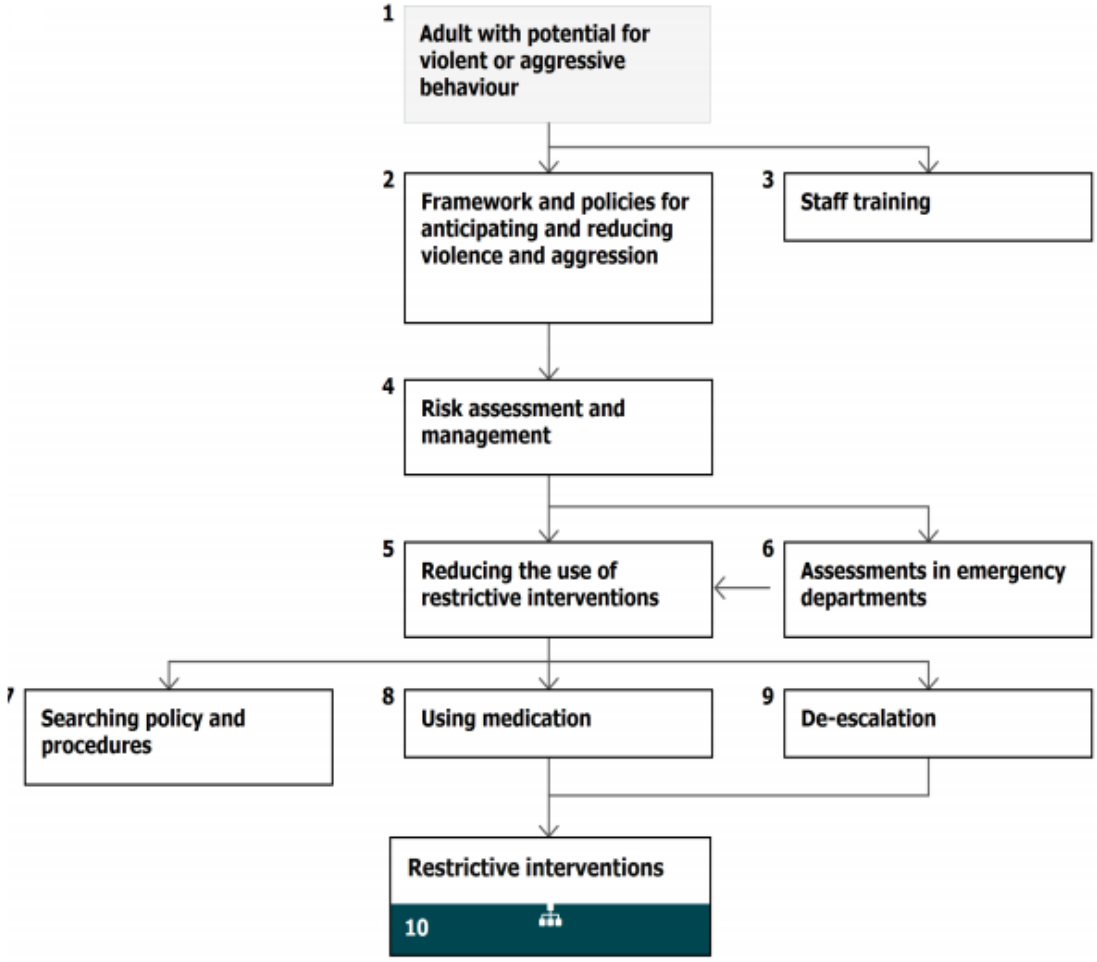
SPJ instruments are presented as attempting to bridge the gap between the clinical and statistical approaches to risk assessment; however, they have also faced criticism, as some hold the belief that the decision of which information to include, exclude or weight more heavily should not be at the discretion of the assessor as it is too subjective. SPJ instruments with a structured formulation process also require more time, energy and resources; often, specific and costly training is necessary, which is less of an issue for actuarial assessment. Therefore, in order to justify the extensive efforts and resources that must be expended on the use of SPJ instruments, it is necessary for them to produce accurate results. Additionally, it may be difficult to accurately assess the findings from studies of SPJ tools; limitations may arise due to poor translation from research to clinical practice caused by the subjective element of judgement. This applies less to actuarial tools, as their accuracy estimates in research contexts are likely to be similar to how well they will perform in a clinical setting.

Both static and dynamic items can be incorporated into actuarial or SPJ tools. Violence risk assessment instruments vary in what factors they include; some include a single type, whilst others draw on a combination of both. For the case of forensic inpatients in particular, evidence suggests that dynamic factors are more relevant for the prediction of inpatient violence and the imported risk from static and historical factors has a weaker association with inpatient violent behaviour (McDermott et al., 2008).

1.3 Current practice

The existing guidelines from the National Institute for Health and Care Excellence (NICE) for the short-term management of violence and aggression in mental health settings

(NG10; May, 2015) recommend a multidisciplinary approach and the use of an actuarial prediction instrument for the assessment of violence risk, rather than unstructured clinical judgement alone. The Brøset Violence Checklist (BVC) (Almvik & Woods, 1998; Linaker & Busch-Iversen, 1995) and the Dynamic Appraisal of Situational Aggression-Inpatient Version (DASA-IV) (Ogloff & Daffern, 2002), both of which are actuarial instruments designed to assess violence risk in inpatient settings, are specifically recommended in the guidelines. Regular re-evaluation of risk assessment and risk management plans is also suggested, with the frequency of this depending on the risk level of the individual. Figure 1 shows a flow chart for anticipating,



reducing the risk of and preventing violence and aggression in adults, taken from NICE NG10 guidelines (May, 2015).

Figure 1 Flow chart for anticipating, reducing the risk of, and preventing violence and aggression in adults (NICE guidelines, NG10; May, 2015)

The National Health Service standard contract guidelines for adult medium and low secure services (NHS, United Kingdom; 2013) state that the assessment of violence risk should be evidence-based and regular in frequency. It is also stated that there should be dynamic risk assessment models implemented in services for support in day-to-day decisions about individual patient care. The guidelines emphasise that the implementation of security measures for forensic inpatients should be based on the risk needs of the individual and should be imposed at the lowest level of restriction possible and only when risks have been identified. This highlights the importance of risk prediction being as accurate as possible to ensure that unnecessary security measures are not applied; this can be a waste of resources and may also interfere with patient recovery and the maintenance of a therapeutic environment. It is required that an initial risk assessment is completed within the first three months of admission, with subsequent repeated, on-going risk assessment then necessary in order for the patient to progress along the forensic care pathway towards eventual release. The instruments recommended by the NHS are the Historical, Clinical, Risk Management 20 (HCR-20) (Douglas et al., 2013; Webster et al., 1997) and the Short-Term Assessment of Risk and Treatability (START) (Webster, Martin, Brink, Nicholls, & Middleton, 2004; Webster, Martin, Brink, Nicholls, & Desmarais, 2009), both of which are SPJ tools. The HCR-20 has a detailed, structured formulation process, whilst the START's clinical judgement section is more basic, with a judgement of low, moderate or high risk for a number of adverse outcomes.

Although both of these sets of guidelines apply only to regions within the United Kingdom and other countries will differ in their guidelines for health professionals, it is likely that the general recommendations (rather than specific tool suggestions) are cross-culturally valid; other countries may have regular violence risk assessment as a requirement for patients in forensic psychiatric services and may have particular locally recommended tools.

1.4 Scope of research to date

A common theme within existing violence risk assessment literature is an emphasis on violent behaviour in a community setting, including antisocial behaviour, violent offending, violent reconviction, sexual offending, sexual reconviction, general criminal offending and general criminal reconviction. Conversely, little attention has been given to research into risk assessment of inpatient violence, perhaps due to the notion that the safety of the public is deemed of greater importance than the risks posed to patients and staff working in secure hospitals.

Furthermore, much of the literature has lacked specific focus; inpatient or institutional violence is often grouped together with community, offending or reconviction outcomes as an overall violent outcome. This lack of distinction between types of violent outcome is particularly common in literature reviews and meta-analyses (Fazel, Singh, Doll, & Grann, 2012; Singh, Grann, & Fazel, 2011; Whittington et al., 2013).

A further related tendency of the literature on violence risk assessment is to combine samples from multiple populations; the forensic population shares characteristics with both the general psychiatric and prison populations, it is often combined with these populations and not treated separately (Campbell, French, & Gendreau, 2009; Fazel et al., 2012; Singh et al., 2011; Whittington et al., 2013). The forensic psychiatric population in most countries is also significantly smaller than both the general psychiatric population and prison populations, which may be another reason it is often overlooked as an individual group; however, it is the unique combination of characteristics from both psychiatric and prison populations that makes the forensic inpatient population of particular interest and worthy of independent investigation. Furthermore, as previously discussed, rates of violence are higher on forensic wards than other psychiatric wards and investigations have found that having a status of not criminally responsible by reason of mental disorder (NCRMD) is a risk factor for institutional violence (Quinsey, Harris, Rice, & Cormier, 2006b); this indicates that within a forensic ward, where

many patients are detained by a specific mental health law, there must be other relevant factors at play.

There have been a number of systematic reviews and meta-analyses of the predictive accuracy of violence risk assessment instruments; however, these reviews have included a range of combined populations or types of violent outcomes (Campbell et al., 2009; Fazel et al., 2012; Singh et al., 2011; Whittington et al., 2013). This reluctance to focus on particular contexts and populations was highlighted by a meta-review of the systematic reviews and meta-analyses of violence risk assessment; 90% of reviews published before 2010 with a clearly defined population included a mixed sample of populations (for example, prison populations, forensic inpatients and forensic outpatients all together) (Singh & Fazel, 2010).

There are exceptions which have focused on a particular context and population. A previous review of inpatient violence in forensic psychiatric populations gave a brief overview of the literature available (Hogan, Ennis, & Assessment, 2010). Whilst this review included clear, limiting inclusion criteria that allowed a focus on the forensic population in inpatient settings, the methodology used was problematic. The use of mean correlation coefficients between violence risk assessment scores and inpatient violence is insufficient for a complete investigation of predictive accuracy; furthermore, this review only included three commonly-used violence risk assessment instruments (the HCR-20, PCL-R and PCL:SV). Therefore, despite some effort, there remains a gap in the literature requiring a comprehensive systematic review and meta-analysis of the accuracy of risk assessment instruments for the prediction of forensic inpatient violence.

1.5 Aims of this review

This study aims to review and summarise the existing literature investigating the efficacy of risk assessment instruments used to predict inpatient violence in the forensic psychiatric population. There are three main questions to address:

1. How accurate are the most commonly used risk assessment instruments for the prediction of inpatient violence on forensic psychiatric wards?
2. Which tools (or types of tools) should be used for forensic inpatient populations in order to maximise the accuracy and efficacy of risk prediction?
3. Are there any study- or sample-related factors that affect the performance of violence risk assessment tools when used to predict forensic inpatient violence?

It is also of interest to examine the existing literature and identify any areas that need more research or any improvements that can be made to methodologies or reporting of findings in order to gain a more comprehensive picture of the accuracy of violence risk assessment tools for the population of forensic patients and the outcome of inpatient violence.

Chapter 2: Methods

2.1 Review protocol

This review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement (Moher, Liberati, Tetzlaff, Altman, & Group, 2009); see Appendix 1 for PRISMA Checklist. A review protocol was published on PROSPERO, an international prospective register of systematic reviews, on 23/11/16 and can be accessed at https://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016049789 or by searching using the registration code "CRD42016049789".

2.2 Risk assessment tools

Using a number of recent reviews and questionnaire surveys (Hurducas, Singh, de Ruiter, & Petrila, 2014; Singh, Desmarais, et al., 2016; Singh et al., 2014), the eleven most commonly used instruments for forensic inpatient violence risk prediction were identified. Actuarial instruments included the Brøset Violence Checklist (BVC) (Almvik & Woods, 1998; Linaker & Busch-Iversen, 1995), the Classification of Violence Risk (COVR) (Monahan et al., 2005; Monahan et al., 2001), the Dynamic Appraisal of Situational Aggression (DASA) (Ogloff & Daffern, 2002), the Level of Service Inventory-Revised (LSI-R) (Andrews & Bonta, 1995), the Psychopathy Checklist Revised (PCL-R) (Hare, 1991), the Psychopathy Checklist Screening Version (PCL:SV) (Hart, Cox, & Hare, 1995), the Violence Risk Appraisal Guide (VRAG) (Quinsey, Harris, Rice, & Cormier, 2006a; Quinsey et al., 2006b) and the Violence Risk Scale (VRS) (Wong & Gordon, 2000). Structured Professional Judgement (SPJ) tools included the Historical Clinical Risk Management-20 (HCR-20) (Douglas et al., 2013; Webster et al., 1997), the Short-Term Assessment of Risk and Treatability (START) (Webster et al., 2004; Webster et al., 2009) and

the Violence Risk Screening-10 (V-RISK-10) (Bjorkly, Hartvig, Heggen, Brauer, & Moger, 2009; Hartvig et al., 2007). Studies testing the predictive validity of unstructured clinical judgement were also searched for, to provide a point of comparison for discussion.

Tools developed specifically for sexual violence (e.g. Sex Offender Risk Appraisal Guide, SORAG (Quinsey et al., 2006a, 2006b)) were excluded. Although the PCL-R and PCL:SV were not developed for risk assessment purposes, but rather to diagnose psychopathic personality disorder, they have become commonly used in inpatient settings to assess risk; numerous studies have found the PCL-R score to have moderate to strong associations with violent, criminal and antisocial outcomes (Grann et al., 1999; Gray et al., 2003; Nicholls, 1997; Tengström et al., 2000) and therefore they warranted inclusion. Although some of these instruments are used for the assessment of both inpatient violence and violent recidivism on release, this review only focuses on the outcome of violence in inpatient settings.

2.2.1 HCR-20

The Historical Clinical Risk Management-20 (HCR-20) (Douglas et al., 2013; Webster et al., 1997) is a structured professional judgment (SPJ) tool and is used in a diverse range of services. It involves ratings of prevalence and relevance for 10 historical factors, 5 clinical factors and 5 risk management factors. These are followed by risk formulation and case formulation sections to be filled in by a clinician familiar with the patient. The individual is scored out of 40 from the prevalence and relevance ratings and an overall risk judgement is formed, placing the individual into the low, moderate or high risk category. Since its initial publication, the HCR-20 has been translated into 20 languages and has been used in over 35 different countries. The most recent version, HCR-20 Version 3 (Douglas et al., 2013), has been developed based on field testing and empirical evaluation and is most commonly used in correctional, forensic and general or civil psychiatric settings (Singh, Bjorkly, & Fazel, 2016).

2.2.2 PCL-R and PCL:SV

The Psychopathy Checklist Revised (PCL-R) (Hare, 1991) and Psychopathy Checklist Screening Version (PCL:SV) (Hart et al., 1995) were originally designed for the purpose of detecting psychopathy, rather than for violence risk assessment specifically. However, they are both widely used in psychiatric services for the prediction of violence as high psychopathy scores have been shown to be good predictors of future violence (Douglas & Skeem, 2005; Gendreau et al., 2002; Grann et al., 1999; Tengström et al., 2000). The PCL-R is a 20-item checklist, while the PCL:SV is a 12-item checklist. Each item is rated on a 3-point scale (0, 1 or 2) and total scores given out of 40 and 24 respectively. The PCL-R is also often incorporated into other risk assessment instruments (e.g. VRAG) and has been translated into 14 languages. It is recommended for use with forensic and correctional populations, with evidence for its applicability to adult males, females, adolescents and sex offenders. The PCL:SV was developed as a brief version of the PCL-R for use in civil or forensic settings in order to determine whether or not to then administer the PCL-R for more detailed information and the potential detection of psychopathy. The PCL:SV has been designed in such a way that it can be easily integrated with the PCL-R, leading to enhanced accuracy.

2.2.3 START

The Short-Term Assessment of Risk and Treatability (START) (Webster et al., 2004; Webster et al., 2009) is an SPJ tool designed to predict aggression or violence over the short-term period of weeks to months. The START tool has two 22-item scales – Strengths and Vulnerabilities – each scored as 0, 1 or 2. This is followed by a section for specific risk estimates for a number of adverse outcomes, each of which is rated as high, moderate or low. The START aims to measure risk across seven domains: violence to others, suicide, self-harm, self-neglect, unauthorised absence, substance abuse and victimization. The START was created for use with

correctional, civil and forensic patients in community or institutional settings. It has been translated into 8 languages and implemented in 15 countries globally.

2.2.4 DASA

The Dynamic Appraisal of Situational Aggression (DASA) (Ogloff & Daffern, 2002) is a short-term risk assessment tool. The DASA consists of 7 items: irritability, impulsivity, unwillingness to follow instructions, sensitive to perceived provocation, easily angered when requests are denied, negative attitudes and verbal threats. Each item is scored as present (1) or absent (0) based on behaviour over the past 24 hours, giving a maximum total score of 7. The DASA was designed primarily for use in an inpatient setting and is predominantly used with forensic patients. The DASA is recommended for use in the short-term management of violence and aggression in NICE guidelines NG10 (May, 2015).

2.2.5 BVC

The Brøset Violence Checklist (BVC) (Almvik & Woods, 1998; Linaker & Busch-Iversen, 1995) is another short-term risk assessment tool. The BVC has 6 items: confused, irritable, boisterous, physically threatening, verbally threatening and attacking objects. Each is scored as present (1) or absent (0) based on behaviour over the preceding 24-hour period, giving a maximum score of 6. The BVC was developed for inpatient assessment and is used on forensic and acute wards. The BVC is also recommended for use in short-term management of violence and aggression in NICE guidelines NG10 (May, 2015).

2.2.6 VRAG

The Violence Risk Appraisal Guide (VRAG) (Quinsey et al., 2006a, 2006b) is a 12-item actuarial tool, with possible total scores ranging from -24 to 38. The VRAG was developed with the aim of assessing risk of violent recidivism amongst correctional and forensic populations and primarily those with a history of criminal violence. However, it is also occasionally used in inpatient settings. The PCL-R is incorporated as an item in the VRAG assessment and all other factors included in the VRAG are static. A version of the VRAG, called the SORAG (Sex Offender Risk Appraisal Guide) has been developed for the assessment of sex offenders.

2.2.7 VRS

The Violence Risk Scale (VRS) (Wong & Gordon, 2000) is an actuarial instrument with 6 static and 20 dynamic factors, each of which are rated as 0, 1, 2 or 3, giving a total score out of 78. The VRS was designed to predict risk of violent recidivism, but is also occasionally used in inpatient settings. Versions of the VRS have since been developed for sex offender (VRS-SO) and youth (VRS-YV) populations.

2.2.8 COVR

The Classification of Violence Risk (COVR) (Monahan et al., 2005; Monahan et al., 2001) is an interactive software that guides the user through a chart review and a short interview with the patient. It was created for the purpose of determining appropriateness of release into the community, but is also occasionally used in inpatient settings. The COVR uses an interactive "classification tree" method, which allows an individualised assessment for each client as the questions asked depend on the answers given to prior questions. The output is a percentage risk between 1% and 76%. The COVR was originally designed to estimate violence risk in acute

civil psychiatric patients after discharge, but has also been used in the assessment and prediction of inpatient violence.

2.2.9 LSI-R

The Level of Service Inventory Revised (LSI-R) (Andrews & Bonta, 1995) is an actuarial tool primarily used in correctional facilities but also with some use in psychiatric settings. The LSI-R consists of 10 "domains" and scores can range from 0 to 54. The LSI-R was developed for the prediction of criminal recidivism with the goal of guiding release and probation decisions, but it has seen some use in inpatient settings with forensic patients.

2.2.10 V-RISK-10

The Violence Risk Screening-10 (V-RISK-10) (Bjorkly et al., 2009; Hartvig et al., 2007) is an SPJ instrument designed for screening of violence risk in acute and general psychiatric services. The tool was developed in response to the demand for an instrument for acute and general psychiatry that was less time-consuming and required less expertise and training than the existing forensic tools. The aim of this tool is to screen individuals and find patients who require a more thorough violence risk assessment. The V-RISK-10 consists of 10 items which are marked as "yes", "maybe", "no" or "don't know". An overall clinical risk judgment is formed and individuals are ranked as low, moderate or high risk and it is indicated whether further risk assessment or preventive measures are required.

2.3 Systematic search

A systematic search was conducted to identify studies that measured the predictive validity of the eleven instruments, specifically in forensic psychiatric settings, for the outcome of

inpatient violence. The databases searched were PsycINFO (1806 – January 2017), PubMed (1809 – January 2017), Embase (1974 – January 2017), CINAHL (1937 – January 2017) and Global Health (1973 – January 2017) for relevant studies, using a key word search of titles and abstracts. Below is an example of the search terms used (specifically for the search conducted on PsycINFO):

((PCL-R OR Psychopathy Checklist Revised OR HCR-20 OR Historical Clinical Risk Management OR PCL:SV OR PCL-SV OR Psychopathy Checklist Screening OR VRAG OR VRAG-R OR Violence Risk Appraisal Guide OR COVR OR Classification Violence Risk OR VRS OR VRS-2 OR Violence Risk Scale OR LSI-R OR Level Service Inventory OR START OR Short Term Assessment Risk Treatability OR BVC OR Br?set Violence Checklist OR DASA OR Dynamic Appraisal Situational Aggression OR V-RISK-10 OR Violence Risk Screening 10 OR risk assess) and inpatient* and violen* and risk and (predict* OR valid*)).ab,ti*

Additional studies were identified through hand-searching references of the identified studies, using the Google Scholar "cited by" function, scanning the annotated bibliographies for each instrument and corresponding with researchers in the field. Studies in all languages, across all dates and those that were unpublished were all considered for inclusion. Studies were excluded if they only measured the predictive validity of select scales of the tools, if they only focused on a specific subgroup of the forensic population (namely, those with a sole diagnosis of learning disability), if instruments were coded retrospectively without blinding to outcomes or if they were calibration studies for the actuarial tools (which may give inflated effects). When studies were using overlapping samples, the sample with the larger number of participants was used in order to avoid double-counting.

Using this search strategy, a total of 448 publications were identified, with 324 remaining after removing duplicates. The inclusion criteria required that the study (a) measured the predictive accuracy of the tool, (b) had a forensic psychiatric sample (disaggregated from correctional or general psychiatric samples) and (c) measured the outcome

of inpatient violence (disaggregated from violence in community settings). 220 texts remained after screening abstracts using these criteria. Full texts were subsequently screened, leaving 52 eligible publications for inclusion; see Figure 3 for study selection flow chart.

Further inclusion criteria were added in order to obtain a range of performance measures for these tools. To be included in the meta-analysis, studies had to report rates of true positives, false positives, true negatives and false negatives at a given tool-specific cut-off score for the outcome of inpatient violence or aggression. Study authors were contacted if this data was unavailable in the manuscript and they were asked to fill in a standardised table, shown in Figure 2.

Figure 2 Exemplar "2x2 table" sent to authors as request for standardised outcome data

	Number who were violent	**Number who were not violent**
Number Classified as High Risk (score = > x)	True Positives	False Positives
Number Classified as Moderate Risk (score = x - y)		
Number Classified as Low Risk (score = < y)	False Negatives	True Negatives

The numbers of true positives, false positives, false negatives and true negatives for each study were derived based on a cut-off score for test positivity for each tool. Cut-off scores were defined using those recommended in the manual of the tools or (if this was unavailable) using the most-commonly used cut-off scores in the literature identified for each particular tool. In the case of the START tool, there were no cut-off scores and the violence risk estimate categorisation of low, medium or high was the outcome measure used. The details of each risk assessment instrument are listed in Table 1.

Instruments used for violence risk assessment in a clinical setting aim to identify individuals that pose a higher risk and may need to be monitored. Therefore, subjects who were classified as moderate or high risk for inpatient violence were combined and compared with those classified as low risk to create the binary outcomes from the "2x2" table (Figure 2).

Table 1 Characteristics of the nine included risk assessment instruments

Instrument type and name	No. of items	Static or Dynamic Items	Cut-off scores [a]	
Actuarial				
BVC	6	All dynamic	*High*	≥ 3
			Low	< 3
COVR	- *	Mainly static	*High*	≥ 26
			Moderate	8 - 26
			Low	< 8
DASA	7	All dynamic	*High*	≥ 4
			Low	< 4
PCL-R	20	Mainly static	*High*	≥ 25
			Moderate	15 -24
			Low	< 15
PCL:SV	12	Mainly static	*High*	≥ 15
			Low	< 15
VRAG	12	All static	*High*	≥ 14
			Moderate	-7 – 13
			Low	< -8
VRS	26	Both	*High*	≥ 42
			Low	< 42
Structured professional judgement				
HCR-20	20	Both	*High*	≥ 30
			Moderate	20-29
			Low	< 20
START	44	Both	- **	

[a] Information on cut-off scores relates only those samples who reported a cut-off score; in some cases cut-off scores were unknown or a clinical risk judgement may have been used instead
* COVR has a varying number of items depending on answer given to previous item
** No cut-off score was used for START classifications, as the low, moderate and high risk categorisation was given from the violence risk estimate section

Figure 3 shows a flow chart diagram of the study selection process. Of the 52 eligible publications, 43 gave a valid overall performance measure (the area under the curve value), so median summary AUC values are reported for all publications that reported it. This gave 78 available samples from the 43 publications.

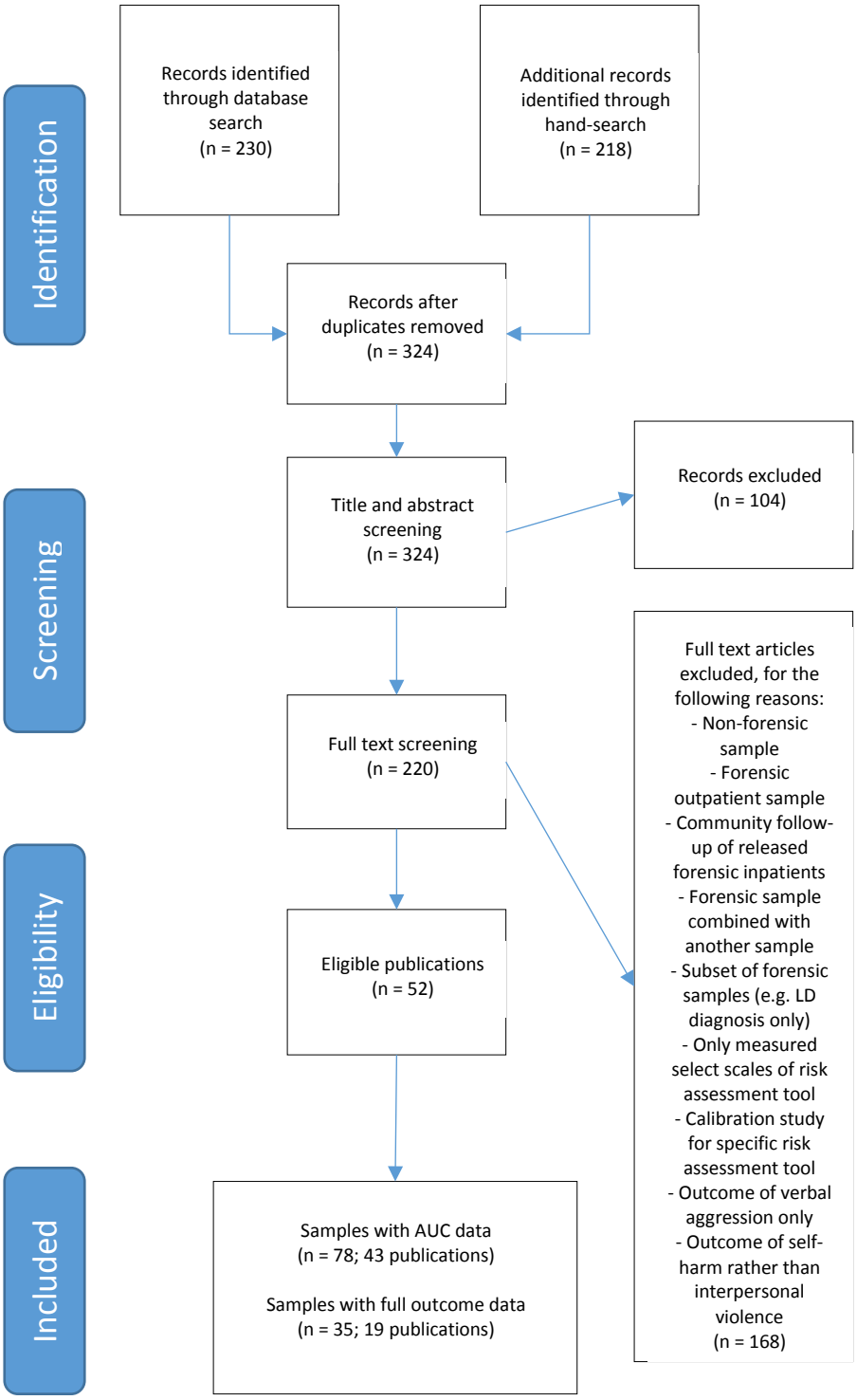


Figure 3 Flow chart diagram displaying the study selection process

The desired full range of outcome data were available in the manuscripts of 11 eligible publications. Further data was requested from the authors of the other 41 publications and data was obtained for 8. Studies for which tabular data could not be obtained were excluded from the meta-analysis; the LSI-R and V-RISK-10 were excluded as no eligible studies assessing these instruments responded.

The number of publications finally included was 19; however, some investigated multiple instruments and/or reported separated data for males and females, resulting in a total of 35 available samples to include in the meta-analysis from the 19 publications.

2.4 Quality assessment

A number of quality assessment tools, such as the QUIPS and QUADAS-2 tools, have been designed particularly for primary studies included in a systematic review and meta-analysis of studies of diagnostic or prognostic test accuracy. The QUADAS-2 tool was chosen for this review as it includes a mechanism of visualisation through flow-chart construction and the QUIPS tool is designed specifically for studies assessing factors within prediction models which is not applicable to the studies included in this review.

The QUADAS-2 tool (Schueler, Schuetz, & Dewey, 2012; Whiting et al., 2011) is recommended by the Cochrane Collaboration and was used to gain a comprehensive idea of the risk of bias for each study. The tool consists of three phases, the first of which requires a summary of the review question and to tailor the QUADAS-2 tool to the review in order to produce review-specific guidance. The second phase involves constructing a flow diagram of the primary study being assessed; this involves patient group-allocations and outcomes. The final section, where risk of bias and concerns regarding applicability are assessed, is of particular importance. This section covers four main domains. The first is patient selection, ensuring an unbiased method of sample recruitment. The second concerns the index test which, for the purposes of this review, is the risk assessment tool in question, how it was administered and

who it was administered by. The third domain covers a reference standard, which, in this study, refers to the outcome of inpatient violence, how this was defined and how it was recorded. The final domain covers flow and timing; flow refers to the reasons for excluding any patients and timing refers to the length of follow-up period used to observe and record violent behaviour following the risk assessment being administered. Each of these domains is assessed in terms of risk of bias and the first three are also assessed with respect to concerns regarding applicability. To aid risk of bias judgements, signalling questions are used in the QUADAS-2 tool. A final judgement is made for each domain regarding whether there is low, moderate or high risk of bias and similarly whether there are low, moderate or high levels of concern regarding applicability. All of the included studies were assessed using this tool and none gave any moderate or high risk of bias or concerns regarding applicability; see Appendix 2 for the QUADAS-2 tool.

2.5 Data extraction

A spreadsheet was compiled to extract data for the main meta-analysis and for further subgroup analyses. This was based on the CHARMS checklist (Moons et al., 2014), which is a recommended checklist for the relevant items to extract from primary studies in a systematic review of prediction models (see Appendix 3). Data was extracted on the primary summary measures (numbers of true positives, false positives, false negatives and true negatives, sensitivities, specificities, positive predictive values, negative predictive values, diagnostic odds ratios and area under the curve values), as well as information on year of study, location of study, study design, recruitment method, sample size, gender distribution of sample, mean age of sample, type of sectioning, psychiatric diagnoses, type of sectioning, criminal history, follow-up period, definition of violent outcome, method of recording violent outcome and risk assessment instrument being used. The primary outcome of the studies included in this meta-analysis was interpersonal inpatient violence *or* interpersonal violence and verbal threat as an

inpatient, depending on how violence was recorded in the study. There were no secondary outcomes.

Data was extracted by myself and a secondary extractor, Dr. Artemis Igoumenou, a forensic psychiatrist who, at the time, was working as a clinical researcher at Queen Mary University of London.

2.6 Data analysis

2.6.1 Instrument groupings

The nine risk assessment tools were divided into two categories for the analysis. One category contained instruments designed to predict imminent violence over a period of 24 hours. The other category contained those designed to predict violence over an unspecified period, longer than 24 hours. The instruments in the imminent violence prediction category were the BVC and the DASA, both of which are designed to predict aggression and violence over a very short-term 24-hour follow-up period. The other tools are all designed to predict violence over a longer period of time so were placed in the "longer-term" prediction category of instruments.

2.6.2 Performance measures

This review followed the current guidance provided by the Cochrane collaboration for systematic reviews of diagnostic and prognostic test accuracy (Macaskill, Gatsonis, Deeks, Harbord, & Takwoingi, 2010). Statistical analyses for reviews of diagnostic test accuracy primarily revolve around two measures of diagnostic accuracy: sensitivity and specificity. The sensitivity, in the context of this study, is a measure of the proportion of violent patients who were correctly identified as higher risk by the risk assessment tool. The specificity is the

proportion of non-violent patients who were correctly identified as low risk by the risk assessment tool. These two statistics are calculated using the numbers of true positives, false positives, true negatives and false negatives for each study. The number of true positives (TP) is the number of violent patients who were correctly classified as higher risk. The number of false positives (FP) is the number of non-violent patients who were incorrectly classified as higher risk. The number of false negatives (FN) is the number of violent patients who were incorrectly classified as low risk. Finally, the number of true negatives (TN) is the number of non-violent patients who were correctly classified as low risk. These can be visualized from the "2x2" table in Figure 2.

Sensitivity and specificity are calculated as follows:

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

The aim of analysis was to both quantify and compare the sensitivity and specificity statistics as well as a number of other accuracy estimates: positive and negative predictive values (PPV and NPV, respectively), the diagnostic odds ratio (DOR) and the area under the curve (AUC) value. Two measures that were considered for inclusion were the positive and negative likelihood ratios (LR+ and LR-, respectively). These refer to the likelihood that a given test result (e.g. categorization as low risk) would be expected in a patient with or without the "target disorder" (in this case, either a violent or non-violent patient) and are primarily used in diagnostic medical testing, where the individual is already known to have a target disorder. As this review focuses on the prediction of future violence, it is an irrelevant statistic and has therefore not be included.

The positive and negative predictive values describe the performance of the diagnostic test with a higher result indicating higher accuracy. The PPV is a measure of the proportion of patients predicted to be violent (higher risk) who were actually violent and is calculated by:

$$\boldsymbol{Positive\ Predictive\ Value} = \boldsymbol{PPV} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

The NPV is a measure of the proportion of patients predicted to be non-violent (low risk) who were actually non-violent and is calculated by:

$$\boldsymbol{Negative\ Predictive\ Value = NPV = \frac{TN}{TN + FN}}$$

The DOR is a measure of the effectiveness of a test and can range from zero to infinity. A DOR above one indicates a useful test and higher DORs reflect better test performance. The DOR is the ratio of the odds of patients who were violent having been predicted to be violent (higher risk) relative to the odds of patients who were not violent having been predicted to be non-violent (low risk). The DOR is calculated by either of the two below equations (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003):

$$\boldsymbol{Diagnostic\ Odds\ Ratio = DOR = \frac{TP/FP}{FN/TN}}$$

$$\boldsymbol{Diagnostic\ Odds\ Ratio = DOR = \frac{Sensivity/(1 - Sensitivity)}{(1 - Specificity)/Specificity}}$$

The area under the curve (AUC) value is derived from a receiver operating characteristics (ROC) curve. An ROC curve is a plot of sensitivity against (1 – specificity) at every possible cut-off threshold. A rising diagonal line of y=x indicates a test that has no informative utility and is "diagnosing" at chance level. Tests with higher accuracy result in an ROC curve approaching the top left-hand corner of the graph, as they would have both a high sensitivity and a high specificity. The AUC value is the area under the plotted ROC curve. Therefore, for uninformative tests, the AUC would be 0.5, while a perfectly accurate test would have an AUC value of 1. In general, an AUC value below 0.7 is interpreted as a poor predictive test, while an AUC of 0.8 or above is considered a good test (Tape, 2006).

During the study selection process there were 52 eligible publications with regards to the initial inclusion criteria, but 33 of these were not included in the meta-analysis due to a lack of the necessary standardised outcome data. Therefore, in addition to the primary analyses, the values for median AUCs were also calculated using data from studies within this group of 52 that reported an AUC value.

2.6.3 Meta-analysis model

For the meta-analysis of studies of diagnostic or prognostic test accuracy the data has two important characteristics. Firstly, there is expected to be a negative correlation between sensitivity and specificity; there is a trade-off between the two as the prediction threshold (i.e. tool cut-off score) varies (Deeks, 2001; Harbord, Deeks, Egger, Whiting, & Sterne, 2007; Moses, Shapiro, & Littenberg, 1993). Secondly, there is expected to be a considerable amount of between-study variation, unlike for the meta-analysis of data from randomised controlled trials, so this must be incorporated into models (Harbord et al., 2007; Lijmer, Bossuyt, & Heisterkamp, 2002). Previously proposed methods of analysis have included generating a summary receiver operating characteristic (SROC) curve using simple linear regression (Littenberg & Moses, 1993). However, this method is only approximate since the assumptions for linear regression are not fulfilled (Harbord et al., 2007) and there are doubts as to the appropriate weighting of the regression. Therefore, this method was discarded as an option for use in this review.

Methods that are more statistically rigorous have been put forward which, although more complex, overcome these problems and are more appropriate for use in this review. These methods are the bivariate random-effects model of analysis (Reitsma et al., 2005) and the hierarchical summary receiver operating characteristic (HSROC) model (Rutter & Gatsonis, 2001). These two hierarchical models jointly analyse pairs of sensitivities and specificities for each study and use the within-study binomial structure of the data whilst incorporating both within- and between-study heterogeneity. The bivariate and HSROC models are recommended and regarded as the optimal methods for producing summary statistics and summary ROC curves in meta-analyses of diagnostic test accuracy (Lee, Kim, Choi, Huh, & Park, 2015; Macaskill et al., 2010; Ochodo, Reitsma, Bossuyt, & Leeflang, 2013; Reitsma et al., 2005; Trikalinos et al., 2012). Without study-level covariates (which is the case for this review), the two are essentially different versions of the same model (Harbord et al., 2007).

The HSROC model directly estimates parameters such as threshold (θ), accuracy (α) and shape (β) as random-effects variables; this model facilitates the direct construction of the HSROC curve, by holding the accuracy parameter (α) fixed at its mean, but allowing the threshold parameter (θ) to vary.

The bivariate model, on the other hand, utilizes five parameters, which are the means (μ_A and μ_B) of the logit-transformed sensitivity and specificity, their variances (σ^2_A and σ^2_B, respectively) and the covariance (σ_{AB}) between them (Lee et al., 2015).

As sensitivity and specificity may be highly correlated, it is therefore recommended in the bivariate model to use an elliptical joint confidence region for both the μ_A and μ_B parameters. It is also possible with the bivariate model to construct a prediction region which describes the total degree of uncertainty of the summary points, reflecting between-study heterogeneity, by denoting values of sensitivity and specificity that have the potential to be observed in future studies.

Of the two, the bivariate model is the preferred model for the estimation of summary values of the sensitivity and specificity (Lee et al., 2015; Reitsma et al., 2005) and is therefore the appropriate analysis required for this review. Moreover, a HSROC curve can be fitted through a recalculation of the HSROC parameters, by transforming the estimated parameters of the bivariate model (Lee et al., 2015).

In this review, HSROC plots were used to present the results of each study in receiver operating characteristic (ROC) space. Each study was plotted as a single sensitivity-specificity point and the bivariate model produced two HSROC curves with summary operating points, 95% confidence regions and 95% prediction regions for imminent tools and longer-term tools, respectively. The summary operating point shows summary values of sensitivity and specificity, plotted as a sensitivity-specificity point on the HSROC curve. The confidence region is associated with the summary sensitivity and specificity estimates in the HSROC space, while simultaneously (based on the included studies) incorporating their inverse relationship. The confidence region, however, is only a measure of within-study uncertainty and does not display

between-study heterogeneity (Dinnes, Deeks, Kirby, & Roderick, 2005; Lee et al., 2015). Assuming the model is appropriate, the prediction region reflects the region within which there is 95% confidence for the true sensitivity and specificity of a future study (Rutter & Gatsonis, 2001). Therefore, the prediction region is useful in the capacity for predicting the summary sensitivity and specificity of a diagnostic accuracy study that is similar to those included (Dinnes et al., 2005; Lee et al., 2015), showing between-study variability.

This meta-analysis was conducted on Stata (StataCorp, 2015), using the *midas* command to generate summary statistics and a summary ROC curve. Summary values for sensitivity, specificity, DOR and AUC were obtained; summary PPVs and NPVs were calculated as medians for each group (imminent vs. longer-term). Summary AUC values for the wider group of samples were calculated as medians.

2.6.4 Assessment of heterogeneity

For meta-analyses of diagnostic or prognostic test accuracy, there is assumed to be some heterogeneity (Macaskill et al., 2010) which is why a random-effects model of analysis is the recommended method of analysis. Before analysis, to see whether pooling study data seems appropriate, a simple, but useful, method is to plot a graph of individual study outcome measures (Devillé et al., 2002) (e.g. plot all sensitivities for imminent tools) along with their 95% confidence intervals or standard errors and make a subjective assessment of the heterogeneity between the studies. These plots showed there to be a fair amount of heterogeneity, but as this is expected for meta-analysing data from prognostic studies, it was deemed acceptable to pool data. See Appendix 4 for plots.

Heterogeneity arising due to variation in diagnostic thresholds (in this case, tool cut-off scores) is likely to be present in a meta-analysis of diagnostic test accuracy. In this study, it has been attempted to make sure that, where possible, cut-off scores are uniform across studies of

the same tool. Furthermore, the pooling of AUC statistics takes into account possible variation in diagnostic thresholds, thereby accounting for the problem of heterogeneity in cut-off scores.

Quantification of heterogeneity in the form of Cochrane's Q statistic or I^2 (as is used in meta-analyses of RCTs) is not the recommended method for bivariate analysis in meta-analyses of prognostic test accuracy and usage is actively discouraged for meta-analyses of diagnostic/prognostic test accuracy (Naaktgeboren et al., 2016). However, the literature is still unclear about what may be an appropriate alternative method of quantifying between-study heterogeneity in bivariate meta-analyses. Alternatives to I^2 have been proposed (Jackson, White, & Riley, 2012; Jackson, White, & Thompson, 2010; Verde, 2008; White, 2011; Zhou & Dendukuri, 2014); however, there is a lack of literature that has researched, validated and provided guidance for their usage (Naaktgeboren et al., 2016).

Instead, the recommended method of assessing heterogeneity is the subjective visual evaluation of the scatter of points from the summary ROC curves and the size of the ellipse of the prediction regions. A greater scatter of points from the ROC curve and a larger prediction region are indicative of greater levels of heterogeneity (Macaskill et al., 2010). Therefore, it was decided to rely on subjective evaluation of heterogeneity, through the visual methods described, for this meta-analysis and not to use a statistic for the quantification of between-study variability.

2.6.5 Meta-regression and subgroup analyses

Meta-regression analysis is a statistically rigorous method used in meta-analyses to assess the effect of moderator variables on the primary outcome measures. In a similar way that regression is used in primary studies to examine relationships between variables, meta-regression is used in meta-analyses with the covariates at the study- or sample-level rather than the subject-level, and the dependent variable is an overall performance measure rather than an individual subject score.

For this review, meta-regression analyses were performed to investigate the relationship between an overall accuracy estimate (area under the curve value) and a number of study or sample characteristics, to see whether any had a moderating effect on the AUC. Sample-related variables included sample size, gender, mean age of participants, proportion of participants with psychotic disorder, proportion of participants with personality disorder and proportion of participants with a violent index offence. Study-related variables included study temporal design (prospective vs. retrospective), type of tool investigated (actuarial vs. SPJ), follow-up period post-assessment and definition of violent outcome used.

Meta-regression was only performed for the studies included in the meta-analysis, and could only be conducted for the longer-term tool samples (n=29), as the number of imminent tools samples was too small (n=6). Any findings of interest from meta-regressions were investigated further using subgroup analyses. Subgroup analyses involve the stratification of the samples based on particular study or sample characteristics and then the comparison of summary accuracy estimates for each subgroup, in order to confirm and elaborate on findings from the meta-regression.

Chapter 3: Results

3.1 Descriptive characteristics

3.1.1 Whole sample

For the meta-analysis, information was collected for 3,306 participants in 35 samples from 19 independent publications (Table 2). Standardised outcome information on numbers of TPs, FPs, FNs and TNs was not reported in manuscripts for 24 of the samples (2,215 [66%] participants) and was obtained directly from study authors.

Of the 3,306 participants, 2,645 (80%) were male and 661 (20%) were female. The overall mean age of participants was 36.6 years (standard deviation [SD] = 3.5). The number of participants diagnosed with a psychotic disorder was 1,586 (48%) and 571 (28%) had a diagnosis of personality disorder, while 1,804 (64%) participants had committed a violent index offence.

The mean sample size was 94.5 (SD = 120.4). With respect to temporal design of the study, prospective design was the most common (21; 60%), followed by retrospective design (12; 34%) and pseudo-prospective design (2; 6%). The overall mean length of follow-up period was 567.4 days (SD = 922.3).

The overall rate of violence over the defined follow-up period for each study ranged from 7% of the sample being violent (Daffern & Howells, 2007) to 52% (Snowden, Gray, Taylor, & Fitzgerald, 2009), with an average rate of violence of 31% (SD = 16.1) of patients in the sample being violent. With regard to the definition of violent outcome, the majority of studies (21; 60%) had a violent outcome restricted to only interpersonal physical violence, while the rest (14; 40%) defined violence to include interpersonal physical violence, as well as verbal aggression.

The final distribution of number of samples per tool was fairly even, with the exception of one instrument. The PCL-R and VRAG each had four samples available for meta-analysis (Snowden et al., 2009; Thomson, Davidson, Brett, Steele, & Darjee, 2008). For the BVC, there were three samples available for meta-analysis (Chan & Chow, 2014; Chu, Daffern, & Ogloff, 2013; Hvidhjelm, Sestoft, Skovgaard, & Bue Bjorner, 2014) and this was also the case for the DASA (Chan & Chow, 2014; Chu et al., 2013; Daffern & Howells, 2007), COVR (McDermott, Dualan, & Scott, 2011; Snowden et al., 2009) and START (Desmarais, Nicholls, Wilson, & Brink, 2012; Gunenc, O'Shea, & Dickens, 2017; Nonstad et al., 2010). The PCL:SV and VRS had only one study each (Dolan, Fullam, Logan, & Davies, 2008; Negredo, Melis, & Herrero, 2015). The HCR-20 had a particularly high number of samples, with thirteen samples examining its predictive accuracy (Arai, Takano, Nagata, & Hirabayashi, 2016; de Vogel & de Ruiter, 2005; De Vogel & De Ruiter, 2006; de Vries Robbe, de Vogel, Wever, Douglas, & Nijman, 2016; Fagan et al., 2009; Jeandarme, Pouls, De Laender, Oei, & Bogaerts, 2017; Mudde, Nijman, van der Hulst, & van den Bout, 2011; Negredo et al., 2015; Snowden et al., 2009).

With regard to instrument type, there were 19 samples investigating the accuracy of actuarial tools, compared to 16 samples investigating the accuracy of SPJ tools. Studies were conducted in 12 different countries: Australia, Belgium, Canada, Denmark, Hong Kong, Ireland, Japan, the Netherlands, Norway, Spain, the UK and the USA.

3.1.2 Comparison between groups

There were a total of 1,394 participants in the 6 imminent tool samples (4 publications), compared to 1,912 participants for the 29 longer-term tools samples (15 publications). For both the imminent and longer-term tool groupings, samples were approximately 80% male and there was little difference in mean age (37.0 and 36.4 years, respectively). For those studies that reported it, the proportion of participants who were diagnosed with a psychotic disorder was much higher for the longer-term tool group (81%) than for the imminent tool group (36%). The

same was true for the proportion of participants diagnosed with a personality disorder; only 9% of participants in the imminent tool sample had this diagnosis, compared to 36% of participants in the longer-term tool group. With regard to the proportion of the sample with a violent index offence, there was less discrepancy, though the proportion was again lower in the imminent tool sample of participants (51%) compared to the longer-term tool sample (73%).

Sample size for imminent tool studies ranged between 38 and 530 participants, with a mean sample size of 232 (SD = 233), while for longer-term tool studies, the range in sample size spanned from 29 to 185, with a mean sample size of 66 (SD = 54) participants. With regard to temporal design, all six of the imminent tool samples were prospective studies. However, in the longer-term tool group, there was more of a mix; 15 (52%) were prospectively designed, 12 (41%) were retrospective and 2 (7%) were pseudo-prospective.

The mean length of follow-up differed greatly between the two groups; for the imminent tool group, the mean follow-up period was 1 day (SD = 0.0), while for longer-term tools, it was 692.2 days (SD = 978.6).

The mean rate of violence over the defined follow-up period was slightly lower in the imminent tool group – 23.8% (SD = 15.3) of patients were violent during the study period post-assessment – compared with the longer-term tool group (32.6% of patients were violent; SD = 16.2). For imminent tool studies, the majority (4; 67%) included verbal aggression in their definition of violent outcome, while the other 2 studies (33%) defined violent outcome as only interpersonal physical violence. For longer-term tool studies, however, the majority (19; 66%) restricted the definition of violent outcome to only interpersonal physical violence, with the rest (10; 34%) incorporating verbal aggression into their definition. See Table 2 for all information on descriptive statistics.

Table 2 Descriptive and demographic characteristics of samples for imminent and longer-term instruments

Note: Data are number (%) of samples, unless stated otherwise. SD = standard deviation.

Category and group	Imminent (n = 6)	Longer-Term (n = 29)
Tool Information		
Type of tool		
Actuarial	6 (100)	13 (45)
Structured professional judgement	0 (0)	16 (55)
Tool used		
BVC	3 (50)	-
COVR	-	3 (10)
DASA	3 (50)	-
HCR-20	-	13 (45)
PCL-R	-	4 (14)
PCL:SV	-	1 (3)
START	-	3 (10)
VRAG	-	4 (14)
VRS	-	1 (3)
Sample characteristics		
Male participants	1115 (80)	1549 (81)
Age (years; mean (SD))	37.0 (2.5)	36.4 (3.8)
Psychotic disorder	508 (37)	931 (81)
Personality disorder	122 (9)	449 (36)
Violent index offence	715 (51)	1089 (73)
Study design		
Sample size (mean (SD))	232 (233)	66 (54)
Temporal design		
Retrospective	0 (0)	12 (41)
Prospective	6 (100)	15 (52)
Pseudo-prospective	0 (0)	2 (7)
Length of follow-up (days; mean (SD))	1.0 (0.0))	692.2 (978.6)
Outcome		
Violent outcome measured		
Only interpersonal physical violence	2 (33)	19 (66)
Including verbal aggression	4 (67)	10 (34)
Rate of violence during study (mean (SD))	23.8 (15.3)	32.6 (16.2)

3.2 Summary statistics

Differences in estimates of predictive accuracy were found between the two groups of instruments (Table 3). Meta-analysis of the studies assessing imminent tools gave a summary sensitivity of 0.59 (95% CI: 0.29-0.83) and a summary specificity of 0.99 (95% CI: 0.80-1.00). For longer-term tools, the sensitivity was 0.75 (95% CI: 0.65-0.83) and specificity was 0.56 (95% CI: 0.46-0.66).

The summary diagnostic odds ratio (DOR) for longer-term instruments was 4.0 (95% CI: 3.0-6.0). A summary DOR for imminent tools was not possible to accurately calculate due to the number of zero-value categories (2 of the 6 samples included had one or more cells with zero values). This resulted in skewed high numbers for the calculated DORs.

Summary estimates for positive and negative predictive values were not computed by the meta-analysis *midas* command on Stata, so median PPVs and NPVs were calculated for the two groups of tools. The median PPV for imminent instruments was 0.36 (IQR: 0.10-0.93) and the median NPV was 0.99 (IQR: 0.85-1.00). The median PPV for longer-term instruments was 0.56 (IQR: 0.30-0.75) and the median NPV was 0.75 (IQR: 0.58-0.95). Thus, the median PPV for longer-term instruments was higher than the median PPV for imminent instruments, but significant overlap in IQRs suggested this difference to be non-significant. For NPVs, the median was higher for imminent tools than for longer-term tools, but again there was an overlap in IQRs.

The final measure of predictive validity used was the AUC value. Two different summary estimates of AUC values are reported, based on different sample sizes. The first were calculated as median AUCs. For the meta-analysis, the included samples were restricted to only those for which information on numbers of true positives, false positives, false negatives and true negatives (n=35) was available. However, there were a further 43 samples (from 24 publications) that reported AUC values and fit the primary inclusion criteria. In order to avoid

limiting the sample results too much due to lack of information, the median AUCs for the wider range of eligible studies have also been included.

With this wider sample, there were 78 samples and a total of 7,705 participants, from 43 publications. This distributed to 10 samples for imminent tools (1,666 participants), compared to 68 samples for longer-term tools (6,039 participants). The resulting median AUC values reflected the results found in the meta-analysis: imminent tools performed better than longer-term tools. The median AUC for imminent tools was 0.83 (IQR: 0.71-0.85), while for longer-term tools it was 0.68 (IQR: 0.62-0.75) (Table 3).

The second summary AUC to report is that from the meta-analysis, using the same samples as for the previously reported range of summary performance measures. With regard to these summary AUC values, imminent tools performed significantly better than the group of longer-term tools. The summary AUC value was very high for imminent tools, at 0.90 (95%CI: 0.87-0.92), but for longer-term tools, it was 0.71 (95% CI: 0.67-0.75). The 95% confidence intervals of these two AUC values do not overlap, indicating a significant difference between them.

Table 3 Summary accuracy estimates produced by two categories of violence risk assessment instruments

	Imminent Instruments (n=6)	**Longer-Term Instruments (n=29)**
Summary estimates (95% confidence interval)		
Sensitivity	0.59 (0.29 – 0.83)	0.75 (0.65 – 0.83)
Specificity	0.99 (0.80 – 1.00)	0.56 (0.46 – 0.66)
PPV*	0.36 (0.10 – 0.93)	0.55 (0.30 – 0.75)
NPV*	0.99 (0.85 – 1.00)	0.75 (0.58 – 0.95)
DOR	-	4.00 (3.00 – 6.00)
AUC*	0.83 (0.71 – 0.85)	0.68 (0.62 – 0.75)

* Median (interquartile range)
Note: median AUC values calculated from wider samples (n = 78) – 10 samples for imminent tools and 68 samples for longer-term tools

3.3 HSROC curves

Figure 4 shows the summary receiver operating characteristics curve formed from the meta-analysis of imminent instruments. The summary ROC curve is approaching the top left-hand corner of the graph, indicating high accuracy. The summary sensitivity, specificity point is plotted on the curve, showing a fairly low sensitivity of 0.59, but a near-perfect specificity of 0.99. Looking at the six studies of imminent tools plotted in ROC space, it can be seen that there is a fair amount of heterogeneity, with samples 1 and 2 (Chan & Chow, 2014; Daffern, 2007) displaying very low sensitivities (0.13 and 0.16, respectively) but very high specificities (0.98 and 1.00, respectively). Samples 3, 5 and 6 (Chan & Chow, 2014; Chu et al., 2013; Hvidhjelm et al., 2014) show higher sensitivities (0.73, 0.76 and 0.66, respectively) and similarly high specificities (1.00, 0.81 and 1.00, respectively). Sample 4 (Chu et al., 2013) shows a slightly opposing pattern of a high sensitivity (0.95) and a lower specificity (0.52). A large amount of the graph is underneath the curve, which is reflected in the high summary AUC value of 0.90. The thick dashed line shows the contour of the 95% confidence region, displaying within-study uncertainty and it is fairly large, showing a relatively high level of within-study uncertainty. The thin dotted line is the prediction region, showing between-study heterogeneity. The prediction region is indicative of the region within which there is 95% confidence that the sensitivity, specificity point of a future study would lie. As can be seen, this prediction contour for imminent instruments covers almost the entirety of the graph, indicating the heterogeneity between the six studies by showing that there are high levels of uncertainty about where a sensitivity-specificity point in any future study could fall in ROC space.

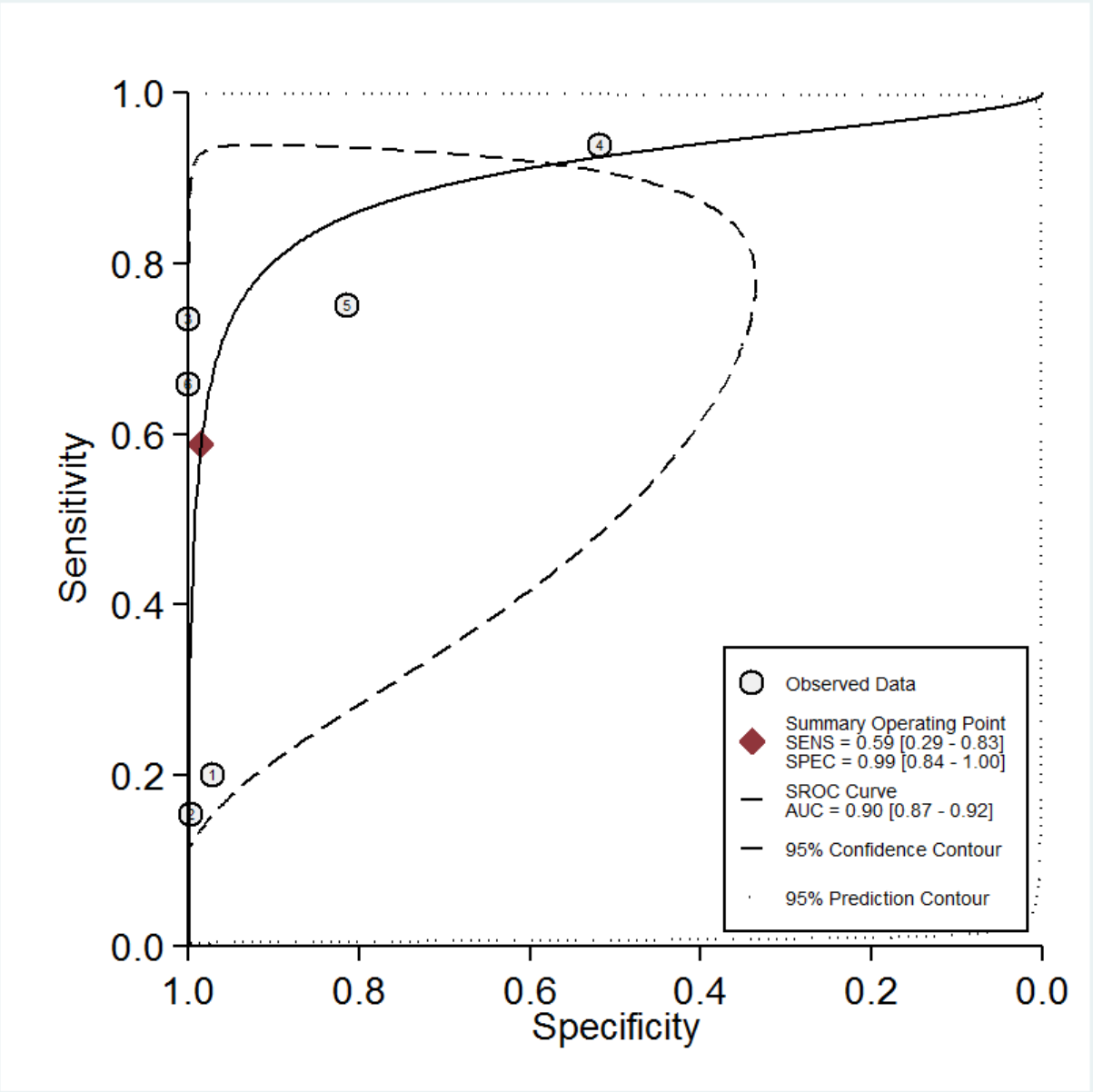


Figure 4 Summary receiver operating characteristics (SROC) curve from bivariate analysis of imminent violence risk assessment instruments for forensic inpatient violence

Note: Summary operating point = best fit for sensitivity and specificity.

Figure 5 displays the summary ROC curve from the meta-analysis of the longer-term instruments. Here, it can be seen that the summary ROC curve is not particularly close to the top left-hand corner of space, but is above the y=x diagonal line that would indicate an uninformative test. The summary sensitivity, specificity point is plotted on the curve, with a high sensitivity of 0.75 and a lower specificity of 0.56. There is a wide spread between the 29

sample points plotted on the graph, indicating moderate levels of heterogeneity. Only sample 8 (Thomson et al., 2008) has a perfect specificity (1.00), though this is paired with a low specificity (0.31). Conversely, sample 7 (Snowden et al., 2009) has a perfect sensitivity (1.00) but is paired with a specificity of 0.00. Sample 9 (Thomson et al., 2008) also has a specificity of 0.00, but a less elevated sensitivity of only 0.69. Samples 1, 5, 10, 11, 12, 14 and 22 (Arai et al., 2016; de Vogel & de Ruiter, 2005; De Vogel & De Ruiter, 2006; Desmarais et al., 2012; Snowden et al., 2009) all have high or perfect sensitivities (1.00, 1.00, 1.00, 1.00, 1.00, 0.96 and 1.00, respectively) and are paired with a range of specificities (0.67, 0.75, 0.58, 0.27, 0.79, 0.49 and 0.14, respectively). The rest of the sample points fall at various points along the curve, with a typical trade-off between sensitivity and specificity values (where sensitivity is low, specificity is high and vice versa). The proportion of the graph that is under the curve does not appear to be high, though it is definitely greater than 50%. This is reflected in the low-medium summary AUC value of 0.70. The 95% confidence contour is fairly small, indicating there not to be much within-study uncertainty. The 95% prediction region, however, is large, implying a high level of between-study heterogeneity, such that a future study could have a sensitivity-specificity point fall anywhere within the contour.

Looking at both Figure 4 and Figure 5 comparatively, a number of observations can be noted. Firstly, with regard to accuracy, it is clear that imminent instruments perform better, with their summary ROC curve much closer to approaching the top-left hand corner than the curve for longer-term tools. These contrastive levels of accuracy are reflected in the proportions of the graphs that fall under the curve, with a notably larger area for imminent, compared to longer-term, instruments. This is shown by their differing summary AUC values, with a high summary AUC of 0.90 (95% CI: 0.87-0.92) for imminent tools and a poor-moderate summary AUC value of 0.70 (95% CI: 0.67-0.75) for longer-term tools. Crucially, there is no overlap in confidence intervals of summary AUC values between the two groups of tools, indicating a significant difference between the summary AUC values.

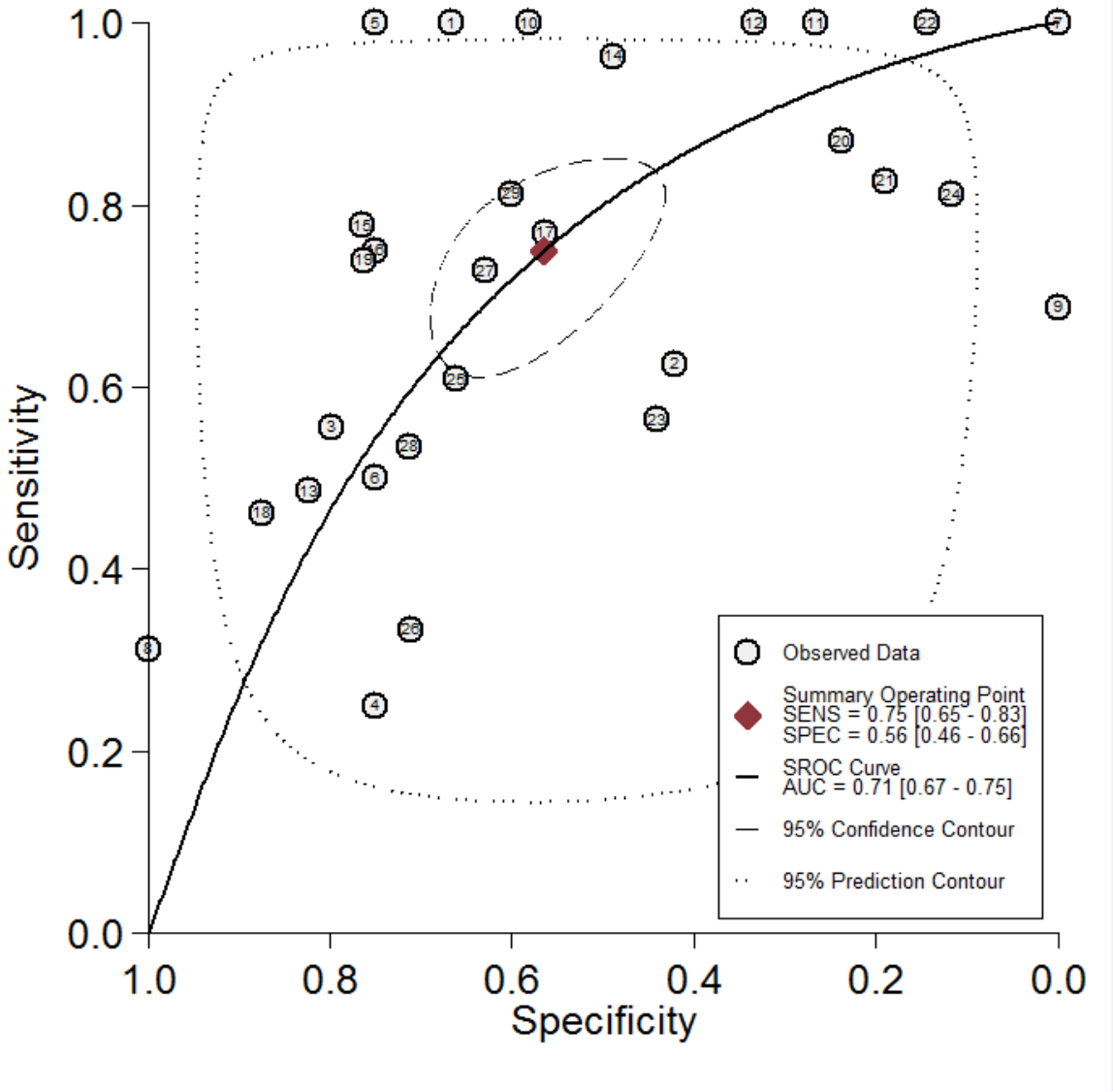


Figure 5 Summary receiver operating characteristics (SROC) curve from bivariate analysis of longer-term violence risk assessment instruments for forensic inpatient violence

Note: Summary operating point = best fit for sensitivity and specificity.

The summary sensitivity-specificity points for each graph suggest similar findings. For longer-term tools, the summary point is relatively far from being at the "gold-standard" (1.00, 1.00) coordinate, while for imminent instruments, the summary point appears to be a lot closer to the "gold-standard" coordinate. However, the summary sensitivity and specificity are closer together in value for longer-term instruments than they are for imminent instruments (0.75, 0.56 for longer-term tools, compared to 0.59, 0.99 for imminent tools, respectively).

The two graphs also differ in terms of the spread of the data points across the ROC space and from the summary curve. However, this is hard to quantify or directly compare, given the difference in sample numbers between the groups (6 samples for imminent tools, compared to 29 for longer-term tools). It appears to be the case that for imminent tools, data points are closer to the curve, but, more generally, further distances from one another; while, for longer-term tools, data points are often further from the curve, but all remain relatively close to a number of other studies.

The confidence contour for longer-term tools is clearly much smaller than that for imminent tools. However, once again this may have been influenced by the difference in number of sample data points between the groups. Similarly, the prediction contour for imminent instruments may be larger than the prediction contour for longer-term instruments for the same reason. However, the discrepancy between the sizes of the prediction contours is less obvious and dramatic, compared with the size difference of the confidence contours. Both groups of studies show fairly large levels of between-study heterogeneity, which is unsurprising, as this is presumed when pooling bivariate data for studies investigating diagnostic/prognostic test accuracy, primarily given possible differences in diagnostic threshold (in the context of this study, cut-off scores).

3.4 Heterogeneity

Detecting and quantifying heterogeneity for meta-analyses of studies of diagnostic test accuracy is distinct from techniques used for meta-analyses of randomized controlled trials (RCTs), due to the bivariate estimates being pooled and investigated. Substantial heterogeneity is assumed, primarily based on differences in diagnostic thresholds (here, cut-off scores) between studies.

As stated in the methods, the statistical quantification of heterogeneity in the form of Cochrane's Q statistic or I^2 is discouraged for meta-analyses of diagnostic test accuracy

(Naaktgeboren et al., 2016). Instead, subjective detection of heterogeneity is recommended (Macaskill et al., 2010) through visual examination of the spread of data points from the summary ROC curve produced in the meta-analysis, as well as the size of the prediction ellipse. As discussed, in both Figure 4 and Figure 5, the prediction ellipses are large, implying large amounts of between-study heterogeneity. However, looking at the spread of data points from the curve for imminent instruments there is relatively little spread from the curve, indicating that perhaps the between-study heterogeneity is not as large or as much of an issue as the size of the prediction ellipse suggests. For longer-term instruments, the points are spread further from the curve; however, this may be a function of the number of data samples included (n=29) compared to that of imminent instruments (n=6). Therefore, it is difficult to say whether the spread of points from the curve aligns with the large prediction ellipse and indicates high levels of between-study heterogeneity or whether the spread of points from the curve is actually not that significant and contradicts the large prediction ellipse, in the same way that the studies of imminent instruments show.

3.4.1 Meta-regression analyses

Meta-regression analyses were performed to investigate whether there were any relationships between accuracy estimates and study or sample characteristics. The variables chosen for the meta-regression analysis were picked based on assumptions that they may have had the capacity to influence accuracy. The AUCs of each individual study in the longer-term tool group were compared, as an overall measure of accuracy. Meta-regression analysis was only performed on the group of longer-term tools, as the sample size of the imminent tools group was too small. Results of the comparison between AUCs and other variables in the meta-regression analyses are shown in Table 4.

Table 4 Results of meta-regression analyses of longer-term instrument samples comparing AUCs to study- and sample-related variables

Variable	Number of samples (n)	Beta coefficient (β)	95% confidence intervals	Significance (p)
Gender	25	0.0002	-0.0015 – 0.0019	0.81
Sample size	25	0.0004	-0.0010 – 0.0019	0.56
Mean age	12	-0.0484	-0.1035 – 0.0067	0.08
Temporal design [a]	25	-0.0431	-0.1201 – 0.0338	0.26
Tool type [b]	25	0.1075	-0.0454 – 0.2604	0.16
Violent outcome [c]	25	-0.0448	-0.2432 – 0.2537	0.65
Psychotic diagnosis	9	0.0016	-0.0089 – 0.0121	0.73
Personality disorder diagnosis	6	0.0022	-0.0085 – 0.0129	0.60
Violent index offence	9	0.0039	-0.0171 – 0.0248	0.68
Follow-up period	23	0.0001	-0.0002 – 0.0004	0.49

[a] Prospective design coded as 1; pseudo-prospective coded as 2; retrospective coded as 3
[b] Actuarial tool coded as 1; structured professional judgement tool coded as 2
[c] Violent outcome including verbal aggression coded as 1; violent outcome only including interpersonal physical violence coded as 2

The first variable examined was gender proportion of the samples because many violence risk assessment instruments have predominantly been developed and validated using samples that are entirely male. Given that there is a notably larger proportion of men than women in the criminal justice system, this is understandable, but it carries the implication that these tools, which have been validated on men, are therefore more likely to be accurate for men. However, no support was found for an association between gender and risk assessment accuracy (β = 0.0002, p = 0.81).

Sample size and mean age were also examined as two variables for meta-regression. Sample size varied a great deal between studies, ranging from 8 patients (Snowden et al., 2009) to 168 patients (Jeandarme et al., 2017). Mean age saw less variation between studies, ranging from 30.7 years (de Vogel & de Ruiter, 2005) to 46.1 years (McDermott et al., 2011). After meta-regression analysis, neither sample size (β = 0.0004, p = 0.56) nor mean age (β = -0.0484, p = 0.08) were shown to have any effect on accuracy estimates in longer-term tools.

Accuracy estimates were also compared between studies that had used a prospective design, compared to those that had used a retrospective design. Given that any retrospectively-designed studies that did not have blinding to outcomes were excluded, it would be unlikely for

there to be any difference between prospective and retrospective studies; this is what was found (β = -0.0431, p = 0.26).

It was also investigated whether there was any clear difference in accuracy between actuarial tools and structured professional judgement (SPJ) tools. In the longer-term tool group, 13 studies examined actuarial tools and 16 studies assessed the predictive accuracy of SPJ tools. However, there was no effect of tool type found in meta-regression (β = 0.0175, p = 0.16).

Another variable of interest was the definition of violent outcome in the studies. 19 studies defined a violent outcome as an instance of physical violence towards another person, while the other 10 studies had a broader definition of violence, including both interpersonal violence *and* verbal aggression. Meta-regression found no effect of violent outcome definition on accuracy estimates (β = -0.0448, p = 0.65).

Diagnoses of psychotic disorders or personality disorders are commonly included as an item on risk assessment tools and are common diagnoses in forensic populations (Fazel & Danesh, 2002). It was examined whether the proportion of patients with a diagnosis of a psychotic disorder or a personality disorder within study samples had an effect on accuracy estimates; however, no such association was revealed (psychotic disorder: β = 0.0016, p = 0.73; personality disorder: β = 0.0022, p = 0.60).

Many patients who are transferred to forensic psychiatric facilities are transferred from prison or due to violent behaviour elsewhere, whether or not that had resulted in a conviction. The effect of the proportion of patients with a violent index offence (i.e. a violent offence, which led to their admission into the forensic psychiatric facility) on accuracy estimates was investigated. There was found to be no effect of the proportions of samples with a violent index offence on estimates of risk assessment test accuracy (β = 0.0039, p = 0.68).

The division of categories for comparison in meta-analysis was done based on the intended length of follow-up period for which the tool in question had been designed. However, studies in the longer-term group displayed a large amount of variation in the follow-up period that was actually used for their sample post-assessment. Follow-up periods for studies using

longer-term tools ranged between 56 days (Dolan et al., 2008) and 8 years (Thomson et al., 2008). Meta-regression analysis for follow-up period indicated there was no significant effect on accuracy estimates of risk assessment tool performance ($\beta = 0.0001$, $p = 0.49$).

Given that there were no significant effects found from the multitude of variables investigated in meta-regression, no subgroup analyses were conducted.

3.5 Comparisons between tools

Our findings showed that, as a general group, imminent tools performed better than longer-term tools for the prediction of inpatient violence for forensic psychiatric populations. Further discrepancies in levels of accuracy can be seen between individual tools (Table 5). Although the wider sample included 78 samples, compared to the 35 samples included in the meta-analysis, the only performance measures reported for the 78 samples were AUC values. AUC values are informative to an extent, but do not give a comprehensive evaluation of the accuracy of the tool. Therefore, when looking at individual tool performance, only the 35 samples used in the meta-analysis are included, so that a range of performance measures can be discussed. Due to small sample sizes, performance measures for instruments that had three or fewer samples have not been discussed here, though they are reported in Table 5; this applies to the BVC, COVR, DASA, PCL:SV, START and VRS.

The HCR-20 had the largest number of samples (n=13); the median AUC value was poor to moderate at 0.71 (IQR: 0.67-0.81). The median sensitivity and median specificity were 0.78 (IQR: 0.56-1.00) and 0.71 (IQR: 0.56-0.76), respectively. The median PPV was low at 0.31 (IQR: 0.26-0.56), while the median NPV was high at 0.94 (IQR: 0.75-1.00). The median DOR was 3.6 (IQR: 1.9-4.5). Overall, taking these statistics into account, the HCR-20 can be said to perform poor-moderate for the prediction of forensic inpatient violence.

The other two instruments with a sufficient number of samples to warrant discussion were the PCL-R (n=4) and the VRAG (n=4). Both the PCL-R and VRAG had poor to moderate

median AUC values of 0.69 (IQR: 0.63-0.75) and 0.71 (IQR: 0.67-0.74), respectively. The VRAG had a median sensitivity of 0.91 (IQR: 0.78-1.00), while the median sensitivity of the PCL-R was low at 0.53 (IQR: 0.45-0.63). With regard to specificity, again the PCL-R median was low at 0.60 (IQR: 0.38-0.81); however, the median specificity of the VRAG was very low at just 0.06 (IQR: 0.00-0.12). The PCL-R and VRAG had PPVs of 0.72 (IQR: 0.63-0.82) and 0.66 (IQR: 0.55-0.79), respectively. The PCL-R also had a low median NPV of 0.55 (IQR: 0.43-0.70); with regard to NPV, the VRAG had a very low NPV with a median of 0.08 (IQR: 0.00-0.37). Finally, with regard to DORs, both instruments had low medians. The median DOR for the PCL-R was 1.1 (IQR: 1.0-2.1), indicating a barely useful test. The median DOR for the VRAG was lower at 0.6 (IQR: 0.3-0.8), indicating it not to be a useful instrument for the prediction of forensic inpatient violence. Overall, these two tools performed poorly for violence prediction for forensic inpatients.

Table 5 Median accuracy estimates and interquartile ranges for each violence risk assessment instrument

	Sample used in meta-analysis (n=35)					***Wider Sample (n=78)***	
	No. of studies	*Sensitivity*	*Specificity*	*PPV*	*NPV*	*DOR*	*No. of studies*	*AUC* [a]
BVC	3	0.66 (0.41-0.80)	1.00 (0.76-1.00)	0.37 (0.25-0.68)	0.99 (0.78-0.99)	43.1 (29.6-56.6) [c]	5	0.83 (0.75-0.87)
COVR	3	0.73 (0.49-0.73)	0.75 (0.69-0.76)	0.50 (0.38-0.64)	0.73 (0.61-0.83)	4.5 (2.8-6.8)	3	0.73 (0.34-0.79)
DASA	3	0.73 (0.43-0.75)	0.98 (0.90-0.99)	0.35 (0.20-0.63)	0.99 (0.97-1.00)	10.9 (9.8-12.1) [c]	5	0.83 (0.65-0.90)
HCR-20	13	0.78 (0.56-1.00)	0.71 (0.56-0.76)	0.31 (0.26-0.56)	0.94 (0.75-1.00)	3.6 (1.9-4.5)	27	0.70 (0.62-0.80)
PCL-R	4	0.53 (0.45-0.63)	0.60 (0.38-0.81)	0.72 (0.63-0.82)	0.55 (0.43-0.70)	1.1 (1.1-2.1)	10	0.64 (0.61-0.69)
PCL:SV	1			- [b]			7	0.68 (0.54-0.72)
START	3	0.81 (0.78-0.89)	0.60 (0.55-0.68)	0.52 (0.44-0.62)	0.86 (0.82-0.92)	9.0 (7.8-17.0)	8	0.81 (0.69-0.88)
VRAG	4	0.91 (0.78-1.00)	0.06 (0.00-0.12)	0.66 (0.55-0.79)	0.08 (0.00-0.37)	0.6 (0.3-0.8)	9	0.63 (0.57-0.64)
VRS	1			- [b]			4	0.63 (0.50-0.69)

[a] Median AUCs calculated using wider sample
[b] Only two samples or fewer, so median or interquartile ranges only given for AUC using wider sample
[c] Based on only two, of three, samples

Chapter 4: Discussion

This systematic review and meta-analysis investigated the predictive validity of nine violence risk assessment instruments from 43 studies, involving 78 samples with a total of 7,705 patients from 14 different countries. The primary finding was that risk assessment instruments designed for imminent violence prediction generally perform better for the prediction of inpatient violence than tools designed to predict violence over a longer period, based on a range of performance measures. As a measure of overall accuracy, imminent tools studies had a higher summary area under the curve (AUC) value; the median AUC for the wider sample (n=10) for the imminent tool group was 0.83 (IQR: 0.71-0.85) which is considered to indicate a test of high accuracy. For longer-term tools (n=68), however, the median AUC was 0.68 (IQR: 0.62-0.75) which is considered to indicate a test with poor accuracy (Tape, 2006).

Imminent tools had a high summary specificity and median negative predictive value (NPV), indicating their ability to identify low risk individuals. With regards to the specificity, 99% (95% CI: 0.80-1.00) of those who went on to not be violent were correctly classified as low risk. This demonstrates that many patients can be screened out as low risk, although they may have other mental health and psychosocial needs. The NPV indicated that 99% (IQR: 0.85-1.00) of individuals classified as low risk went on to not be violent; however, the NPV is affected by the prevalence of violence. A low prevalence of violence would indicate that when a patient has been classified as low risk, there is a higher likelihood that this classification will be correct due to the fact that fewer patients are actually violent. For imminent tool samples, an average of 23.8% (SD = 15.3) of patients in a sample would actually go on to be violent; this rate of violence is not low and so does not explain why the NPV is high.

The summary sensitivity and median positive predictive value (PPV) for imminent tools, however, were not high. The sensitivity showed that, of those who went on to be violent, only

59% (95% CI: 0.29-0.83) were correctly classified as moderate or high risk[1], meaning that 41% of violent incidents would be "unexpected" as they would be committed by patients who had been classified as low risk. This lower sensitivity is perhaps unsurprising, given that the specificity is so high. There is often a trade-off between sensitivity and specificity, which is dependent on the positioning of the prognostic threshold (i.e. the score above which the individual will be classified as higher risk and below which will be classified as low risk). In this case for imminent tools, if the threshold was lowered, the sensitivity would increase and the specificity would decrease. Of those who were classified as higher risk, the PPV showed that only 36% (IQR: 0.10-0.93) actually go on to be violent. This indicates some level of "over-prediction", where more patients are predicted to be violent than actually are violent. However, as with the NPV, the PPV is also affected by prevalence of violence; a low rate of violence would indicate that a prediction of high risk would statistically be less likely to be correct, which could explain the low PPV.

Longer-term tools performed less well across the range of performance measures. The summary sensitivity indicated that 75% (95% CI: 0.65-0.83) of those went on to be violent were correctly classified as higher risk; this therefore demonstrates that around a quarter of violent patients had incidents "unexpectedly". However, the summary specificity for longer-term tools was low: only 56% (95% CI: 0.46-0.66) of those who did not go on to be violent were predicted to be low risk. This indicates that time, energy and resources would potentially be wasted on monitoring patients who are not actually going on to be violent, assuming these instruments were being used for prediction. Of those who were classified as higher risk, the PPV shows that only 56% (IQR: 0.30-0.75) actually went on to be violent and of those classified as lower risk, the NPV shows that 75% (IQR: 0.58-0.95) actually went on to not be violent. Once again, the low PPV indicates a level of over-prediction of violence: predicting more patients will be violent than actually are. Both NPV and PPV are affected by prevalence rates of violence over the study period, but once again, the mean rate of violence for longer-term tools was not low; on average,

[1] From here on, classification as moderate or high risk will be referred to as "higher" risk

32.1% (SD = 16.2) of patients within a sample were violent, so does not account for the low PPV or moderate-high NPV. The summary DOR for longer-term tools suggested some overall accuracy on a crude measure, namely that a patient who goes on to be violent is 4 times more likely to be classified as higher risk than a patient who does not go on to be violent.

No associations were found between study- or sample-related variables and discrimination estimates (i.e. the AUC value). However, there may be variables, for which information was not available, that could have influenced tool performance. One of these may be the level of security of the forensic ward in which the primary study was conducted. Wards with higher levels of security impose many more restrictions on their patients and, as such, violent incidents are less likely to occur as there are higher staff-to-patient ratios, and more careful observation. These low rates of violence could therefore affect the performance of these tools. Other variables that it may have been useful to have information on would have been proportion with substance use disorder or learning difficulties, which was not reported in enough included studies to warrant inclusion in the meta-regression. Another variable that may have been informative for the sub-analyses was time to first violent incident, particularly for those studies with longer follow-up periods. This information may have provided further information on whether there were any patterns in the predictive accuracy of the tool in question by different time periods.

4.1 Individual tool performance

In general, these findings suggest that there are large variations in whether violence risk assessment instruments work adequately for the prediction of inpatient violence. Of those included in the meta-analysis, the HCR-20 had the most samples (n=13). The HCR-20 is the most widely used tool internationally (Singh et al., 2014), yet across a range of performance measures it had a low to moderate accuracy. Given that it is so widely used, it is perhaps surprising to note that it performs relatively poorly for predicting forensic inpatient violence.

The HCR-20 is a general violence risk assessment instrument, with applications and recommendations for use in a broad range of contexts, for a variety of populations and for a range of follow-up periods for violence prediction. Therefore, it is often used in settings that it has not been specifically tailored to and thus lower levels of accuracy are a likely consequence of how it has been developed.

With respect to the other two instruments with a sufficient number of studies to warrant discussion – the PCL-R and the VRAG – both performed poorly for the prediction of forensic inpatient violence. The VRAG has been revealed to be particularly poor at screening out low risk individuals, with an especially low median specificity and NPV. As can be seen in Table 1, the VRAG comprises only static factors (i.e. factors that do not change over time), like criminal history and environmental and behavioural factors during childhood. This may be a reason why it performs relatively poorly for inpatient violence, as on an inpatient ward more dynamic assessment on a day-to-day basis is required. With respect to the PCL-R, the original purpose of this instrument was to detect psychopathy, not violence risk per se, so its limitations for the prediction of forensic inpatient violence is likely due to the fact that this was not the intended function of the tool. Psychopathy, as discussed, has been identified as a static risk factor for violence, so it is therefore unsurprising that it has been shown to perform well for the prediction of longer-term violent tendencies but not for shorter follow-up periods in a dynamic inpatient setting (Grann et al., 1999; Nicholls, 1997).

An important point to note is that of the nine instruments included in this review, only three had a sufficient number of samples to justify individual discussion of summary performance measures. Furthermore, all three of these tools do not perform well for the prediction of forensic inpatient violence, especially when compared to summary statistics obtained for imminent tools. The reasoning for this is likely due to the fact that they were not developed using inpatient settings; yet, the literature indicates that they still dominate inpatient research. Though these instruments may perform adequately for violent outcomes in the

community, the current evidence does not support their use for the prediction of inpatient violence in forensic psychiatry.

4.2 Comparisons with unstructured clinical judgement

After unstructured clinical judgement became unpopular for violence risk assessment, due to the emergence of actuarial (and subsequently SPJ) tools, the literature on clinical judgement dwindled. However, there have been some studies which have investigated the accuracy of clinical judgement in forensic psychiatry for predicting inpatient violence. One such study assessed 183 male forensic psychiatric patients and monitored assaultive behaviour over the following 12 weeks (Hoptman, Yates, Patalinjug, Wack, & Convit, 1999). Of the total sample, 33% were violent over the 12-week period and 60% of these patients were correctly predicted to be high risk (sensitivity), while 76% of those who were not violent were correctly predicted to be low risk. The PPV of the clinical risk categorisation in this study showed that 55% of those classified as high risk actually went on to be violent and the NPV showed that 80% of those classified as low risk went on to not be violent. If this sensitivity, specificity, PPV and NPV are compared with the summary statistics from this meta-analysis, it can be seen that unstructured clinical judgement does not do badly at predicting forensic inpatient violence. The sensitivity in this study is approximately the same as the summary sensitivity for imminent tools and the specificity is higher than the summary specificity for longer-term tools. The PPV is higher than the median PPV for imminent tools and the NPV is higher than the median NPV for longer-term tools. This is the only available study investigating the predictive accuracy of clinical judgement for forensic inpatient violence that provides a range of performance measures and it suggests that clinical judgement does not perform significantly worse than some of the imminent or longer-term tools reported here. However, this is a small sample of 183 male participants, compared to the summary statistics of samples in this review totalling over 3,000.

4.3 Summary of the literature

The literature search conducted for this review yielded 52 studies that matched the criteria for inclusion in this review. Of these, 43 reported at least one predictive accuracy measure of interest for the population in question. Given that the focus of this review is particularly and purposefully narrow, this is a not an insignificant number of studies to have retrieved.

As has been discussed, there are a number of types of violent outcomes that risk assessment instruments are used to predict (e.g. violent reconviction), so narrowing the focus to only inpatient violence will narrow the literature available. Furthermore, predicting violence upon discharge and violence in the community are issues that are potentially of more interest to the general public, which may explain why these issues constitute a significantly greater proportion of the research than institutional violence. This literature bias may have been overlooked because previous reviews have often combined institutional and community outcomes of violence, rather than focusing on one or the other (Fazel et al., 2012; Singh et al., 2011; Whittington et al., 2013).

Another focus of our review was the chosen sample population, which was restricted to only forensic psychiatric inpatients. As discussed, violence risk assessment tools are used for a number of, predominantly correctional and psychiatric, populations. Forensic psychiatry falls in between these two fields, encompassing those who have psychiatric problems and a history of offending or violence; therefore, a focus on the inpatient forensic psychiatric population further reduces the amount of available literature. Given that both of these restrictions are applied, 52 studies spanning a 15-year period (2002 to 2017) is a reasonable number. These studies were also conducted across 14 different countries, showing a range of multinational investigations.

Despite this, there is a need to improve research in this area. One improvement would be to report a full range of true and false positives and negatives. Only 11 of the total 78 samples reported these. AUC values were reported for all 78 samples; however, when reported in isolation, the AUC value is not indicative of the overall accuracy of the instrument so should be reported in conjunction with a number of other measures of predictive accuracy, including sensitivity, specificity, PPV, NPV and DOR (Singh, 2013). Furthermore, in order to carry out a meta-analysis with these performance measures, the numbers of true and false positives and negatives is required.

There is a clear discrepancy in the distribution of studies investigating imminent instruments compared to longer-term instruments; of the 78 samples, only 10 samples investigated imminent tools while 68 examined the predictive accuracy of longer-term tools. The groupings were already unevenly distributed with regards to the tools used; only two imminent tools were included, compared to seven longer-term tools, which may initially appear to be a factor causing the uneven sample distribution. However, this is not quite the case if the number of samples for each individual tool is broken down. With regard to the wider sample, every tool had between three and ten samples investigating predictive accuracy, apart from the HCR-20 which had a larger sample number of 27. This bias remains when looking at the smaller sample used in the meta-analysis, with 13 of the 35 samples focusing on the HCR-20.

As noted, the HCR-20 is a widely-used tool and is the most commonly used violence risk assessment tool on a number of continents (Singh et al., 2014). It is therefore unsurprising that it features heavily throughout the literature on risk assessment and has been examined in numerous validation studies. There are, of course, issues with the literature bias towards investigating the HCR-20; the amount of research regarding the predictive accuracy of the HCR-20 for forensic inpatient violence is markedly different than for other instruments. Whilst it does not seem unreasonable for the literature to reflect clinical practice, the findings from this review clearly show that the HCR-20 performs relatively poorly at predicting inpatient violence

in forensic psychiatric samples. Clinical practice should reflect what has been discovered and validated through research, as opposed to clinical practice dictating the direction of research. It is therefore evident that the focus needs shift from the HCR-20 and longer-term tools and towards imminent and possibly newer tools.

Further imbalances can be seen from the instruments being used for the prediction of inpatient violence; there are a significantly greater number of tools designed for longer-term prediction than for imminent prediction. Again, a large proportion of the violence risk assessment literature focuses on longer-term risk outside of an inpatient setting, with many instruments being developed for the prediction of community violence. The BVC and DASA have been developed more recently as the negative consequences of inpatient violence and aggression have become more recognised, necessitating better prediction and management. Given the awareness of the problem of inpatient violence and the findings of this review, it would be advantageous for more tools to be developed for the specific function of predicting inpatient violence. Although this review indicates that the BVC and DASA perform adequately for the prediction of forensic inpatient violence, it would be beneficial for clinicians to have a range of externally validated instruments, similar to the BVC and DASA, for inpatient use.

One area in which the literature on inpatient violence prediction appears to be fairly divided is the broadness of their definition of violent outcome. Some studies only include interpersonal physical violence as the measured outcome, while others include interpersonal physical violence as well as any form of verbal aggression and/or damage to property. Broadening the definition of violent outcome results in higher reported rates of violence as more incidents are counted as violent outcomes. Both the positive and negative predictive values (PPV and NPV) are violence rate-dependent, such that higher rates of violence will increase the PPV and decrease the NPV. Sensitivity and specificity, though commonly believed otherwise, are also somewhat influenced by rates of violence (Brenner & Gefeller, 1997; Li & Fine, 2011; Singh, 2013). Despite there being no exact functional relationship between the rate

of violence and either sensitivity or specificity (Kraemer & Gibbons, 2009), an instrument that discriminates fairly well between those who were violent and those who were not violent would see an increase in sensitivity and a decrease in specificity as the prevalence (and therefore the rate) of violence increases (Singh, 2013). This demonstrates that a difference in rate of violence caused solely by the definition of violent outcome that is used could have an effect on accuracy measures. However, rates of violence are taken into account when calculating the area under the curve (AUC) value and so do not affect the overall accuracy estimate. The inclusion of verbal aggression and/or property damage in a study's definition of violent outcome may be due to an attempt to report higher accuracy estimates to aid publication; however, the meta-regression analysis performed in this review based on the violent outcome definition used did not show there to be any effect on accuracy estimates. Interpersonal physical violence is generally of greater interest than verbal aggression and property damage, due to the serious physical and psychological consequences it can have on staff and other patients.

4.4 Implications for clinical practice

The findings of this review indicate that, for the prediction of inpatient violence, the use of imminent prediction tools over the 24-hour period post-assessment is currently the best-performing method. In clinical practice, consideration should be given to the use of the BVC and DASA. Interestingly, they are the tools that are recommended in the NICE guidelines (NG10; May, 2015) for short-term management of violence and aggression in inpatient mental health settings. Both the BVC and DASA gave high median AUC values of 0.83 and both had high median specificities. Furthermore, the BVC and DASA involve violence prediction over the following 24-hour period, so the window within which violence is predicted to happen is much narrower than for many other risk assessment instruments. This means that violence

prevention and management strategies can be implemented at the exact time when they are necessary, rather than needing to be implemented at some point within a longer timeframe. The goal of risk prediction is to be able to effectively manage and intervene; it is therefore important for the risk assessment tool to provide a useful estimate of when the violence will occur in order for it to be anticipated and planned for.

The BVC and DASA are actuarial risk assessment instruments, with 6 and 7 items respectively; such brief assessments can easily be integrated into daily routine practice on an inpatient ward. They are much less time consuming than instruments like the HCR-20 and can be completed within a matter of minutes, as they take into account current factors rather than requiring an extensive review of background information. Nurses or clinicians, who encounter the patients every day and are very familiar with them, can quickly complete these risk assessments.

A randomised controlled trial (RCT) was carried out using the Swiss version of the BVC in acute wards in the German-speaking part of Switzerland (Abderhalden et al., 2008). The intervention group involved conducting a BVC assessment twice a day for each patient, followed by the discussion and implementation of preventative measures for individuals with high or very high risk scores. The intervention group showed a large decrease in the frequency of severe events of patient aggression (adjusted risk reduction: 42%), compared to the wait-list control group with no risk assessment performed (adjusted risk reduction: 7%). The reduction in the intervention wards was significantly larger ($p<0.001$) than the decline in control wards. The implication of this finding is that the regular administration of risk assessment instruments, in combination with discussion and (where necessary) implementation of preventative measures for high risk individuals, can lead to significant reductions in the prevalence of violent incidents. It cannot be affirmed whether the causal mechanism of violence reduction is the risk assessment administration itself, the discussion and implementation of preventative measures or a combination of the two. This RCT highlights an important message for the field; the use of

violence risk assessment is not the final stage and violence management and prevention are also important for violence reduction, perhaps even more so than risk assessment.

Although the implementation of imminent risk assessment instruments is the recommendation from the findings of this review, it is important to acknowledge that there is also a need for effective and, where possible, non-forceful methods for managing and preventing violent behaviour. Once the risk assessment has been completed and certain patients have been classified as higher risk and others as lower risk, it is important that the next step is to act on this in an effective manner. The pilot study of the previously discussed RCT attempted to separate the components of risk assessment and violence management (Needham et al., 2004). A number of wards implemented use of a risk assessment tool only (the Swiss version of the BVC), while another group of wards implemented both risk assessment tools and a staff training course. The aim of the training course was to keep risk of injury to a minimum when dealing with violent individuals, avoid aggressive outbursts, decrease the usage of harsh coercive measures (i.e. restraint and seclusion) and to help the patient control their behaviour. The findings showed that the wards using only risk assessment had a higher prevalence of interpersonal aggressive incidents (equivalent to baseline) than those that used both risk assessment and a training course. The use of a risk assessment instrument reduced the usage of coercive measures relative to baseline and the combined use of risk assessment and training further reduced the use of coercive measures. This indicates that, whilst use of risk assessment alone causes no difference in prevalence of violence compared to baseline, the addition of training as a proxy for structured violence management reduces prevalence. This emphasises the importance of the implementation of strategies for the management and prevention of violence following risk assessment.

An additional point to consider is the future of violence risk assessment instruments in an age of advancing technology. For example, the primary method of violence risk monitoring may shift from systematic administration of tools to continuous ongoing assessment, in some

cases self-administered by the patients, with links to online databases for constant monitoring of risk (Gulati et al., 2016). Automatic detection of physiological markers for intoxication or abnormal mood states (Large & Nielssen, 2017) could be possible, along with more subjective self-reporting of mood states or psychiatric symptoms. This large and constant stream of data could be used to predict future patient behaviour through easily traceable patterns; with alerts set up for staff to intervene at appropriate moments, this could lead to a reduction in the prevalence of violent incidents.

4.5 Strengths and limitations

To my knowledge, this is the first comprehensive review and meta-analysis of violence risk assessment instruments in the context of their predictive accuracy for inpatient violence in forensic psychiatric populations. Unlike previous reviews of risk assessment tools, this choice of a focused context and outcome makes the findings significantly more useful in application and generalizable to the particular setting in question. Previous reviews have either focused on a wider, more general population (Whittington et al., 2013), used insufficient reporting of results (Hogan et al., 2010), investigated a small selection of tools (Hogan et al., 2010) or specified a wider range of violent outcomes (e.g. not just inpatient violence, but also recidivism on release) (Fazel et al., 2012; Singh et al., 2011). Recent criticism of risk assessment literature has stated that there is insufficient focus on subpopulations in a specific context and the resulting predictive accuracy of instruments for this said context. This review has addressed this issue by concentrating solely on inpatient violence in forensic patients (Douglas, Pugh, Singh, Savulescu, & Fazel, 2017).

Furthermore, the literature on predictive accuracy of violence risk assessment has often been criticised for reporting and relying upon only one or two accuracy measures (Douglas et al., 2017; Singh, 2013). This review has investigated a wide range of accuracy measures,

including a discussion of the drawbacks of each of the accuracy measures reported (Singh, 2013); this gives a broader and more comprehensive picture of the performance of instruments for the prediction of inpatient violence. For example, many studies report the AUC value in isolation as a measure of accuracy; however, the AUC value does not actually measure how well a risk assessment instrument's predictions concur with actual future violence (Singh, 2013), but rather fulfils the role of a rank sum measure of discrimination between those who did and did not go on to be violent (Steyerberg et al., 2010). As such, it does not give an adequate amount of information with which to evaluate the accuracy and effectiveness of the tool in question. It is recommended that a number of accuracy measures – sensitivity, specificity, positive predictive value, negative predictive value and diagnostic odds ratio – should be reported together in order to provide the most thorough evaluation of the tool's predictive performance (Singh, 2013). The reporting of all of these accuracy measures in this review shows a complete picture of the predictive accuracy of the instruments being assessed.

One limitation of this review is that only studies which reported numbers of true and false positives and negatives were included in our meta-analysis, resulting in the exclusion of 33 of the 52 eligible studies. However, median AUC values have been reported for the wider sample of 43 studies (9 studies did not report AUC values for the forensic inpatient sample). Furthermore, the use of numbers of true and false positives and negatives is the input required to produce the range of summary statistics reported, which, as discussed, provides a comprehensive overview of tool performance. The lower number of studies included in this meta-analysis is a reflection of this aim to provide a comprehensive picture and highlights the problems with the literature and reporting of results within the field of risk assessment tool predictive accuracy. Of the 35 samples included in the meta-analysis, information on true and false positives and negatives was only reported for 11 samples within the manuscript itself. Information for the remaining 24 samples was retrieved from author responses on request; this meta-analysis therefore includes a reasonable amount of unpublished data.

Another limitation of this review is the moderate amount of heterogeneity between the studies included in this meta-analysis, perhaps due to variations in cut-off scores used for risk classifications, given that a number of other possible confounding variables were investigated in meta-regression. This variation is generally not avoidable, especially in reviews of prognostic (as opposed to diagnostic) studies, due to the unknown outcome, but the use of a random-effects meta-analytic model has helped to account for this variation. Furthermore, the same cut-off scores for each sample of the same instrument were applied where possible.

4.6 Conclusions and future research

This review is a comprehensive summary and meta-analysis of research regarding violence risk prediction in forensic psychiatric inpatient samples. Future research on violence risk assessment in these settings should focus more on imminent tools and less on the predictive accuracy of longer-term tools, in particular the HCR-20. For inpatient violence, prediction is more informative for violence management if it is for a shorter follow-up period; there should be more focus on the validation of the BVC, DASA and any newly developed short-term prediction tools.

With regards to clinical recommendations, the findings from this review support the use of the BVC and the DASA for nurses and other clinical staff in daily practice on forensic psychiatric wards to assess the risk of inpatient violence over a 24-hour period. However, they are not directly linked to risk management strategies, which clinical teams will need to consider alongside these risk assessments. These tools are also brief checklists that can be easily integrated into daily routine on the ward and do not require expertise or training. The more time-intensive tools that have been validated in outpatient settings, such as the HCR-20 and VRAG, should not be used as they do not perform well for the prediction of forensic inpatient violence.

From a general methodological perspective, future work in the area of violence risk assessment should report as many different estimates of predictive accuracy as possible in order to form a complete evaluation of a tool. Each accuracy estimate has a limitation of some kind, so cannot be solely trusted as a reliable or valid measure of the overall accuracy of an instrument. Studies should aim to report the sensitivity, specificity, positive and negative predictive values, diagnostic odds ratio and AUC value. Furthermore, reporting of numbers of true and false positives and negatives in primary studies would be useful for future summative analysis.

Future studies should also attempt to focus on a single population in order to ascertain the best particular risk assessment instrument to be used for that specific population. It is clear from this review that tools aimed for general application to a wide range of populations do not necessarily perform as well for specific populations. In this case, the HCR-20 is an instrument that is widely used for community settings, but for forensic inpatients it performs poorly. A focus on the predictive accuracy of specific tools designed for specific contexts is required (Douglas et al., 2017).

In order for clinical practice concerning violence risk management to be conducted effectively, it is necessary for nurses and clinicians to take into account the limitations of the violence risk assessment tools that they are using, based on validated evidence from research. When risk assessments are conducted and violence prevention or management strategies are subsequently put in place (or not), their limitations should be widely understood. The predictive ability of violence risk assessment tools is not perfect and is presented poorly in the literature due to inadequate reporting of accuracy estimates. Therefore, it is possible that those who are assessing violence risk may be more reliant on the scores of the violence risk assessment than may be appropriate given the tool's accuracy (Douglas et al., 2017).

Finally, it is the case that the goal is not violence prediction, but rather violence management and prevention. Many studies have focused on prediction models and risk

assessment tools, but very few focus on the subsequent practices that are implemented in order to manage this future violence. Instruments such as the BVC and DASA consist entirely of dynamic, changeable risk factors. Therefore, there is scope for targeted interventions to minimise violence on the ward once an individual is predicted to be high risk. This should be the primary endeavour following risk assessment. In conclusion, it is not enough to merely "be aware" of the risk posed by certain individuals; there should be specific plans of action and strategies in place to target the factors that are showing up as particularly problematic. Future work should seek to emphasis the importance of the link between violence prediction and risk management in both practical environments and research.

References

Abderhalden, Christoph, Needham, Ian, Dassen, Theo, Halfens, Ruud, Haug, Hans-Joachim, & Fischer, Joachim E. (2008). Structured risk assessment and violence in acute psychiatric wards: randomised controlled trial. *The British Journal of Psychiatry, 193*(1), 44-50.

Almvik, R, & Woods, P. (1998). The Brøset Violence Checklist (BVC) and the prediction of inpatient violence: some preliminary results. *Psychiatric Care, 5*(6), 208-211.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC.

Anderson, Tanya R, Bell, Carl C, Powell, Traci E, Williamson, Johnny L, & Blount, Morris A. (2004). Assessing psychiatric patients for violence. *Community Mental Health Journal, 40*(4), 379-399.

Andrews, DA, & Bonta, J. (1994). Psychology of criminal behavior. *Cincinnati, OH: Anderson*.

Andrews, DA, & Bonta, J. (1995). LSI-R: The Level of Service Inventory-Revised. Toronto, Ontario, Canada: Multi-Health Systems: Inc.

Arai, K., Takano, A., Nagata, T., & Hirabayashi, N. (2016). Predictive accuracy of the Historical-Clinical-Risk Management-20 for violence in forensic psychiatric wards in Japan. *Crim Behav Ment Health*. doi: 10.1002/cbm.2007

Arango, Celso, Calcedo Barba, Alfredo, González-Salvador, Teresa, & Calcedo Ordóñez, Alfredo. (1999). Violence in inpatients with schizophrenia: A prospective study. *Schizophrenia Bulletin, 25*(3), 493.

Bartels, Stephen J, Drake, Robert E, Wallach, Michael A, & Freeman, Daniel H. (1991). Characteristic hostility in schizophrenic outpatients. *Schizophrenia bulletin, 17*(1), 163.

Bjorkly, S., Hartvig, P., Heggen, F. A., Brauer, H., & Moger, T. A. (2009). Development of a brief screen for violence risk (V-RISK-10) in acute and general psychiatry: An introduction with

emphasis on findings from a naturalistic test of interrater reliability. *Eur Psychiatry, 24*(6), 388-394. doi: 10.1016/j.eurpsy.2009.07.004

Bowers, Allan, Teresa, Simpson, Alan, Nijman, Henk, & Warren, Jonathan. (2007). Adverse incidents, patient flow and nursing workforce variables on acute psychiatric wards: The Tompkins Acute Ward Study. *International Journal of Social Psychiatry, 53*(1), 75-84.

Bowers, Stewart, D, Papadopoulos, C, Dack, C, Ross, J, & Khanom, H. (2011). Inpatient violence and aggression: A literature review.

Brenner, HERMANN, & Gefeller, OLAF. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in medicine, 16*(9), 981-991.

Caldwell, Michael F. (1992). Incidence of PTSD among staff victims of patient violence. *Psychiatric Services, 43*(8), 838-839.

Campbell, Mary Ann, French, Sheila, & Gendreau, Paul. (2009). The prediction of violence in adult offenders a meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36*(6), 567-590. doi: 10.1177/0093854809333610

Chan, Oliver, & Chow, Kavin Kit-wan. (2014). Assessment and determinants of aggression in a forensic psychiatric institution in Hong Kong, China. *Psychiatry research, 220*(1), 623-630. doi: 10.1016/j.psychres.2014.08.008

Chaplin, Robert, McGeorge, Maureen, & Lelliott, Paul. (2006). The National Audit of Violence: in-patient care for adults of working age. *The Psychiatrist, 30*(12), 444-446.

Cheung, Peter, Schweitzer, Isaac, Crowley, Kathleen, & Tuckwell, Virginia. (1997). Violence in schizophrenia: role of hallucinations and delusions. *Schizophrenia research, 26*(2), 181-190.

Chu, C. M., Daffern, M., & Ogloff, J. R. P. (2013). Predicting aggression in acute inpatient psychiatric setting using BVC, DASA, and HCR-20 Clinical scale. 2. Retrieved (Chu) Clinical and Forensic Psychology Branch, Ministry of Social and Family Development, Singapore, Singapore, 24,

from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed15&NEWS=N&AN=52473733

Daffern, Michael. (2007). The predictive validity and practical utility of structured schemes used to assess risk for aggression in psychiatric inpatient settings. *Aggression and Violent Behavior, 12*(1), 116-130. doi: 10.1016/j.avb.2006.03.005

Daffern, Michael, & Howells, Kevin. (2002). Psychiatric inpatient aggression: A review of structural and functional assessment approaches. *Aggression and violent behavior, 7*(5), 477-497.

Daffern, Michael, & Howells, Kevin. (2007). The prediction of imminent aggression and self-harm in personality disordered patients of a high security hospital using the HCR-20 Clinical Scale and the Dynamic Appraisal of Situational Aggression. 2. Retrieved Personality Disorders [3217], 6, from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc5&NEWS=N&AN=2008-17312-005

de Vogel, & de Ruiter. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. 3. Retrieved (de Vogel) Dr. Henri van der Hoeven Kliniek, Department of Research, P.O. Box 174, 3500 AD Utrecht, Netherlands, 12, from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed10&NEWS=N&AN=40873362

De Vogel, & De Ruiter, Corine. (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*(3), 321-336.

de Vries Robbe, Michiel, de Vogel, Vivienne, Wever, Edwin C., Douglas, Kevin S., & Nijman, Henk L. I. (2016). Risk and protective factors for inpatient aggression. 10. Retrieved Abidin, Z., Davoren, M., Naughton, L., Gibbons, O., Nulty, A., & Kennedy, H. G. (2013). Susceptibility (risk and protective) factors for in-patient violence and self-harm: Prospective study of structured professional judgement instruments START and SAPROF, DUNDRUM-3 and

DUNDRUM-4 in forensic mental health services. BMC Psychiatry, 13http://dx.doi.org/10.1186/1471-244X-13-197, 43, from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc13&NEWS=N&AN=2016-42126-005

Deeks, Jonathan J. (2001). Systematic reviews of evaluations of diagnostic and screening tests. *BMJ: British Medical Journal, 323*(7305), 157.

Desmarais, Sarah L, Nicholls, Tonia L, Wilson, Catherine M, & Brink, Johann. (2012). Using dynamic risk and protective factors to predict inpatient aggression: reliability and validity of START assessments. *Psychological assessment, 24*(3), 685. doi: 10.1037/a0026668

Devillé, Walter L, Buntinx, Frank, Bouter, Lex M, Montori, Victor M, De Vet, Henrica CW, Van der Windt, Danielle AWM, & Bezemer, P Dick. (2002). Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC medical research methodology, 2*(1), 9.

Dinnes, J, Deeks, J, Kirby, J, & Roderick, P. (2005). A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy.

Dolan, M., Fullam, R., Logan, C., & Davies, G. (2008). The Violence Risk Scale Second Edition (VRS-2) as a predictor of institutional violence in a British forensic inpatient sample. *Psychiatry Res, 158*(1), 55-65. doi: 10.1016/j.psychres.2006.08.014

Douglas, Hart, SD, Webster, CD, & Belfrage, H. (2013). HCR-20 version 3: assessing risk for violence. *Burnaby, BC, Canada: Mental Health, Law and Policy Institute, Simon Fraser University*.

Douglas, Pugh, J., Singh, I., Savulescu, J., & Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. *European Psychiatry, 42*, 134-137.

Douglas, & Skeem, Jennifer L. (2005). Violence risk assessment: getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3), 347.

Douglas, Vincent, Gina M, & Edens, John F. (2006). Risk for criminal recidivism: The role of psychopathy.

Fagan, J., Papaconstantinou, A., Ijaz, A., Lynch, A., O'Neill, H., & Kennedy, H. G. (2009). The Suicide Risk Assessment and Management Manual (S-RAMM) Validation Study II. *Irish Journal of Psychological Medicine, 26*(3), 107-113.

Fazel, Seena, & Danesh, John. (2002). Serious mental disorder in 23 000 prisoners: a systematic review of 62 surveys. *The lancet, 359*(9306), 545-550.

Fazel, Seena, Singh, Jay P, Doll, Helen, & Grann, Martin. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *Bmj, 345*, e4692.

Ferguson, Christopher J, Averill, Patricia M, Rhoades, Howard, Rocha, Donna, Gruber, Nelson P, & Gummattira, Pushpa. (2005). Social isolation, impulsivity and depression as predictors of aggression in a psychiatric inpatient population. *Psychiatric Quarterly, 76*(2), 123-137.

Gendreau, Paul, Goggin, Claire E, & Law, Moira A. (1997). Predicting prison misconducts. *Criminal Justice and behavior, 24*(4), 414-431.

Gendreau, Paul, Goggin, Claire, & Smith, Paula. (2002). Is the PCL-R really the "unparalleled" measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior, 29*(4), 397-426.

Gendreau, Paul, Little, Tracy, & Goggin, Claire. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*(4), 575-608.

Glas, Afina S, Lijmer, Jeroen G, Prins, Martin H, Bonsel, Gouke J, & Bossuyt, Patrick MM. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology, 56*(11), 1129-1135.

Grann, Martin, Långström, Niklas, Tengström, Anders, & Kullgren, Gunnar. (1999). Psychopathy (PCL-R) predicts violent recidivism among criminal offenders with personality disorders in Sweden. *Law and human behavior, 23*(2), 205.

Gray, Nicola S, Hill, Charlotte, McGleish, Andrew, Timmons, David, MacCulloch, Malcom J, & Snowden, Robert J. (2003). Prediction of violence and self-harm in mentally disordered

offenders: a prospective study of the efficacy of HCR-20, PCL-R, and psychiatric symptomatology. *Journal of consulting and clinical psychology, 71*(3), 443.

Grisso, Thomas, Davis, Jeffrey, Vesselinov, Roumen, Appelbaum, Paul S, & Monahan, John. (2000). Violent thoughts and violent behavior following hospitalization for mental disorder. *Journal of Consulting and Clinical Psychology, 68*(3), 388.

Gulati, Gautam, Cornish, Robert, Al-Taiar, Hasanen, Miller, Christopher, Khosla, Vivek, Hinds, Christopher, . . . Fazel, Seena. (2016). Web-based violence risk monitoring tool in psychoses: pilot study in community forensic patients. *Journal of forensic psychology practice, 16*(1), 49-59.

Gunenc, Cevher, O'Shea, Laura E, & Dickens, Geoffrey L. (2017). Structured risk assessment for reduction of multiple risk outcomes in a secure mental health setting: Use of the START. *Criminal Behaviour and Mental Health*.

Harbord, Roger M, Deeks, Jonathan J, Egger, Matthias, Whiting, Penny, & Sterne, Jonathan AC. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8*(2), 239-251.

Hare, Robert D. (1991). *The Hare psychopathy checklist-revised: Manual*: Multi-Health Systems, Incorporated.

Hart, Stephen David, Cox, David Neil, & Hare, R D. (1995). *The Hare psychopathy checklist: Screening version (PCL: SV)*: MSH-Multi-Health Systems, Incorporated.

Hartvig, P, Østberg, B, Alfarnes, S, Moger, TA, Skjønberg, M, & Bjørkly, S. (2007). Violence Risk Screening-10 (V-RISK-10). *Oslo, Norway: Centre for Research and Education in Forensic Psychiatry*.

Hogan, Neil, Ennis, Liam, & Assessment, FA. (2010). Assessing risk for forensic psychiatric inpatient violence: A meta-analysis. *Open Access Journal of Forensic Psychology, 2*, 137-147.

Hoptman, Matthew J, Yates, Kathy F, Patalinjug, Marilou B, Wack, Renate C, & Convit, Antonio. (1999). Clinical prediction of assaultive behavior among male psychiatric patients at a maximum-security forensic facility. *Psychiatric Services, 50*(11), 1461-1466.

Hurducas, Claudia C, Singh, Jay P, de Ruiter, Corine, & Petrila, John. (2014). Violence risk assessment tools: A systematic review of surveys. *International Journal of Forensic Mental Health, 13*(3), 181-192.

Hvidhjelm, Jacob, Sestoft, Dorte, Skovgaard, Lene Theil, & Bue Bjorner, Jakob. (2014). Sensitivity and specificity of the Brøset Violence Checklist as predictor of violence in forensic psychiatry. *Nordic Journal of Psychiatry, 68*(8), 536-542. doi: 10.3109/08039488.2014.880942

Inoue, Makoto, Tsukano, Ken, Muraoka, Mitsutaro, Kaneko, Fumiko, & Okamura, Hitoshi. (2006). Psychological impact of verbal abuse and violence by patients on nurses working in psychiatric departments. *Psychiatry and Clinical Neurosciences, 60*(1), 29-36.

Jackson, Dan, White, Ian R, & Riley, Richard D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in medicine, 31*(29), 3805-3820.

Jackson, Dan, White, Ian R, & Thompson, Simon G. (2010). Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in medicine, 29*(12), 1282-1297.

Jeandarme, Inge, Pouls, Claudia, De Laender, Jan, Oei, TI, & Bogaerts, Stefan. (2017). Field validity of the HCR-20 in forensic medium security units in Flanders. *Psychology, Crime & Law, 23*(4), 305-322.

Kay, Stanley R, Wolkenfeld, Fred, & Murrill, Lisa M. (1988). Profiles of aggression among psychiatric patients: II. Covariates and predictors. *The Journal of nervous and mental disease, 176*(9), 547-557.

Kraemer, Helena Chmura, & Gibbons, Robert D. (2009). Where do we go wrong in assessing risk factors, diagnostic and prognostic tests? The problems of two-by-two association. *Psychiatric Annals, 39*(7), 711-718.

Large, Matthew, & Nielssen, Olav. (2017). The limitations and future of violence risk assessment. *World psychiatry, 16*(1), 25-26.

Lauvrud, Christian, Nonstad, Kåre, & Palmstierna, Tom. (2009). Occurrence of post traumatic stress symptoms and their relationship to professional quality of life (ProQoL) in nursing staff at a forensic psychiatric security unit: a cross-sectional study. *Health and quality of life outcomes, 7*(1), 31.

Lee, Juneyoung, Kim, Kyung Won, Choi, Sang Hyun, Huh, Jimi, & Park, Seong Ho. (2015). Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part II. statistical methods of meta-analysis. *Korean journal of radiology, 16*(6), 1188-1196.

Li, Jialiang, & Fine, Jason P. (2011). Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics*, kxr008.

Lijmer, Jeroen G, Bossuyt, Patrick MM, & Heisterkamp, Siem H. (2002). Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in medicine, 21*(11), 1525-1537.

Linaker, Olav M, & Busch-Iversen, H. (1995). Predictors of imminent violence in psychiatric inpatients. *Acta Psychiatrica Scandinavica, 92*(4), 250-254.

Lipsey, Mark W, Wilson, David B, Cohen, Mark A, & Derzon, James H. (2002). Is there a causal relationship between alcohol use and violence? *Recent developments in alcoholism* (pp. 245-282): Springer.

Littenberg, Benjamin, & Moses, Lincoln E. (1993). Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making, 13*(4), 313-321.

Macaskill, Petra, Gatsonis, Constantine, Deeks, Jonathan, Harbord, Roger, & Takwoingi, Yemisi. (2010). Cochrane handbook for systematic reviews of diagnostic test accuracy. *Version 0.9. 0. London: The Cochrane Collaboration*.

McDermott, Dualan, Isah V, & Scott, Charles L. (2011). The predictive ability of the Classification of Violence Risk (COVR) in a forensic psychiatric hospital. *Psychiatric Services, 62*(4), 430-433. doi: 10.1176/appi.ps.62.4.430

10.1176/ps.62.4.pss6204_0430

McDermott, Edens, J. F., Quanbeck, C. D., Busse, D., & Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: variable- and person-focused analyses. *Law Hum Behav, 32*(4), 325-338. doi: 10.1007/s10979-007-9094-8

McNiel, Dale E, & Binder, Renee L. (1994). Screening for risk of inpatient violence: Validation of an actuarial tool. *Law and Human Behavior, 18*(5), 579.

Menzies, Robert, & Webster, Christopher D. (1995). Construction and validation of risk assessments in a six-year follow-up of forensic patients: A tridimensional analysis. *Journal of Consulting and Clinical Psychology, 63*(5), 766.

Moher, David, Liberati, Alessandro, Tetzlaff, Jennifer, Altman, Douglas G, & Group, Prisma. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med, 6*(7), e1000097.

Monahan, John, Steadman, Henry J, Robbins, Pamela Clark, Appelbaum, Paul, Banks, Steven, Grisso, Thomas, . . . Silver, Eric. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric services, 56*(7), 810-815.

Monahan, John, Steadman, Henry J, Silver, Eric, Appelbaum, Paul S, Robbins, Pamela Clark, Mulvey, Edward P, . . . Banks, Steven. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*: Oxford University Press.

Moons, Karel GM, de Groot, Joris AH, Bouwmeester, Walter, Vergouwe, Yvonne, Mallett, Susan, Altman, Douglas G, . . . Collins, Gary S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med, 11*(10), e1001744.

Moses, Lincoln E, Shapiro, David, & Littenberg, Benjamin. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in medicine, 12*(14), 1293-1316.

Mudde, N., Nijman, H., van der Hulst, W., & van den Bout, J. (2011). [Predicting aggression during the treatment of forensic psychiatric patients by means of the HCR-20]. *Tijdschr Psychiatr, 53*(10), 705-713.

Naaktgeboren, Christiana A., Ochodo, Eleanor A., Van Enst, Wynanda A., de Groot, Joris A. H., Hooft, Lotty, Leeflang, Mariska M. G., . . . Reitsma, Johannes B. (2016). Assessing variability in results in systematic reviews of diagnostic studies. *BMC Medical Research Methodology, 16*, 6. doi: 10.1186/s12874-016-0108-4

Needham, Ian, Abderhalden, Christoph, Meer, R, Dassen, Theo, Haug, HJ, Halfens, RJG, & Fischer, Joachim E. (2004). The effectiveness of two interventions in the management of patient violence in acute mental inpatient settings: report on a pilot study. *Journal of psychiatric and mental health nursing, 11*(5), 595-601.

Negredo, Laura, Melis, Francesca, & Herrero, Óscar. (2015). Riesgo de violencia institucional y comunitaria en delincuentes con trastorno mental. *Anuario de Psicología Jurídica, 25*(1), 21-27.

Nicholls, Tonia Lee. (1997). *Comparing risk assessments with female and male civil psychiatric patients: the utility of the HCR-20 and PCL: SV.* Simon Fraser University.

Nonstad, Kåre, Nesset, Merete B, Kroppan, Erik, Pedersen, Truls W, Nøttestad, Jim Aa, Almvik, Roger, & Palmstierna, Tom. (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*(4), 294-299.

Ochodo, Eleanor A, Reitsma, Johannes B, Bossuyt, Patrick M, & Leeflang, Mariska MG. (2013). Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. *Journal of clinical epidemiology, 66*(11), 1281-1288.

Ogloff, J, & Daffern, M. (2002). Dynamic appraisal of situational aggression: Inpatient version. *Melbourne, Victoria, Australia: Monash University and Forensicare*.

Quinsey, Harris, Rice, & Cormier. (2006a). Actuarial prediction of violence.

Quinsey, Harris, Rice, & Cormier. (2006b). Violent offenders: appraising and managing risk. *Washington, DC*.

Reitsma, Johannes B, Glas, Afina S, Rutjes, Anne WS, Scholten, Rob JPM, Bossuyt, Patrick M, & Zwinderman, Aeilko H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology, 58*(10), 982-990.

Richter, Dirk, & Berger, Klaus. (2006). Post-traumatic stress disorder following patient assaults among staff members of mental health hospitals: a prospective longitudinal study. *BMC psychiatry, 6*(1), 15.

Rutter, Carolyn M, & Gatsonis, Constantine A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine, 20*(19), 2865-2884.

Schueler, Sabine, Schuetz, Georg M, & Dewey, Marc. (2012). The revised QUADAS-2 tool. *Annals of internal medicine, 156*(4), 323.

Schwartz, David, Dodge, Kenneth A, Coie, John D, Hubbard, Julie A, Cillessen, Antonius HN, Lemerise, Elizabeth A, & Bateman, Helen. (1998). Social-cognitive and behavioral correlates of aggression and victimization in boys' play groups. *Journal of abnormal child psychology, 26*(6), 431-440.

Singh. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law, 31*(1), 8-22.

Singh, Bjorkly, Stal, & Fazel, Seena. (2016). International perspectives on violence risk assessment. *International perspectives on violence risk assessment.* Retrieved Forensic Psychology & Legal Issues [4200],

from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc13&NEWS=N&AN=2016-40455-000

Singh, Desmarais, Otto, Nicholls, Petersen, & Pritchard. (2016). The International Risk Survey: Use and Perceived Utility of Structured Violence Risk Assessment Tools in 44 Countries. *International Perspectives on Violence Risk Assessment*, 101.

Singh, Desmarais, Sarah L, Hurducas, Cristina, Arbach-Lucioni, Karin, Condemarin, Carolina, Dean, Kimberlie, . . . Grann, Martin. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13*(3), 193-206.

Singh, & Fazel, Seena. (2010). Forensic risk assessment: A metareview. *Criminal Justice and Behavior, 37*(9), 965-988.

Singh, Grann, Martin, & Fazel, Seena. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical psychology review, 31*(3), 499-513.

Snowden, R. J., Gray, N. S., Taylor, J., & Fitzgerald, S. (2009). Assessing risk of future violence among forensic psychiatric inpatients with the classification of violence risk (COVR). 11. Retrieved (Snowden, Gray, Taylor, Fitzgerald) School of Psychology, Cardiff University, Park Place, Cardiff, CF10 3AT, United Kingdom, 60, from http://psychservices.psychiatryonline.org/cgi/reprint/60/11/1522http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=355573242

StataCorp. (2015). Stata Statistical Software: Release 14: College Station, TX: StataCorp LP.

Steyerberg, Ewout W, Vickers, Andrew J, Cook, Nancy R, Gerds, Thomas, Gonen, Mithat, Obuchowski, Nancy, . . . Kattan, Michael W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.), 21*(1), 128.

Swanson, Jeffrey W, Borum, Randy, Swartz, Marvin S, & Monahan, John. (1996). Psychotic symptoms and disorders and the risk of violent behaviour in the community. *Criminal Behaviour and Mental Health, 6*(4), 309-329.

Swanson, Jeffrey W, Holzer III, Charles E, Ganju, Vijay K, & Jono, Robert Tsutomu. (1990). Violence and psychiatric disorder in the community: evidence from the Epidemiologic Catchment Area surveys. *Psychiatric Services, 41*(7), 761-770.

Tape, Thomas G. (2006). The area under an ROC curve. *Interpreting diagnostic tests*.

Tengström, Anders, Grann, Martin, Långström, Niklas, & Kullgren, Gunnar. (2000). Psychopathy (PCL-R) as a predictor of violent recidivism among criminal offenders with schizophrenia. *Law and Human Behavior, 24*(1), 45.

Thomson, Lindsay, Davidson, Michelle, Brett, Caroline, Steele, Jonathan, & Darjee, Rajan. (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. 2. Retrieved Psychological Disorders [3210], 7, from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc6&NEWS=N&AN=2008-17325-007

Trikalinos, Thomas A, Balion, Cynthia M, Coleman, Craig I, Griffith, Lauren, Santaguida, Pasqualina L, Vandermeer, Ben, & Fu, Rongwei. (2012). meta-analysis of test performance when there is a "gold standard". *Journal of general internal medicine, 27*(1), 56-66.

Verde, Pablo Emilio. (2008). Meta-analysis of diagnostic test data: modern statistical approaches. *Deutsche Nationalbibliothek*.

Wang, Eugene W, & Diamond, Pamela M. (1999). Empirically identifying factors related to violence risk in corrections. *Behavioral Sciences & the Law, 17*(3), 377-389.

Webster, Douglas, KS, Eaves, D, & Hart, SD. (1997). HCR-20: Assessing Risk for Violence (Version 2). Burnaby, British Columbia, Canada, Simon Fraser University. *Mental Health, Law, and Policy Institute*.

Webster, Martin, M, Brink, Johann, Nicholls, T, & Middleton, C. (2004). START: The Short-term assessment of risk and treatability. *Hamilton: St Joseph's Healthcare*.

Webster, Martin, ML, Brink, J, Nicholls, TL, & Desmarais, SL. (2009). Manual for the Short-Term Assessment of Risk and Treatability (START)(Version 1.1). *Coquitlam, Canada: British Columbia Mental Health & Addiction Services*.

White, Ian R. (2011). Multivariate random-effects meta-regression: updates to mvmeta. *Stata Journal, 11*(2), 255.

Whiting, Penny F, Rutjes, Anne WS, Westwood, Marie E, Mallett, Susan, Deeks, Jonathan J, Reitsma, Johannes B, . . . Bossuyt, Patrick MM. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine, 155*(8), 529-536.

Whittington, R, Hockenhull, JC, McGuire, J, Leitner, M, Barr, W, Cherry, MG, . . . Dickson, R. (2013). A systematic review of risk assessment strategies for populations at high risk of engaging in violent behaviour: update 2002–8.

Wildgoose, Joanna, Briscoe, Martin, & Lloyd, Keith. (2003). Psychological and emotional problems in staff following assaults by patients. *The Psychiatrist, 27*(8), 295-297.

Witt, Katrina, Van Dorn, Richard, & Fazel, Seena. (2013). Risk factors for violence in psychosis: systematic review and meta-regression analysis of 110 studies. *PloS one, 8*(2), e55942.

Wong, S, & Gordon, A. (2000). Violence Risk Scale (VRS). *Saskatoon, Saskatchewan*.

Zhou, Yan, & Dendukuri, Nandini. (2014). Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: The case of meta-analyses of diagnostic accuracy. *Statistics in medicine, 33*(16), 2701-2717.

Appendix

Appendix 1 PRISMA checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	9
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	20
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	21
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	23
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	29
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	29
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	29

Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	29
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	34
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	34
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	33
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	35
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	38
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	41
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	41
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	32
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	43
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	33
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	43
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	47
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	53
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	54

DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	62
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	71
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	73
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	3

Appendix 2 QUADAS-2 – quality assessment tool for systematic reviews using primary studies of diagnostic test accuracy

QUADAS-2

Phase 1: State the review question:

Patients (setting, intended use of index test, presentation, prior testing):
Index test(s):
Reference standard and target condition:

Phase 2: Draw a flow diagram for the primary study

Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION

A. Risk of Bias

Describe methods of patient selection:

- ❖ Was a consecutive or random sample of patients enrolled? Yes/No/Unclear
- ❖ Was a case-control design avoided? Yes/No/Unclear
- ❖ Did the study avoid inappropriate exclusions? Yes/No/Unclear

Could the selection of patients have introduced bias? **RISK: LOW/HIGH/UNCLEAR**

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting)**:**

Is there concern that the included patients do not match the review question? **CONCERN: LOW/HIGH/UNCLEAR**

DOMAIN 2: INDEX TEST(S)

If more than one index test was used, please complete for each test.

A. Risk of Bias

Describe the index test and how it was conducted and interpreted:

- ❖ Were the index test results interpreted without knowledge of the results of the reference standard? Yes/No/Unclear
- ❖ If a threshold was used, was it pre-specified? Yes/No/Unclear

Could the conduct or interpretation of the index test have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

B. Concerns regarding applicability

Is there concern that the index test, its conduct, or interpretation differ from the review question? **CONCERN: LOW /HIGH/UNCLEAR**

DOMAIN 3: REFERENCE STANDARD

A. Risk of Bias

Describe the reference standard and how it was conducted and interpreted:

- ❖ Is the reference standard likely to correctly classify the target condition? Yes/No/Unclear
- ❖ Were the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question? **CONCERN: LOW /HIGH/UNCLEAR**

DOMAIN 4: FLOW AND TIMING

A. Risk of Bias

Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):

Describe the time interval and any interventions between index test(s) and reference standard:

- ❖ Was there an appropriate interval between index test(s) and reference standard? Yes/No/Unclear
- ❖ Did all patients receive a reference standard? Yes/No/Unclear
- ❖ Did patients receive the same reference standard? Yes/No/Unclear
- ❖ Were all patients included in the analysis? Yes/No/Unclear

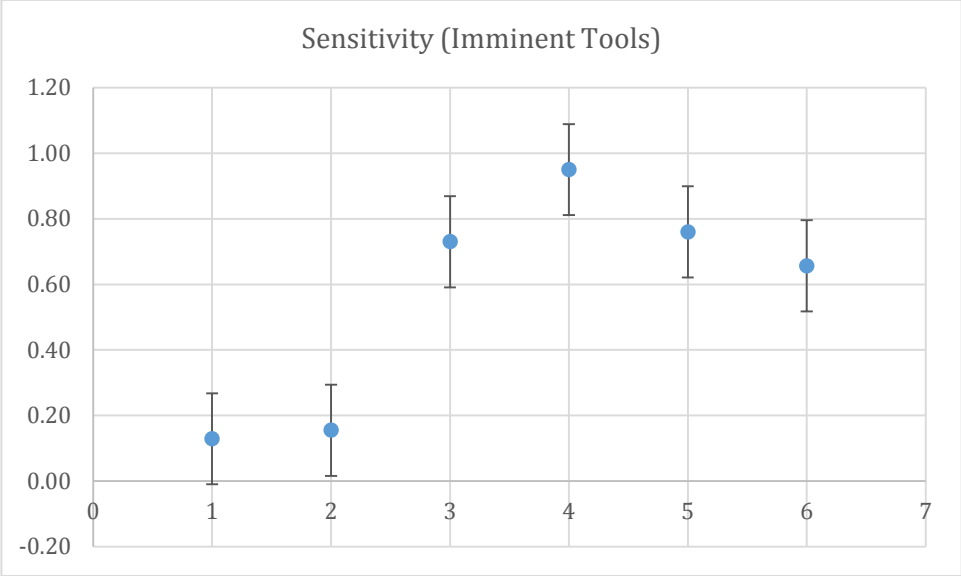
Could the patient flow have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

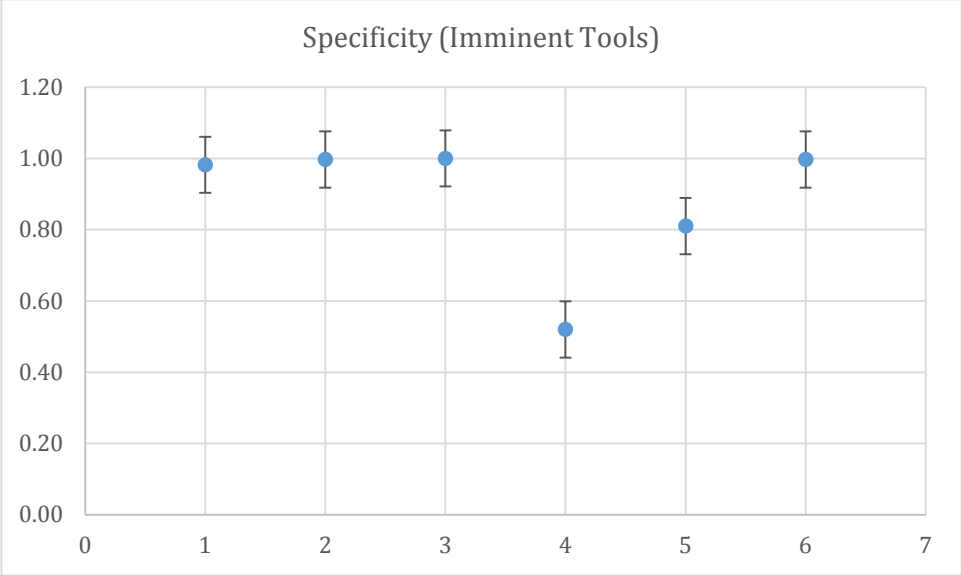
Appendix 3 CHARMS checklist

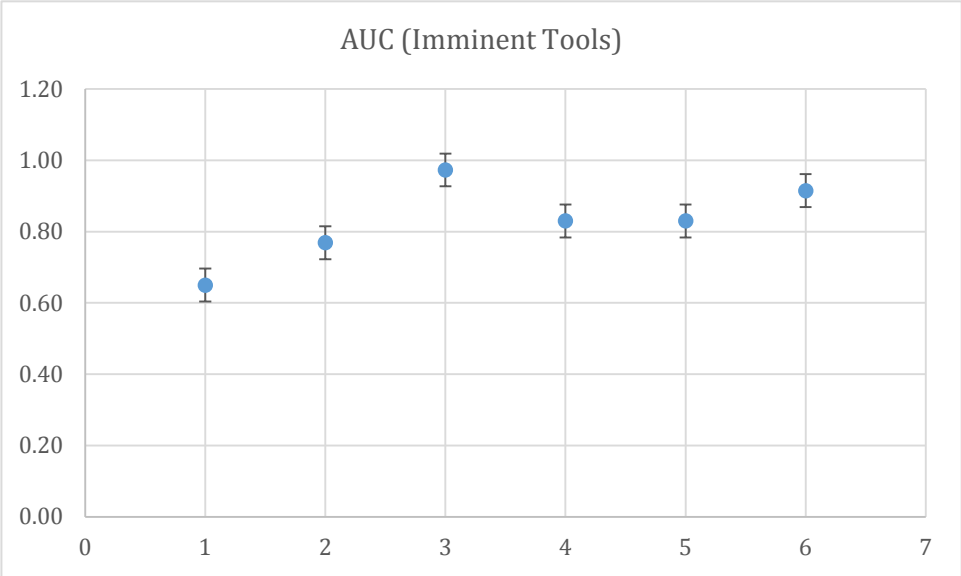
Domain	Key items	Reported on page #
SOURCE OF DATA	Source of data (e.g., cohort, case-control, randomized trial participants, or registry data)	
PARTICIPANTS	Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria)	
	Participant description	
	Details of treatments received, if relevant	
	Study dates	
OUTCOME(S) TO BE PREDICTED	Definition and method for measurement of outcome	
	Was the same outcome definition (and method for measurement) used in all patients?	
	Type of outcome (e.g., single or combined endpoints)	
	Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?	
	Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?	
	Time of outcome occurrence or summary of duration of follow-up	
CANDIDATE PREDICTORS (OR INDEX TESTS)	Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)	
	Definition and method for measurement of candidate predictors	
	Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)	
	Were predictors assessed blinded for outcome, and for each other (if relevant)?	
	Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or categorised)	
SAMPLE SIZE	Number of participants and number of outcomes/events	
	Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable)	
MISSING DATA	Number of participants with any missing value (include predictors and outcomes)	
	Number of participants with missing data for each predictor	
	Handling of missing data (e.g., complete-case analysis, imputation, or other methods)	
MODEL DEVELOPMENT	Modelling method (e.g., logistic, survival, neural network, or machine learning techniques)	
	Modelling assumptions satisfied	
	Method for selection of predictors **for inclusion** in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)	
	Method for selection of predictors **during multivariable modelling** (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)	
	Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)	
MODEL PERFORMANCE	Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals	
	Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used	
MODEL EVALUATION	Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators)	
	In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)	
RESULTS	Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)	
	Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance	
	Comparison of the distribution of predictors (including missing data) for development and validation datasets	
INTERPRETATION AND DISCUSSION	Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)	
	Comparison with other studies, discussion of generalizability, strengths and limitations.	

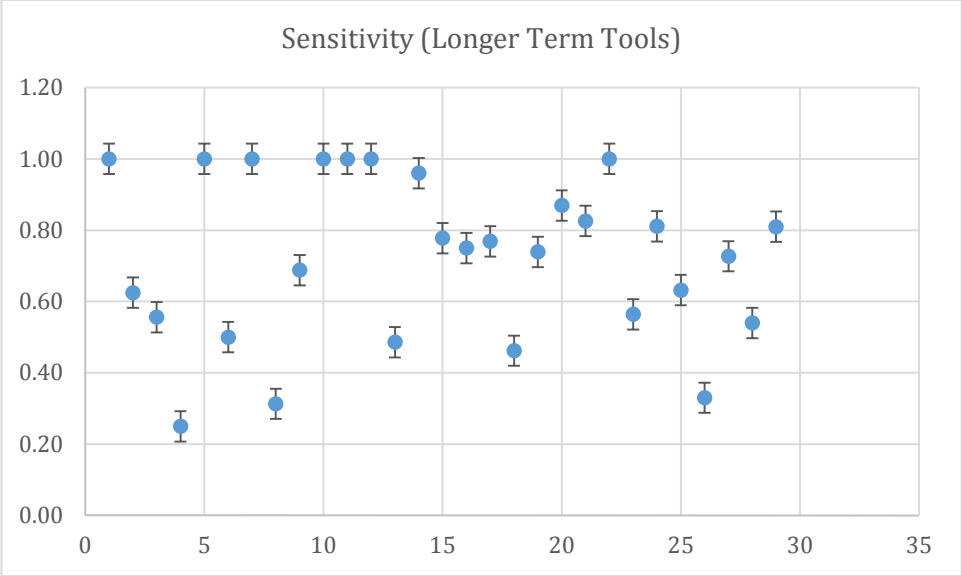
Appendix 4 Plots of accuracy measures for each study in the two groups for meta-analysis.

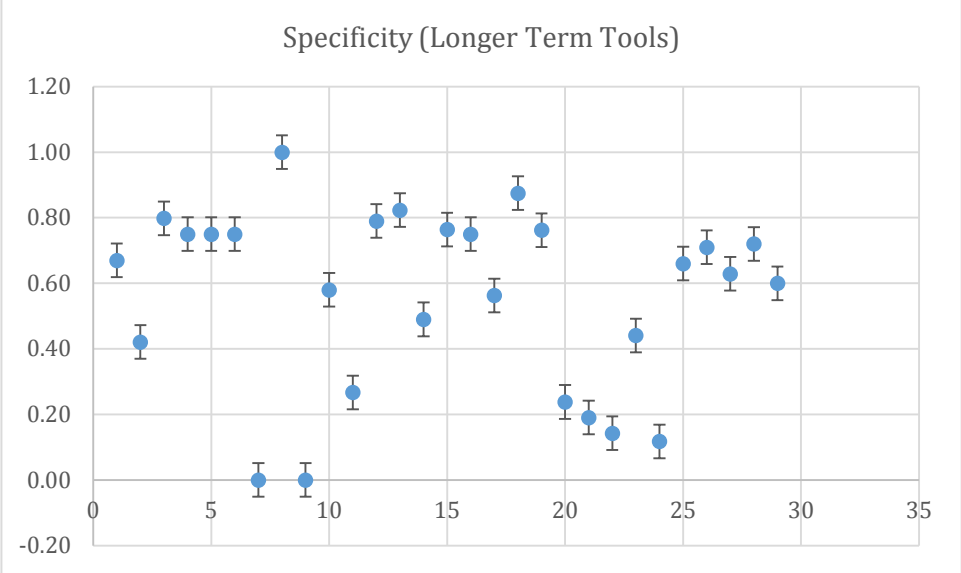
Accuracy measure (Y axis) plotted against study number (X axis), with standard error bars

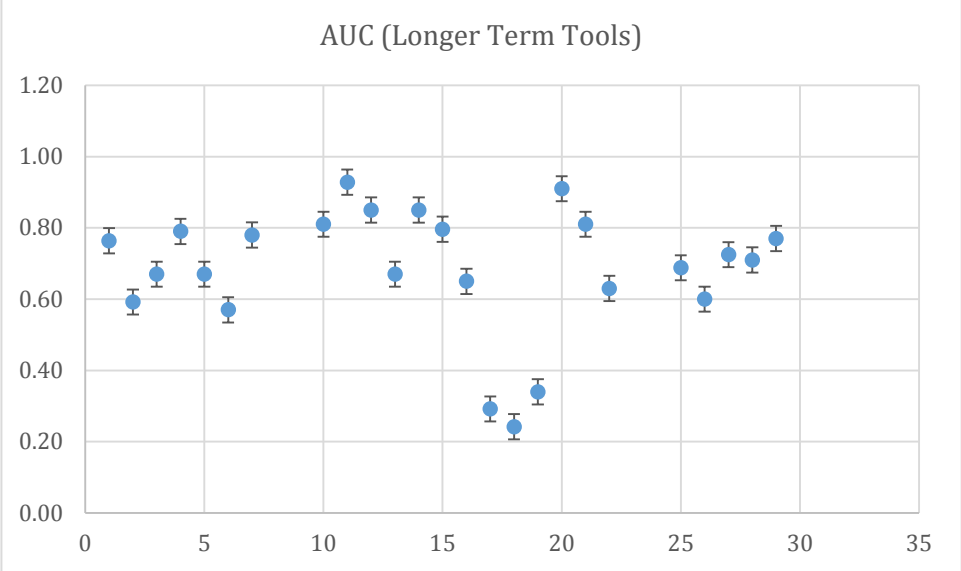







Sensitivity (Longer Term Tools)
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0
5
10
15
20
25
30
35


Specificity (Longer Term Tools)
1.20
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
0
5
10
15
20
25
30
35


AUC (Longer Term Tools)
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0
5
10
15
20
25
30
35

Appendix 5 Table showing studies included in meta-analysis with study- and sample-related variables

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Arai (2016)	HCR-20	93	Male	n/a	12.9	Interpersonal	n/a	n/a	180	Sens: 1.00 Spec: 0.58 PPV: 0.26 NPV: 1.00 DOR: n/a AUC: 0.81	Low
	HCR-20	15	Female	n/a	20.0	Interpersonal	n/a	n/a	180	Sens: 1.00 Spec: 0.67 PPV: 0.43 NPV: 1.00 DOR: n/a AUC: 0.76	Low
Chan (2014)	BVC	530	Mixed	39.2	17.7	Including verbal threat	Psychotic disorder (25%); Personality disorder (12%)	46.2	1	Sens: 0.16 Spec: 1.00 PPV: 0.98 NPV: 0.56 DOR: 70.09 AUC: 0.77	Low
	DASA	530	Mixed	39.2	17.7	Including verbal threat	Psychotic disorder (25%)	46.2	1	Sens: 0.73 Spec: 1.00 PPV: 0.91 NPV: 0.99 DOR: n/a AUC: 0.97	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Chu (2013)	BVC	70	Mixed	34.3	22.9	Interpersonal	Psychotic disorder (80%)	65.7	1	Sens: 0.95 Spec: 0.52 PPV: 0.12 NPV: 0.99 DOR: 16.15 AUC: 0.83	Low
	DASA	70	Mixed	34.3	22.9	Interpersonal	Psychotic disorder (80%); Personality disorder (20%)	65.7	1	Sens: 0.76 Spec: 0.81 PPV: 0.04 NPV: 1.00 DOR: 13.20 AUC: 0.83	Low
Daffern (2007)	DASA	38	Male	n/a	7.0	Including verbal threat	Personality disorder (100%)	n/a	1	Sens: 0.13 Spec: 0.98 PPV: 0.35 NPV: 0.94 DOR: 8.64 AUC: 0.65	Low
de Vogel (2005)	HCR-20	21	Male	30.7	28.6	Interpersonal	n/a	74	561	Sens: 1.00 Spec: 0.27 PPV: 0.35 NPV: 1.00 DOR: n/a AUC: 0.93	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
de Vogel (2005)	HCR-20	27	Female	33.2	29.6	Interpersonal	n/a		306	Sens: 0.63 Spec: 0.42 PPV: 0.31 NPV: 0.73 DOR: 1.21 AUC: 0.59	Low
de Vogel (2006)	HCR-20	127	Male	32.9	15.0	Interpersonal	Personality disorder (66%)	93	645	Sens: 1.00 Spec: 0.79 PPV: 0.21 NPV: 1.00 DOR: n/a AUC: 0.85	Low
de Vries Robbe (2016)	HCR-20	146	Male	n/a	11.3	Including verbal threat	n/a	n/a	365	Sens: 0.49 Spec: 0.82 PPV: 0.26 NPV: 0.93 DOR: 4.38 AUC: 0.67	Low
	HCR-20	39	Female	n/a	10.2	Including verbal threat	n/a	n/a	365	Sens: 0.56 Spec: 0.80 PPV: 0.24 NPV: 0.94 DOR: 4.92 AUC: 0.67	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Desmarais (2012)	START	120	Male	38	22.7	Interpersonal	Psychotic disorder (85%)	80	365	Sens: 0.96 Spec: 0.49 PPV: 0.36 NPV: 0.98 DOR: 24.89 AUC: 0.85	Low
Dolan (2008)	VRS-2	147	Mixed	36	52.1	Including verbal threat	Psychotic disorder (90%); Personality disorder (4%)	51	56	Sens: 0.63 Spec: 0.66 PPV: 0.68 NPV: 0.61 DOR: 3.04 AUC: 0.69	Low
Fagan (2009)	HCR-20	81	Male	n/a	11.1	Including verbal threat	n/a	n/a	184.8	Sens: 0.78 Spec: 0.76 PPV: 0.29 NPV: 0.97 DOR: 11.32 AUC: 0.80	Low
Gunenc (2017)	START	44	Male	34.3	45.5	Interpersonal	Psychotic disorder (48%)	n/a	180	Sens: 0.75 Spec: 0.75 PPV: 0.71 NPV: 0.78 DOR: 9.00 AUC: 0.65	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Hvidhjelm (2014)	BVC	156	Mixed	38	48.7	Including verbal threat	Psychotic disorder (83%); Personality disorder (3%)	85.3	1	Sens: 0.66 Spec: 1.00 PPV: 0.37 NPV: 1.00 DOR: n/a AUC: 0.92	Low
Jeandarme (2017)	HCR-20	168	Mixed	36.1	19.6	Interpersonal	Psychotic disorder (43%); Personality disorder (76%)	75.6	675.4	Sens: 0.33 Spec: 0.71 PPV: 0.22 NPV: 0.81 DOR: 1.23 AUC: 0.60	Low
McDermott (2011)	COVR	146	Mixed	46.1	15.1	Interpersonal	Psychotic disorder (78%)	63	140	Sens: 0.73 Spec: 0.63 PPV: 0.26 NPV: 0.93 DOR: 4.52 AUC: 0.73	Low
Mudde (2011)	HCR-20	102	Mixed	37.1	42.2	Including verbal threat	Psychotic disorder (73%)	70	1278	Sens: 0.54 Spec: 0.72 PPV: 0.58 NPV: 0.68 DOR: 2.84 AUC: 0.71	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Negredo (2015)	HCR-20	29	Male	38	44.8	Interpersonal	Psychotic disorder (79%); Personality disorder (21%)	74.7		Sens: 0.77 Spec: 0.56 PPV: 0.59 NPV: 0.75 DOR: 4.29 AUC: 0.29	Low
	PCL:SV	29	Male	38	44.8	Interpersonal	Psychotic disorder (79%); Personality disorder (21%)	74.7		Sens: 0.46 Spec: 0.88 PPV: 0.75 NPV: 0.67 DOR: 6.00 AUC: 0.24	Low
Nonstad (2010)	START	47	Mixed	36	34.8	Including verbal threat	Psychotic disorder (96%); Personality disorder (15%)	n/a	90	Sens: 0.81 Spec: 0.60 PPV: 0.52 NPV: 0.86 DOR: 6.50 AUC: 0.77	Low
Snowden (2009)	COVR	44	Male	n/a	52.3	Interpersonal	n/a	n/a	180	Sens: 0.74 Spec: 0.76 PPV: 0.77 NPV: 0.73 DOR: 9.07 AUC: 0.34	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Snowden (2009)	COVR	8	Female	n/a	50.0	Interpersonal	n/a	n/a	180	Sens: 0.25 Spec: 0.75 PPV: 0.50 NPV: 0.50 DOR: 1.00 AUC: 0.79	Low
	HCR-20	44	Male	n/a	52.3	Interpersonal	n/a	n/a	180	Sens: 0.87 Spec: 0.24 PPV: 0.56 NPV: 0.63 DOR: 2.08 AUC: 0.91	Low
	HCR-20	8	Female	n/a	50.0	Interpersonal	n/a	n/a	180	Sens: 1.00 Spec: 0.75 PPV: 0.80 NPV:1.00 DOR: n/a AUC: 0.67	Low
	PCL-R	44	Male	n/a	52.3	Interpersonal	n/a	n/a	180	Sens: 0.83 Spec: 0.19 PPV: 0.53 NPV: 0.50 DOR: 1.12 AUC: 0.81	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
Snowden (2009)	PCL-R	8	Female	n/a	50.0	Interpersonal	n/a	n/a	180	Sens: 0.50 Spec: 0.75 PPV: 0.67 NPV: 0.60 DOR: 3.00 AUC: 0.57	Low
	VRAG	44	Male	n/a	52.3	Interpersonal	n/a	n/a	180	Sens: 1.00 Spec: 0.14 PPV: 0.56 NPV: 1.00 DOR: n/a AUC: 0.63	Low
	VRAG	8	Female	n/a	50.0	Interpersonal	n/a	n/a	180	Sens: 1.00 Spec: 0.00 PPV: 0.50 NPV: 0.00 DOR: n/a AUC: 0.78	Low
Thomson (2008)	PCL-R	144	Male	n/a	20.1	Including verbal threat	Psychotic disorder (100%)	n/a	2920	Sens: 0.56 Spec: 0.44 PPV: 0.77 NPV: 0.24 DOR: 1.02 AUC: n/a	Low

First author (year)	Tool	Sample size	Gender	Mean age (years)	% of sample that were violent	Violent outcome measured	Diagnoses	% with violent index offence	Length of follow-up (days)	Accuracy measures	QUADAS (Low or High)
	PCL-R	145	Male	n/a	20.0	Including verbal threat	Psychotic disorder (100%)	n/a	2920	Sens: 0.31 Spec: 1.00 PPV: 1.00 NPV: 1.00 DOR: n/a AUC: n/a	Low
Thomson (2008)	VRAG	17	Female	n/a	17.7	Including verbal threat	Psychotic disorder (100%)	n/a	2920	Sens: 0.81 Spec: 0.12 PPV: 0.75 NPV: 0.16 DOR: 0.57 AUC: n/a	Low
	VRAG	17	Female	n/a	17.7	Including verbal threat	Psychotic disorder (100%)	n/a	2920	Sens: 0.69 Spec: 0.00 PPV: 0.92 NPV: 0.00 DOR: n/a AUC: n/a	Low