Reliability and validity of the FORUM-P and FORUM-C: two novel instruments for outcome measurement in forensic mental health

Howard Ryland, Jonathan Cook, Oana Ciobanasu, Oluwasegun Oluwabamise, Robert Cornish, Hasanen Al-Taiar, Eberechi Chris-Okoro, Catherine Hoggart, Alex Vallakalil, Ray Fitzpatrick & Seena Fazel

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS   Check for updates

Reliability and validity of the FORUM-P and FORUM-C: two novel instruments for outcome measurement in forensic mental health

Howard Ryland [a], Jonathan Cook[b], Oana Ciobanasu[c], Oluwasegun Oluwabamise[d], Robert Cornish[a,d], Hasanen Al-Taiar[d], Eberechi Chris-Okoro[d], Catherine Hoggart[d], Alex Vallakalil[d], Ray Fitzpatrick[e] and Seena Fazel[a]

[a]Department of Psychiatry, University of Oxford, Oxford, UK; [b]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK; [c]Gloucestershire Health and Care NHS Trust, Oxford, UK; [d]Oxford Health NHS Foundation Trust, Oxford, UK; [e]Nuffield Department of Population Health, University of Oxford, Oxford, UK

ABSTRACT

We conducted a series of tests on the FORensic oUtcome Measure (FORUM), a novel tool for measuring outcomes in forensic mental health services, which consists of complementary patient-reported (FORUM-P) and clinician-reported instruments (FORUM-C). Inpatients and outpatients at a UK forensic psychiatric regional service completed the FORUM-P and members of their clinical teams completed the FORUM-C. Patients and clinicians also provided qualitative feedback on the instruments. We assessed test-retest and inter-rater reliability in standard ways. Sixty-two patients participated with a mean age of 41.0 years (standard deviation 11.3). Thirty-five clinicians provided information about these patients. For internal consistency, Cronbach's alpha for FORUM-P was 0.87 (95% confidence interval 0.80–0.93) and for FORUM-C was 0.93 (95% confidence interval 0.91–0.96). For test-retest reliability, weighted kappa for FORUM-P was 0.44 (95% confidence interval 0.24–0.63) and for FORUM-C was 0.78 (95% confidence interval 0.73–0.85). For inter-rater reliability, Spearman correlation coefficient for overall FORUM-C score between the first rating by clinician 1 and clinician 2 was 0.47 (95% confidence interval 0.18–0.69). For comprehensiveness, comprehensibility, and relevance FORUM-P and FORUM-C were both rated as good. FORUM-P and FORUM-C provide a novel, robust set of complementary instruments with promising psychometric properties for monitoring outcomes in forensic mental health.

Introduction

Forensic mental health services provide care for people with a mental illness who are considered to have an elevated risk of violence towards others on admission (Crocker

Supplemental data for this article can be accessed here https://doi.org/10.1080/1068316X.2022.2076855.
This article has been republished with minor changes. These changes do not impact the academic content of the article.

et al., 2017). Patients in such services may also have been convicted of serious crimes and can receive compulsory treatment orders in lieu of a prison sentence (Vollm et al., 2018). Patients typically progress through inpatient services of decreasing levels of security, before being discharged to the community, where they may continue to be subject to restrictions and surveillance (Latham & Williams, 2020). To be discharged, patients will usually need to demonstrate sufficiently low levels of risk to others and themselves.

Forensic mental health services are resource-intensive and are estimated to cost between 10–20% of the total mental health budget in high-income countries (Senior et al., 2020). The implications of admission are typically considerable for patients, with average lengths of stay more than 12 months and in conditions that are frequently highly restrictive (Edworthy & Vollm, 2016). There are high rates of repeat offending, which are around 5% per year after hospital discharge, with deleterious impact on discharged patients, their victims and families, and wider society (Fazel et al., 2016). For these reasons, measuring outcomes of care in forensic mental health services is necessary to ensure that services are clinically effective, cost-effective and provide a positive experience for patients and their carers. Repeated longitudinal measurement of progress can also facilitate the identification of patient needs and care planning between patients and their clinical teams (Thomas et al., 2008).

Systematic reviews have consistently shown that existing outcome measures do not cover key domains and fail to adequately capture the patient perspective (Fitzpatrick et al., 2010; Ryland, Cook, Yukhnenko, et al., 2021; Shinkfield & Ogloff, 2014). Many of these measures focus on risk and clinical symptoms, neglecting quality of life and functional outcomes, and these are not viewed as comprehensive by clinicians (Ryland, Cook, Fitzpatrick, et al., 2021). The administrative burden on staff working in forensic mental health services is high (Dickinson & Wright, 2008; Newman et al., 2020). Patients are unlikely to be able to complete complex assessments due to high levels of symptoms and cognitive impairment (Shumlich et al., 2019). Therefore, a new outcome measure is needed that is both comprehensive and easy to use, and which is reported by both clinicians and patients.

We designed a new tool called the Forensic oUtcome Measure (FORUM) that was developed iteratively using best practice techniques (Ryland, Cook, Ferris, et al., 2021; see Appendix 1). This process included qualitative interviews, focus groups, a Delphi survey and cognitive debriefing interviews. The FORUM consists of two complementary instruments that were developed in parallel. The FORUM-P is a patient-reported outcome measure consisting of 20 items, while the FORUM-C is a clinician-reported outcome measure with 23 items. Twelve items in the FORUM-P correspond to 13 items in the FORUM-C, with the item concerning relationships in the FORUM-P broken down into two questions in FORUM-C: one on relationships with staff and the other on relationships with other individuals. The remaining eight items in the FORUM-P and 10 items in the FORUM-C are unique to each instrument.

In this paper, we describe the first psychometric investigation of FORUM in a sample of patients who use forensic mental health services. It assesses a full range of reliability and validity measures (de Vet et al., 2011).

Materials and methods

Setting

The study took place in the forensic mental health service of one regional public healthcare provider in the UK, which covers a population of 2.1 million. This service consists of four male medium secure wards, two male low secure wards, two women's enhanced low secure wards and one mixed-sex low secure pre-discharge ward. Participants were also recruited from the forensic community team. Community patients were in different types of accommodation, ranging from 24 hour specialised forensic community units to independent, private accommodation.

Participants

Recruitment of patients on inpatient units was undertaken by researchers, who were qualified psychiatrists. All available inpatients in the service were invited to participate. The researchers arranged with clinical staff to attend the ward on a suitable day, when patients were likely to be available. The researchers liaised with nursing staff to identify any patients they thought would be unsuitable to participate, for example, because their behaviour had been erratic, and they may pose a risk to the researchers. All other available patients were then approached by the researchers to ask if they would be interested in participating in the study. If the patient said that they would be interested, then the purpose of the study was explained in more detail and informed consent to participate was obtained. Some patients were unavailable, for example, because they had an extended period of leave off the ward. The proportion of patients approached who participated was recorded. For each patient who participated, at least one clinician was identified who knew that patient well enough to complete the FORUM-C. A second clinician per patient was also identified where possible. Clinicians could be from any professional discipline and the professional background of the clinicians was noted.

Recruitment of patients in the community was undertaken by a researcher (HR) who accompanied clinicians from the community forensic team during their visits to patients in the community. The patient was introduced to the researcher and asked if they would like to take part in the research. The number of patients agreeing to participate was recorded. For patients who consented to participate, the accompanying clinicians were asked to complete the FORUM-C.

Data collection

Patients were asked to complete the FORUM-P and clinicians familiar with those patients the FORUM-C. Additional feedback forms were developed for both patients and clinicians which were designed with input from a dedicated patient and public advisory group (MacInnes et al., 2011). These included questions on the comprehensiveness, ease of use and relevance of the new measures, as well as the opportunity to provide general feedback (Terwee et al., 2018). The feedback form for clinicians also included two questions about the perceived practicality and usefulness of the new instrument for their clinical practice (Bauer et al., 2015).

Patients who consented to participate were asked to complete the FORUM-P and the additional feedback form by a researcher. Participants were able to complete the questionnaires themselves, or with the aid of a researcher, who could read out and mark the questionnaires, if preferred. Patients were then invited to complete the FORUM-P again between 1 and 7 days after their initial participation. This was the same questionnaire, with the addition of a question asking them if anything had changed in the interim that would affect their answers (Prinsen et al., 2018). Some participants completed the second FORUM-P with a researcher, while others chose to complete it independently and either leave it for the researcher to collect or return in by post in a stamped, addressed envelope.

Clinicians were asked to complete the FORUM-C on the same day or the following day to when the first FORUM-P was completed. They were then asked to complete the FORUM-C a second time between 1 and 14 days later. One clinician for each patient was also asked to provide a score on the Global Assessment of Functioning (GAF) (Aas, 2011).

Demographic data and other patient characteristics, including sex, age, ethnicity, length of stay, diagnosis, index offence and Mental Health Act status were collected from the electronic patient record. The most recent scores on the Health of the Nation Outcome Scale – secure version (HoNOS-Secure) and the Historical, Clinical, Risk 20 (HCR-20) were also recorded (Dickens et al., 2007; Douglas, 2014). These two tools are currently used as outcome measures and form Key Performance Indicators that services in England must report to commissioners. If there were multiple HoNOS-Secure or HCR-20 ratings, then the one closest to the time of the first FORUM-P was used, regardless of whether this was completed before or after the FORUM-P.

The GAF, HCR-20 and HoNOS-Secure were selected to compare with the FORUM, as this combination of tools covers a range of important outcome domains, including risk, symptoms and functioning.

Ethical approval

Following the approved local pathway, the protocol for this study was submitted to the Clinical Trials and Research Governance Committee at the University of Oxford and to the Research and Development Committee at Oxford University Hospitals NHS Trust, which confirmed that it should be classified as a service evaluation. The protocol was subsequently reviewed and approved by the Forensic Clinical Governance Committee and Clinical Audit Committee of Oxford Health NHS Foundation Trust.

Data analysis

Descriptive characteristics were summarised using standard statistics. The initial FORUM-P and the initial FORUM-C of the first clinician approached were used for all analyses except for test-retest and inter-rater reliability. The numbers of missing individual items and whole scales with one or more missing items were calculated. To facilitate statistical analysis, numerical values between 0 and 4 were assigned to the corresponding response options ranging from never to always. The frequency of responses for individual items and the whole scales were tabulated and graphed as histograms to visually inspect the

distribution of responses (de Vet et al., 2011). Any items or aggregate scores with missing data were excluded from statistical analyses.

Exploratory factor analysis (EFA) was performed separately for FORUM-P and FORUM-C using several commonly used approaches (principal factors, principal component factors, maximum likelihood and iterated principal factors), to identify the most appropriate solution. Factors with eigenvalues greater than one were considered for retention. Additionally, a qualitative assessment of items was carried out to decide how many factors should be retained. Varimax rotation was performed to increase the interpretability of the retained factors (Watson, 2017). Internal consistency was assessed using Cronbach's alpha for the FORUM-P and FORUM-C as a whole and the individual factors identified using the EFA (Cronbach, 1951), with 95% confidence intervals calculated.

The test-retest reliability for the FORUM-P and FORUM-C scores was tested using a weighted kappa statistic (Cohen, 1960) using linear weights. The correlation of the overall FORUM-C scores of the first and second clinician was determined using Spearman's rank correlation and 95% confidence intervals were calculated (Zar, 2005).

The correlation of the total FORUM-P and FORUM-C score was calculated for the whole sample and for each level of security using Spearman's rank correlation. The correlation of the total scores of the mirrored items that appeared in both the FORUM-P and FORUM-C was also calculated. The correlation of the total FORUM-P and FORUM-C scores with the GAF, total HoNOS-Secure, total HCR-20 and the sum of the dynamic scales (Clinical and Risk) of the HCR-20 was calculated using Spearman's rank correlation. A higher score on the GAF indicates superior functioning, while lower scores on the HoNOS-Secure and HCR-20 indicate better outcomes and lower risk respectively (Aas, 2011; Dickens et al., 2007; Douglas, 2014). It was hypothesised that the FORUM, and especially the FORUM-C as another clinician-rated instrument, would therefore correlate positively with the GAF and negatively with the HoNOS-Secure and HCR-20. A difference in the total scores between the three levels of security was assessed using a Kruskal–Wallis test (Sawilowsky & Fahoome, 2005). Nominal statistical significance was assessed at the 2-sided 5% significance level.

The number of responses and mean scores were calculated for the additional feedback questions for patients and clinicians.

All analyses were conducted in Stata version 16.1.

Results

Sample characteristics

A total of 62 patients participated out of 122 who were approached, giving an overall participation rate of 51%. The mean age of participants was 41.0 years (standard deviation 11.3 years). The median length of stay for inpatients ($n = 51$) was 667 days (inter-quartile range 219-1479). The other characteristics of the participants are described in Table 1.

Item response rates and distribution

No items were consistently omitted, with a maximum omission rate of 2/62 (3%) for any one item in the FORUM-P and 4/62 (6%) for the FORUM-C. Some clinicians did raise

Table 1. Sample characteristics.

	N	%
Sex		
Male	52	83.9
Female	10	16.1
Ethnicity		
White	46	74.2
Black	7	11.3
Asian	6	9.7
Mixed	3	4.8
Level of security		
Medium	24	38.7
Low	27	43.6
Community	11	17.7
Diagnosis		
Schizophrenia spectrum	46	74.2
Personality disorder	10	16.1
Other	6	9.7
Index offence		
Serious violent crime	24	38.7
Sexual offence	6	9.7
Other	32	51.6
Legal detention status (Mental Health Act)		
Forensic	47	75.8
Prison	8	12.9
Other	7	11.3
Source of admission (Inpatients only $n = 51$)		
Prison	13	25.5
Mental health unit	9	17.6
Community	29	56.9

concerns about their ability to answer particular items, for example, the ability of occupational therapists or psychologists to comment on medication usage. There was no clear link between these concerns and the items with higher omission rates (Krosnick, 2018). See Appendix 2 for full details of missing items.

Overall, the distribution of responses for individual items appeared acceptable. The distribution of responses was negatively skewed for many items in both the FORUM-P and FORUM-C with 'often' the most popular response for 13/20 (65%) FORUM-P items and 13/23 (57%) FORUM-C items. There were no noticeable floor or ceiling effects for any items.

Factor analysis and internal consistency

Exploratory factor analysis was carried out initially using the principal factors approach which produced three factors with eigenvalues of greater than one for the FORUM-P and four for the FORUM-C. The corresponding factors accounted for 76% and 82% of the variance respectively. Findings were similar for the other factor analysis methods. For both questionnaires, once the factors to be retained was determined, varimax rotation was carried out for default factor analysis. The items that loaded on to each of the retained factors (after rotation) are described in Tables 2 and 3 for FORUM-P and FORUM-C respectively. Table 4 describes the internal consistency of the FORUM-P, FORUM-C and all retained factors.

For the FORUM-P, factor one included items that appeared to be related to general wellbeing, with the exception of item 13 'I have been actively working on reducing my risk of

Table 2. Items loading most strongly onto the three retained factors following exploratory factor analysis of the FORUM-P with varimax rotation.

Items	Factor loading		
	Factor 1	Factor 2	Factor 3
Loading to factor 1			
Item 1 – I have felt good about myself	0.8620	0.1125	−0.0270
Item 3 – My life has felt meaningful	0.8258	0.2030	0.0639
Item 5 – I have had the relationships with others I have wanted	0.6696	−0.0030	0.3759
Item 10 – I have slept well	0.6208	0.1777	−0.2670
Item 7 – My mental and emotional health has been good	0.6111	−0.2854	0.2015
Item 19 – I have made progress towards greater independence	0.5318	0.3926	0.2261
Item 4 – I have taken part in activities I have found enjoyable	0.4638	0.2325	0.1716
Item 11 – I have felt safe	0.4224	0.3767	−0.1102
Item 13 – I have been actively working on reducing my risk of harm to others	0.0688	0.0202	0.0322
Loading to factor 2			
Item 20 – I have made use of the support offered by mental health services	0.1582	0.7353	0.1815
Item 18 – I have had the opportunity to develop the skills I want	0.2093	0.7033	−0.0519
Item 15 – I have been able to ask for help when I have needed it	−0.0275	0.6922	0.2316
Item 6 – I have been actively involved in planning my care	0.3106	0.6408	0.3970
Item 14 – I have been able to trust others	0.2706	0.5952	0.0583
Item 9 – I have found taking my medication helpful	0.0152	0.5739	0.0140
Item 8 – I have been helped by talking therapies	−0.2255	0.5384	0.3095
Item 2 – I have felt hopeful	0.5123	0.5377	−0.2797
Loading to factor 3			
Item 16 – I have been able to take decisions when I have needed to	−0.0149	0.1499	0.8226
Item 17 – I have been able to make the plans that I have wanted to	0.3252	0.2095	0.5753
Item 12 – I have been aware of how my behaviour could affect others	0.0428	0.1599	0.3621

Table 3. Items loading most strongly on to the four retained factors following exploratory factor analysis of the FORUM-C with varimax rotation.

Items	Factor loading			
Loading to factor 1	Factor1	Factor 2	Factor 3	Factor 4
Item 17 – … has actively participated in planning their care	0.6974	0.2078	0.3517	0.1752
Item 22 – … has made progress towards greater independence	0.6828	0.1828	0.3678	0.2475
Item 23 – … has made use of the support offered by mental health services	0.6622	0.2759	0.1219	0.1614
Item 6 – … has benefitted from talking therapies	0.6544	0.3402	0.0297	−0.1216
Item 20 – … has asked for help when needed	0.6012	0.1361	0.1072	0.4309
Item 21 – … has taken the decisions they have needed to	0.5940	0.3522	0.1114	0.2566
Item 16 – … has taken part in activities	0.5730	0.1664	0.2297	0.2951
Item 7 – … appears to have slept a good amount (not too little or too much)	0.4901	−0.0579	0.3724	−0.1412
Item 3 – … has had good physical health	0.4602	0.0654	0.2570	0.1068
Loading to factor 2				
Item 9 – … has avoided behaviours that reduce their safety	0.2333	0.8757	0.0973	0.0830
Item 10 – … has managed difficult situations well	0.2942	0.7182	0.1322	0.2747
Item 11 – … has accepted rules that enable everyone to remain safe	0.2590	0.7123	0.4133	0.1346
Item 13 – Others feel safe around them	−0.1861	0.7095	0.3714	−0.0444
Item 14 – … has had positive relationships with staff	0.3584	0.6280	0.4483	0.0380
Item 8 – … has avoided misuse of substances (alcohol and drugs)	0.1627	0.5126	−0.2745	0.3130
Loading to factor 3				
Item 15 – … has had positive relationships with other people	0.1175	0.4211	0.6526	0.0998
Item 1 – … has had good mental health	0.3349	0.2593	0.6192	0.3331
Item 12 – … has actively worked on reducing their risk of harm to others	0.3284	0.4261	0.6091	0.0907
Item 19 – … has appeared able to trust others	0.2530	0.2708	0.6051	0.3825
Item 2 – … has had a good understanding of their mental health	0.5257	0.1684	0.5487	0.0272
Loading to factor 4				
Item 18 – … has appeared confident, with good self-esteem	0.2462	−0.0210	0.2192	0.7551
Item 5 – … has managed any side effects from their medication	0.0281	0.3150	−0.0218	0.6283
Item 4 – … has taken their prescribed medication	0.0782	0.2233	0.2154	0.6108

Table 4. Internal consistency of the FORUM-P, FORUM-C and retained factors from exploratory factor analysis.

Items	Cronbach's *α* (95% confidence interval)
FORUM-P	
All FORUM-P items	0.87 (0.80–0.93)
Factor 1	0.78 (0.68–0.87)
Factor 1 excluding item 13	0.82 (0.73–0.92)
Factor 2	0.84 (0.77–0.90)
Factor 3	0.65 (0.51–0.80)
FORUM-C	
All FORUM-C items	0.93 (0.91–0.96)
Factor 1	0.87 (0.82–0.93)
Factor 2	0.88 (0.83–0.93)
Factor 3	0.86 (0.81–0.92)
Factor 4	0.75 (0.59–0.90)

harm to others'. This item did not load strongly on to any of the three retained factors and had a very high uniqueness. As it loaded most strongly on to factor 1, it was grouped with these items. The internal consistency of factor 1 in the FORUM-P was analysed with and without item 13 included due to its low loading on to that factor. Cronbach's alpha increased only marginally from 0.78 to 0.82 when item 13 was excluded (see Table 4). Factor two contained items which all seemed related to engagement with, and response to, treatment. This thematic coherence was reflected in the good internal consistency, with Cronbach's alpha of 0.84 (see Table 4). The final factor contained only three items, with two dealing with aspects of agency and one about insight into the effect of one's behaviour on others. These appeared to be less clearly linked conceptually and this was reflected in the questionable internal consistency with Cronbach's alpha of 0.65 (see Table 4).

For the FORUM-C, the first factor concerned themes of engagement with treatment and physical well-being. The second factor contained items which concerned management of risk. The third factor contained items that were related to themes of insight and trust. The first three factors all had good internal consistency, with values of Cronbach's alpha in excess of 0.8 (see Table 4). The final factor appeared to contain items that were less coherently grouped into two themes of medication and confidence. This was reflected in a slightly lower internal consistency, with Cronbach's alpha of 0.75 (see Table 4).

Reliability

The weighted kappa for test-retest reliability for the overall FORUM-P score between time 1 and time 2 (*n* = 32) was 0.44 with a 95% confidence interval of 0.24–0.63. The weighted kappa for test-retest reliability for the overall FORUM-C score between clinician 1 at time 1 and time 2 (*n* = 39) was 0.78 with a 95% confidence interval (CI) of 0.73–0.85. The Spearman correlation coefficient for the overall FORUM-C score between the first rating by clinician 1 and clinician 2 (*n* = 38) was 0.47 (95% CI 0.18–0.69).

Hypothesis testing

The Spearman correlation coefficient between the overall FORUM-P and FORUM-C scores for the full sample of complete responses (*n* = 45) was 0.14 (95% CI −0.16 to 0.42). Table 5 describes the correlations between the FORUM-P and FORUM-C at different levels of security. Eleven items directly matched between the FORUM-P and the FORUM-C. The

Table 5. Correlation between the FORUM-P and FORUM-C scores at different levels of security.

Level of security	N	Spearman's ρ (95% confidence interval)
Medium secure unit	19	−0.24 (−0.63 to 0.24)
Low secure unit	21	0.32 (−0.13 to 0.66)
Community	5	0.82 (−0.22 to 0.99)

Table 6. Correlation of the FORUM-P and FORUM-C with the GAF, HoNOS-Secure, HCR-20 total and HCR-20 dynamic.

	N	Spearman's ρ (95% confidence interval)
GAF		
FORUM-P	54	0.17 (−0.10 to 0.42)
FORUM-C	51	0.33 (0.06 to 0.58)
HoNOS-Secure		
FORUM-P	50	−0.28 (−0.51 to 0.00)
FORUM-C	49	−0.29 (−0.53 to −0.01)
HCR-20 total score		
FORUM-P	41	0.09 (−0.22 to 0.39)
FORUM-C	41	−0.40 (−0.63 to −0.11)
HCR-20 dynamic score		
FORUM-P	48	−0.10 (−0.37 to 0.19)
FORUM-C	46	−0.31 (−0.55 to −0.03)

Spearman correlation coefficient for the total score of these 11 items for the full sample of complete responses ($n = 50$) was 0.24 with a 95% confidence interval of −0.05 to 0.48. Table 6 describes the correlation of the FORUM-P and FORUM-C with the GAF, HoNOS-Secure, HCR-20 total score and HCR-20 dynamic score.

When the correlation was analysed by level of security, the correlation was small and negative for medium security, moderate and positive for low security and high and positive for the community, however the 95% confidence intervals contained zero for the full sample and each level of security subsample. Using the Kruskal-Wallis test, no evidence of significant difference for the FORUM-P (Chi-square^2 = 0.93, $p = 0.63$) were found between the three levels of security (medium secure unit $n = 22$, low secure unit $n = 23$, community $n = 9$), or for the FORUM-C (Chi-square^2 = 0.84, $p = 0.66$) (medium secure unit $n = 20$, low secure unit $n = 24$, community $n = 7$).

Summary of participant feedback

Feedback was received from 61 out of 62 patients who participated and all 35 clinicians, including 3 psychiatrists, 25 nurses, 6 occupational therapists and 1 psychologist. Patients gave an average score of 4.0 out of 5 for comprehensiveness, 4.6 for ease of use and 3.9 for relevance. Clinicians gave average scores of 4.1 for comprehensiveness, 4.5 for ease of use and 4.3 for relevance. Clinicians gave an average score of 4.0 out of 5 for practicality in clinical practice and 4.2 for usefulness.

Discussion

Main findings

This study is the first evaluation of the performance of the FORUM, a new instrument for measuring outcomes in forensic mental health services. Sixty-two patients and 35

clinicians completed the FORUM. A full range of measures of reliability and validity were tested.

The acceptability of individual items in the FORUM appeared to be good. The FORUM-P and FORUM-C were both positively received by patients and clinicians respectively. The negative skew of responses may be a result of the constitution of the sample, as those patients who agreed to participate may have been less unwell and therefore have higher ratings on both questionnaires (Spencer et al., 2018). Other possible causes include the lack of a high secure sample and respondents giving more favourable responses to please researchers. High average scores were recorded for all feedback questions, particularly for ease of use. Patients' comments were mostly supportive of the tool's usefulness and highlighted the positive framing of the questions as appealing (Sonderen et al., 2013). They also noted that the areas covered by the questionnaire may not otherwise be routinely considered by staff.

The internal consistency of the FORUM-P and FORUM-C scales was good, indicative of a unifying conceptual underpinning to each of the two instruments (Bland & Altman, 1997). The results of exploratory factor analysis suggested that certain items in the FORUM-P and FORUM-C appeared to be statistically related and could represent potential subscales (Watson, 2017). Given the relatively small numbers who participated in the study, caution is needed when interpreting the findings and these should be viewed as preliminary, pending further analyses with larger samples (Prinsen et al., 2018).

The test-retest reliability was moderate for the FORUM-P and good for FORUM-C (Cohen, 1960). The lack of perfect agreement may reflect the fact that the scales are not anchored in vignettes and measure complex concepts (King & Wand, 2007). The higher reliability for the FORUM-C may be partly explained by the fact that clinicians would not necessarily see the patient being rating again between the first and second times they completed the FORUM-C. This meant that their impression of the patient's condition was likely to remain stable (Powers et al., 2017). Patients were not asked to repeat the measure if they believed that their answers would definitely be different the second time. However, some did say they could not be sure their answers would not change and these responses were included in the analysis. The gap of up to six days between patients' first and second ratings could mean that events occurred in the interim that resulted in a change in responses (de Vet et al., 2011). It is possible that patients may not have been aware of the effect of such events on their responses. The correlation between the two clinician raters was high, indicating good inter-rater reliability for the FORUM-C (Zar, 2005).

The expected direction of correlation was observed with all other measures, with the exception of the FORUM-P and the HCR-20 total score, which had a very small positive correlation. No significant differences between total score by level of security were observed, but the numbers in each group, especially in the community were low (Pillay et al., 2008).

Implications for policy and practice

The items in the FORUM-P and FORUM-C are not all reflections of each other, with several questions that are unique to both. The limited correlation between the overall FORUM-P

and FORUM-C was unsurprising given these differences. This may well reflect the different underlying items and thereby the divergence in perspective between patients and clinicians (Chandwani et al., 2017). Services, therefore, need to ask patients for their views on outcomes and the questions may be different to those posed to clinicians (Collins, 2019). If there is truly an increase in concordance between these scores as patients move through the pathway, this could be useful as an indicator of recovery, although the importance of concordance should not be over-emphasised. For FORUM, this would need to be confirmed in a larger sample which could provide more precise results. A lack of concordance however does not mean that either perspective is incorrect (Collins, 2019). Greater concordance, especially where the patient rating is lower than that of their clinical team, could provide an indication of unmet need (Thomas et al., 2008). In such cases, the clinical team may be unaware of aspects of the patient's experience of their life and care within services. This could offer valuable signals about how patients view their own mental health, potentially concerning important concepts not assessed by other existing measures (Shinkfield & Ogloff, 2014).

The positive engagement of patients and clinicians with the questionnaires in this study suggests that the FORUM would be acceptable for use in forensic services (Mezey et al., 2013). Care is required during implementation to gain the confidence of patients and encourage uptake (Bauer et al., 2015). This should focus on ensuring that patients see it as a meaningful tool in their care (Collins, 2019). Equally important will be to support clinicians to use it, as they will be responsible for offering patients the opportunity to complete the FORUM-P. The burden of completing the tool was low for most participants, including both patients and clinicians. Most participants were able to complete all the questions rapidly, with high support for it being easy to use (Terwee et al., 2018). This supports its suitability for use in routine clinical practice, where it could be applied repeatedly at regular intervals, to provide longitudinal information about a patient's progress (Wolf et al., 2018).

The substantial but imperfect correlation between clinician raters may mean that it is better for the FORUM-C to be completed by one consistent clinician. Alternatively, it may be more appropriate for the multidisciplinary team to undertake a joint assessment, which would have the advantage of drawing on the wider knowledge, perspectives and expertise of the various team members. A potential disadvantage of this approach could be that it would take much longer to rate a patient due to the time taken to reach consensus (Maddock, 2015). The moderate FORUM-P test-retest reliability indicates that careful attention should be paid to the timing of questionnaire completion (de Vet et al., 2011). Particular events may have a short-lived effect on the way that someone answers the questionnaire, which may not be representative of how they would respond otherwise. For example, it may be better for patients to avoid answering questions immediately after receiving good or bad news, or following an incident with another patient.

Correlations in the expected direction between FORUM-C and the other three clinician-reported measures provide support for the construct validity of this questionnaire (Prinsen et al., 2018). None of these other instruments can be considered a gold standard outcome measure in forensic mental health services, therefore this cannot be considered evidence of criterion validity (Fazel & Wolf, 2017; Shinkfield & Ogloff, 2014). While there are some common themes between their content and the items in FORUM-C, there are other concepts in the FORUM-C which are not considered by the other measures.

Further work is needed to determine how the new instrument could be best used in routine care (Bauer et al., 2015). In this study, the FORUM-P and FORUM-C were administered separately to patients and their clinicians sequentially. Patients and clinicians did not meet to discuss their respective answers or utilise the responses to identify need or formulate a care plan. Repeated measures were only gathered in a very short time span to assess test-retest reliability and this was not intended to assess responsiveness (Terwee et al., 2003). The ability for an individual to track their progress over time is one of the most important features of an outcome measure, however, this property was not examined in the current study. Further studies to establish the ability of the FORUM to measure change over time are therefore needed.

This study employed a pragmatic approach to gathering data, for example by using various clinicians from the patients' regular teams to complete the FORUM-C. This should mean that the data more accurately reflects the situation in clinical practice (Holtrop & Glasgow, 2020). It also collected data on content validity and this included clinicians' views on the potential use of the FORUM-C in routine practice (Terwee et al., 2018).

Limitations

Only around a half of the patients approached agreed to participate, although it is recognised that the level of participation of patients in forensic settings in research is often low (Völlm et al., 2017). The number of participants limited precision for certain statistical tests such as the exploratory factor analysis. Although the sample size was similar to some other validation studies of new outcome measures, larger studies are needed to estimate psychometric parameters with greater precision (Ryland, Cook, Yukhnenko, et al., 2021; Watson, 2017). Most participants were male, although this does reflect the characteristics of patients in forensic mental health services. The sample did not include any participants from a high-security setting (Tapp et al., 2013). Although community patients were included, their numbers were small. Participants were from one state-funded hospital provider in England. Cultural issues and differences in forensic mental health pathways may limit the generalisability to other countries. The limited ethnic diversity of the sample, which contained mostly white participants may also limit the relevance of findings to regions and countries with a different range and proportion of ethnicities.

Items were all given equal weight and no consideration was given to assigning varying weights to score the scale. In future studies, a more definitive scoring system could be developed. Some important psychometric properties were not considered in this study. Confirmatory factor analysis was not performed to establish the goodness of fit of the factor structure identified through the exploratory factor analysis. Responsiveness was not determined due to the limited time available for follow-up and the long anticipated timescale for meaningful change to occur in this population. A difference in score between patients by level of security could provide some proxy evidence of responsiveness (Pillay et al., 2008). Ultimately though, evidence is required of responsiveness within individuals over time (de Vet et al., 2011). This would ideally span moves between levels of security as an objective indicator of change (Kennedy, 2002). This would require caution about the interpretation of score immediately before and after a move, due to disruption around moves between levels, and especially on discharge,

that could temporarily impact patients' responses. Assessing what score would constitute a meaningful difference to responders should be determined by future studies.

Participants' completion of the feedback form for both the FORUM-P and FORUM-C was not anonymous. Respondents therefore may have provided more positive answers, to avoid being seen to criticise the researchers who developed the instrument (Krosnick, 2018). Over 70% of the clinicians who participated were nursing staff. This meant that there were limited views from other professional groups, with some groups, such as social workers, not represented at all.

Conclusions

Overall, we found that the psychometric properties of FORUM-P and FORUM-C were acceptable and, on some measures, good. FORUM-P and FORUM-C provide a novel, psychometrically robust set of complementary instruments to monitor a comprehensive range of outcomes in forensic mental health. Further studies are needed in larger samples to confirm these findings with greater precision (Prinsen et al., 2018). Furthermore, it is necessary to conduct additional studies to assess other psychometric parameters, in particular responsiveness within individuals over time (Terwee et al., 2003) and what degree of change would constitute a meaningful difference to stakeholders. More work is needed to determine how the FORUM could be most effectively utilised within a routine clinical context, including the system used to score items (Bauer et al., 2015).

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Howard Ryland ⓘ http://orcid.org/0000-0002-5287-555X

References

Aas, I. H. M. (2011). Guidelines for rating Global Assessment of Functioning (GAF). *Annals of General Psychiatry*, *10*(2). https://doi.org/10.1186/1744-859X-10-2.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology*, *3*(1), 32. https://doi.org/10.1186/s40359-015-0089-9

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*(7080), 572. https://doi.org/10.1136/bmj.314.7080.572

Chandwani, K. D., Zhao, F., Morrow, G. R., Deshields, T. L., Minasian, L. M., Manola, J., & Fisch, M. J. (2017). Lack of patient-clinician concordance in cancer patients: Its relation with patient variables. *Journal of Pain and Symptom Management*, *53*(6), 988–998. https://doi.org/10.1016/j.jpainsymman.2016.12.347

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collins, B. (2019). *Outcomes for mental health services. What really matters?* King's Fund.

Crocker, A. G., Livingston, J. D., & Leclair, M. C. (2017). Forensic mental health systems internationally. In R. Roesch & A. N. Cook (Eds.), *Handbook of forensic mental health services* (pp. 3–76). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315627823-2

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge University Press.

Dickens, G., Sugarman, P., & Walker, L. (2007). HoNOS-secure: A reliable outcome measure for users of secure and forensic mental health services. *Journal of Forensic Psychiatry & Psychology*, *18*(4), 507–514. https://doi.org/10.1080/14789940701492279

Dickinson, T., & Wright, K. M. (2008). Stress and burnout in forensic mental health nursing: A literature review. *British Journal of Nursing*, *17*(2), 82–87. https://doi.org/10.12968/bjon.2008.17.2.28133

Douglas, K. S. (2014). Version 3 of the historical-clinical-risk management-20 (HCR-20V3): Relevance to violence risk assessment and management in forensic conditional release contexts [review]. *Behavioral Sciences & The Law*, *32*(5), 557–576. https://doi.org/10.1002/bsl.2134

Edworthy, R., & Vollm, B. (2016). Long-stay in high and medium secure forensic psychiatric care - prevalence, patient characteristics and pathways in England. *European Psychiatry*, *33*(SUPPL.), S180. https://doi.org/10.1016/j.eurpsy.2016.01.385

Fazel, S., Fiminska, Z., Cocks, C., & Coid, J. (2016). Patient outcomes following discharge from secure psychiatric hospitals: Systematic review and meta-analysis. *British Journal of Psychiatry*, *208*(1), 17–25. https://doi.org/10.1192/bjp.bp.114.149997

Fazel, S., & Wolf, A. (2017). Selecting a risk assessment tool to use in practice: A 10-point guide [review]. *Evidence Based Mental Health*, *21*, 21. https://doi.org/10.1136/eb-2017-102861

Fitzpatrick, R., Chambers, J., Burns, T., Doll, H., Fazel, S., Jenkinson, C., Kaur, A., Knapp, M., Sutton, L., & Yiend, J. (2010). A systematic review of outcome measures used in forensic mental health research with consensus panel opinion. *Health Technology Assessment (Winchester, England)*, *14*(18), 1–94. https://doi.org/10.3310/hta14180

Holtrop, J. S., & Glasgow, R. E. (2020). Pragmatic research: An introduction for clinical practitioners. *Family Practice*, *37*(3), 424–428. https://doi.org/10.1093/fampra/cmz092

Kennedy, H. G. (2002). Therapeutic uses of security: Mapping forensic mental health services by stratifying risk. *Advances in Psychiatric Treatment*, *8*(6), 433–443. https://doi.org/10.1192/apt.8.6.433

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(1), 46–66. https://doi.org/10.1093/pan/mpl011

Krosnick, J. A. (2018). Questionnaire design. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 439–455). Springer.

Latham, R., & Williams, H. K. (2020). Community forensic psychiatric services in England and Wales. *CNS Spectrums*, *25*(5), 604–617. https://doi.org/10.1017/S1092852919001743

MacInnes, D., Beer, D., Keeble, P., Rees, D., & Reid, L. (2011). Service-user involvement in forensic mental health care research: Areas to consider when developing a collaborative study. *Journal of Mental Health*, *20*(5), 464–472. https://doi.org/10.3109/09638231003728109

Maddock, A. (2015). Consensus or contention: An exploration of multidisciplinary team functioning in an Irish mental health context. *European Journal of Social Work*, *18*(2), 246–261. https://doi.org/10.1080/13691457.2014.885884

Mezey, G., White, S., Thachil, A., Berg, R., Kallumparam, S., Nasiruddin, O., Wright, C., & Killaspy, H. (2013). Development and preliminary validation of a measure of social inclusion for use in people with mental health problems: The SInQUE. *International Journal of Social Psychiatry*, *59* (5), 501–507. https://doi.org/10.1177/0020764012443752

Newman, C., Jackson, J., Macleod, S., & Eason, M. (2020). A survey of stress and burnout in forensic mental health nursing. *Journal of Forensic Nursing*, *16*(3), 161–168. https://doi.org/10.1097/jfn.0000000000000271

Pillay, S. M., Oliver, B., Butler, L., & Kennedy, H. G. (2008). Risk stratification and the care pathway. *Irish Journal of Psychological Medicine*, *25*(4), 123–127. https://doi.org/10.1017/S0790966700011228

Powers, J. H., Patrick, D. L., Walton, M. K., Marquis, P., Cano, S., Hobart, J., Isaac, M., Vamvakas, S., Slagle, A., Molsen, E., & Burke, L. B. (2017). Clinician-reported outcome assessments of treatment benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. *Value in Health*, *20*(1), 2–14. https://doi.org/10.1016/j.jval.2016.11.005

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Ryland, H., Cook, J., Ferris, R., Markham, S., Sales, C., Fitzpatrick, R., & Fazel, S. (2021). Development of the FORUM: A new patient and clinician reported outcome measure for forensic mental health services. *Psychology, Crime & Law*, 1–18. https://doi.org/10.1080/1068316X.2021.1962873

Ryland, H., Cook, J., Fitzpatrick, R., & Fazel, S. (2021). Ten outcome measures in forensic mental health: A survey of clinician views on comprehensiveness, ease of use and relevance. *Criminal Behaviour & Mental Health*, *31*(6). https://doi.org/10.1002/cbm.2221

Ryland, H., Cook, J., Yukhnenko, D., Fitzpatrick, R., & Fazel, S. (2021). Outcome measures in forensic mental health services: A systematic review of instruments and qualitative evidence synthesis. *European Psychiatry*, *64*(1), e37. https://doi.org/10.1192/j.eurpsy.2021.32

Sawilowsky, S., & Fahoome, G. (2005). Kruskal-Wallis test. In B. Everitt & D. Howell (Eds.), *Encyclopaedia of statistics in behavioral science*. John Wiley and Sons. https://doi.org/10.1002/0470013192.bsa333

Senior, M., Fazel, S., & Tsiachristas, A. (2020). The economic impact of violence perpetration in severe mental illness: A retrospective, prevalence-based analysis in England and Wales. *Lancet Public Health*, *5*(2), e99–e106. https://doi.org/10.1016/S2468-2667(19)30245-2

Shinkfield, G., & Ogloff, J. (2014). A review and analysis of routine outcome measures for forensic mental health services. *International Journal of Forensic Mental Health*, *13*(3), 252–271. https://doi.org/10.1080/14999013.2014.939788

Shumlich, E. J., Reid, G. J., Hancock, M., & Hoaken, P. (2019). Executive dysfunction in Criminal populations: Comparing forensic psychiatric patients and correctional offenders. *International Journal of Forensic Mental Health*, *18*(3), 243–259. https://doi.org/10.1080/14999013.2018.1495279

Sonderen, E. V., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLOS ONE*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Spencer, B. W. J., Gergel, T., Hotopf, M., & Owen, G. S. (2018). Unwell in hospital but not incapable: Cross-sectional study on the dissociation of decision-making capacity for treatment and research in in-patients with schizophrenia and related psychoses. *British Journal of Psychiatry*, *213*(2), 484–489. https://doi.org/10.1192/bjp.2018.85

Tapp, J., Warren, F., Fife-Schaw, C., Perkins, D., & Moore, E. (2013). What do the experts by experience tell us about 'what works' in high secure forensic inpatient hospital services? *Journal of Forensic Psychiatry & Psychology*, *24*(2), 160–178. https://doi.org/10.1080/14789949.2012.760642

Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research*, *12*(4), 349–362. https://doi.org/10.1023/A:1023499322593

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research*, *27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Thomas, S. D., Slade, M., McCrone, P., Harty, M. A., Parrott, J., Thornicroft, G., & Leese, M. (2008). The reliability and validity of the forensic Camberwell Assessment of Need (CANFOR): A needs assessment for forensic mental health service users. *International Journal of Methods in Psychiatric Research*, *17*(2), 111–120. https://doi.org/10.1002/mpr.235

Völlm, B., Foster, S., Bates, P., & Huband, N. (2017). How best to engage users of forensic services in research: Literature review and recommendations. *International Journal of Forensic Mental Health*, *16*(2), 183–195. https://doi.org/10.1080/14999013.2016.1255282

Vollm, B. A., Clarke, M., Herrando, V. T., Seppanen, A. O., Gosek, P., Heitzman, J., & Bulten, E. (2018). European psychiatric Association (EPA) guidance on forensic psychiatry: Evidence based assessment and treatment of mentally disordered offenders. *European Psychiatry*, *51*, 58–73. https://doi.org/10.1016/j.eurpsy.2017.12.007

Watson, J. C. (2017). Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development*, *50*(4), 232–238. https://doi.org/10.1080/07481756.2017.1336931

Wolf, A., Fanshawe, T., Sariaslan, A., Cornish, R., Larsson, H., & Fazel, S. (2018). Prediction of violent crime on discharge from secure psychiatric hospitals: A clinical prediction rule (FoVOx). *European Psychiatry*, *47*, 88–93. https://doi.org/10.1016/j.eurpsy.2017.07.011

Zar, J. H. (2005). Spearman rank correlation. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics*. John Wiley & Sons. https://doi.org/10.1002/0470011815.b2a15150